Routledge
Taylor & Francis Group

European Journal of Work and Organizational Psychology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/pewo20

Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework

Anna-Sophie Ulfert, Eleni Georganta, Carolina Centeio Jorge, Siddharth Mehrotra & Myrthe Tielman

To cite this article: Anna-Sophie Ulfert, Eleni Georganta, Carolina Centeio Jorge, Siddharth Mehrotra & Myrthe Tielman (2024) Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework, European Journal of Work and Organizational Psychology, 33:2, 158-171, DOI: 10.1080/1359432X.2023.2200172

To link to this article: https://doi.org/10.1080/1359432X.2023.2200172

Published online: 20 Apr 2023.

EUROPEAN JOURNAL OF WORK AND ORGANIZATIONAL PSYCHOLOGY
2024, VOL. 33, NO. 2, 158–171
https://doi.org/10.1080/1359432X.2023.2200172

Routledge
Taylor & Francis Group

Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework

Anna-Sophie Ulfert[a], Eleni Georganta[b], Carolina Centeio Jorge[c], Siddharth Mehrotra[c] and Myrthe Tielman[c]

[a]Human Performance Management Group, Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, Eindhoven, the Netherlands; [b]Programme group Work and Organizational Psychology, Faculty of Social and Behavioural Sciences, University of Amsterdam, Amsterdam, the Netherlands; [c]Interactive Intelligence Group, Faculty of Electrical Engineering, Mathematics & Computer Science, Delft University of Technology, Delft, the Netherlands

ABSTRACT

Intelligent systems are increasingly entering the workplace, gradually moving away from technologies supporting work processes to artificially intelligent (AI) agents becoming team members. Therefore, a deep understanding of effective human-AI collaboration within the team context is required. Both psychology and computer science literature emphasize the importance of trust when humans interact either with human team members or AI agents. However, empirical work and theoretical models that combine these research fields and define team trust in human-AI teams are scarce. Furthermore, they often lack to integrate central aspects, such as the multilevel nature of team trust and the role of AI agents as team members. Building on an integration of current literature on trust in human-AI teaming across different research fields, we propose a multidisciplinary framework of team trust in human-AI teams. The framework highlights different trust relationships that exist within human-AI teams and acknowledges the multilevel nature of team trust. We discuss the framework's potential for human-AI teaming research and for the design and implementation of trustworthy AI team members.

ARTICLE HISTORY
Received 31 August 2021
Accepted 3 April 2023

KEYWORDS
Artificial intelligence; human-AI teams; trust; team trust; multidisciplinary

Introduction

Artificial intelligence (AI) is playing an increasingly important role within the work context and has been described as one of the most impactful current trends (Kaplan & Haenlein, 2020). Instead of fully replacing employees, current research highlights that AI can be most beneficial when humans and AI systems closely collaborate, complementing each other's strengths and weaknesses (Centeio Jorge et al., 2022). With their continuously improving capabilities, AI systems are moving away from being a technology used as a tool towards becoming members of work teams (Larson & DeChurch, 2020). As AI systems become as agentic as their human colleagues, the future promises even more interaction with autonomous technologies (Seeber et al., 2020). These developments are beginning to show in many areas of organizational life, for example, in algorithmic leadership (e.g., at Uber; Leicht Deobald et al., 2019), in personnel selection (Mirowska, 2020), or even as direct team members in a panel discussion at a conference (Cummings et al., 2021).

A human-AI team can be defined as a collection of human individuals and one or more AI agents (van den Bosch et al., 2019), where an AI agent is a computer entity that perceives and acts in an environment[1] (Russell & Norvig, 2016) and possesses "a partial or high degree of self-governance with respect to decision-making, adaptation, and communication" (O'Neill et al., 2022, p. 904). AI agents as team members are typically defined by their function and capabilities rather than their physical features. Thus, depending on their degree of autonomy and interdependence, chatbots (e.g., ChatGPT), virtual assistants, or robots can act as AI team members. Within a team, AI and human team members, interact virtually, possess common goals, and are brought together to perform a task (Kozlowski & Ilgen, 2006; McNeese et al., 2019). Further, human-AI teams "exhibit interdependencies with respect to workflow, goals, and outcomes, have different roles and responsibilities, and are together embedded in an encompassing organizational system, with boundaries and linkages to the broader system context and task environment." (Kozlowski & Ilgen, 2006, p. 79). Successful human-AI team collaboration can become extremely powerful. For instance, in a team consisting of human nurses and an AI agent, the AI team member may be able to support the team by providing direct access to a patient's medical records and recommend treatment plans via its connection to a central system. At the same time, human nurses may be better at socially interacting with patients and making health-related decisions. In this way, each team member can support how the team operates as a whole and the team's performance (with their individual competences). Although this may currently seem like a futuristic scenario, human-AI teams, such as the one described, can already be found in different sectors of the healthcare domain or in aviation (Hengstler et al., 2016; McNeese et al., 2019; O'Neill et al., 2022). With the continuous development of such technologies, AI agents will soon take over more complex tasks, increasing

CONTACT Anna-Sophie Ulfert a.s.ulfert.blank@tue.nl

This article has been republished with minor changes. These changes do not impact the academic content of the article.

the interdependence between them and their human team members (Seeber et al., 2020). Yet, there are still central challenges to address before humans and AI agents can be truly complementary – with all parties benefiting from their collaboration. Especially a better understanding of how human-AI teams function and the main mechanisms that support collaboration, such as trust, is needed (Centeio Jorge et al., 2022).

Trust describes the willingness to rely on and be vulnerable to another party – a central prerequisite for effective collaboration between users and AI agents as well as between human team members (B. A. De Jong et al., 2016; Breuer et al., 2020; Schaefer et al., 2016; Sheridan, 2019). For instance, distrust, over-trust, or under-trust in human-AI teamwork can critically hinder team effectiveness (de Visser et al., 2020; Glikson & Woolley, 2020). In the human-AI nursing team, this could mean that the human nurses do not rely on the AI team member's medication suggestions, reject correct suggestion by the AI, or accept incorrect ones. Not trusting as well as overly relying on the AI team member can have critical effects on the patient's safety (Liao et al., 2015) and may negatively impact team decision-making. For instance, an AI team member may support doctors with predictions of likely hospital readmission based on correlated care decisions (e.g., Bayati et al., 2014; Caruana et al., 2015). Yet, whether or not the team members will trust each other and rely on this information will impact their overall team performance. At the same time, the AI team member may not always be reliable. In such cases, overtrust may lead to negative consequences such as faulty decision making.

Initial conceptualizations of team trust in human-AI teams have started to emerge (e.g., de Visser et al., 2020; Ulfert & Georganta, 2020), due to its central function for collaboration between humans and AI agents (Schaefer et al., 2016). Yet, research in this field is still scarce and faces significant limitations. First, current research on trust in human-AI teams tends to be limited in focus, usually only considering trust from a single team member. Specifically, both psychology and human-computer interaction (HCI) research mainly explore how a single human operator (trustor) builds trust towards an AI team member (trustee) and perceives the AI agent's trustworthiness.[2] Computer science literature also tends to focus on trust as a relation between different AI agents (e.g., in multi-agent systems) or on AI agent's trust towards humans (Azevedo-Sa et al., 2021). However, approaches are missing that combine these perspectives, acknowledging that both human and AI agents as team members may be enabled to evaluate each other's trustworthiness, making them both trustors and trustees.

Second, current approaches to trust in human-AI teams tend to focus only on the trust relationships of a single human towards an AI team member. This disregards the different trust relationships that exist in a team (e.g., human-human, human-AI, and AI-AI trust; de Visser et al., 2020; Huang et al., 2021). Research already suggests that both humans and AI team members make trust evaluations of their team members, and thus will be required to trust and to be trusted (Ulfert & Georganta, 2020). For example, in the human-AI nursing team, in an emergency situation after a night shift, it is critical for the AI team member to estimate which human team member can be trusted or not (e.g., estimating whether the human team member is physically or cognitively still able to perform their task). Similarly, a human team member needs to be aware when an AI team member can be trusted or when information provided by this team member may be not relevant, not useful, or even not reliable.

Third, the multilevel nature of human-AI teamwork (i.e., individual-, dyadic-, and team-level) has been neglected when exploring team trust in human-AI teams. As highlighted in the human team literature, the perceptions of all team members towards the different entities (individuals, dyads and team) need to be taken into consideration to understand team trust (Fulmer & Ostroff, 2021). For example, in the human-AI nursing team, a human nurse develops trust not only towards each individual team member (human and AI), but also towards its dyad within the team (human-human; human-AI; AI-AI) and towards the overall team.

Fourth, it remains unclear to what extent trust perceptions (i.e., trust beliefs) and relationships between human team members are indeed similar to those between humans and AI team members. For example, a nurse may perceive a doctor's trustworthiness differently from an AI team member's trustworthiness, simply because the first is a human and the second is not. At the same time, an AI team member may hold different beliefs about another AI team member's trustworthiness than about a human team member's. Fifth, our understanding of whether trust towards other entities (i.e., individual, dyad, team) is affected by how much the other team members trust these entities is limited. For example, if a human nurse does not trust an AI team member, this may also impact the extent to which a human doctor trusts the AI team member.

In an effort to comprehensively describe trust in human-AI teams and overcome the above limitations, we aim to move towards a multidisciplinary and multilevel conceptualization of team trust in human-AI teams considering the differences and commonalities in building trust towards human and AI team members. To do so, we develop a multidisciplinary framework of team trust in human-AI teams by integrating literature on psychological trust, teamwork, HCI, and computer science. Specifically, we recognize that in a human-AI team, AI agents do not only have the role of a technology but also the role of a team member. Furthermore, we consider all the aspects that shape team trust in human-AI teams: the human-AI team composition, the roles (trustee/trustor) of all team members (humans and AI agents), the different perceptions (beliefs) about each team member's trustworthiness, the dyadic relationships between team members (human-human, human-AI, AI-AI), and team trust. With our paper, we contribute to current literature and overcome the existing limitations by: (1) bringing organizational research, HCI, and computer science literature together in a framework that may be used and understood as an initial guideline by diverse disciplines; (2) overcoming existing construct inaccuracies by clearly defining and differentiating the different trust components from both the trustee and trustor, human and AI agent, as well as individual-, dyadic- and team-level; (3) providing an overview of what we know about human-AI (team) trust so far; (4) introducing a first multilevel-

Table 1. Articles included in the review.

Authors	Article Type	Team trust definition
Correia et al. (2018)	Experimental study	Team trust; perception of trust by the human regarding the team as a whole
de Visser et al. (2020)	Theoretical Paper	Interpersonal trust between team members impacts team interaction
McNeese et al. (2019)	Experimental Study	Team trust is "the extent to which a person is confident in, and willing to act on the basis of, the words, actions, and decisions of another" (McAllister, 1995, p. 25)
Ulfert and Georganta (2020)	Theoretical Paper	Team trust; shared perception (based on Salas et al., 2005)

multidisciplinary conceptual framework of team trust in human-AI teams to inspire future research and computational modelling.

We recognize the complexity and dynamics of team trust and trust relationships in human-AI teams. Therefore, we understand our framework as a first step towards understanding team trust in these new types of teams. Specifically, our aim is to identify the diverse trust beliefs that may exist in a human-AI team that will shape initial team trust. In research, our framework can initiate communication between disciplines and guide first empirical investigations of the way team trust in human-AI teams emerges. In practice, the framework can help identify targeted approaches for evaluating or improving trust in human-AI teams (e.g., by identifying incorrect trust beliefs), which can for example be used by HR professionals. For designers of human-AI teams (e.g., technology developers) the framework can help identify core factors to consider in the design of systems (e.g., displays of trustworthiness). For individuals responsible for implementing human-AI teams, understanding the complexities of team trust and factors important for its initial emergence, is crucial for successfully guiding implementation processes.

Team trust in human-AI teams

Current perspectives on team trust in human-AI teams

Team research highlights that for team trust to develop, all team members need to recognize the interests of other team members to engage in a joint endeavour (Costa et al., 2018; Webber, 2002). Thus, team trust incorporates both trust perceptions of individual team members and trust perceptions of the overall team, incorporating simultaneous individual- and team-level processes (Breuer et al., 2020; Fulmer & Ostroff, 2021). In human-AI teaming research, some initial work highlights these multiple perspectives when studying team trust. Specifically, researchers have pointed out that all team members (both humans and AI agents) develop trust towards each other and the team (Centeio Jorge et al., 2022; Huang et al., 2021; Ulfert & Georganta, 2020).

To gain an overview of research that specifically addresses team trust in human-AI teams, we conducted a scoping review (PRISMA; Page et al., 2021), combining both psychology and computer science databases (PsycInfo, PsychArticles, ACM, and Scopus databases). Specifically, we aimed to investigate whether prior research incorporates the suggestion of differentiating and considering trust towards the different entities that exist in a team (i.e., individual, dyad, team). As the primary search term, we used Human AND (Agent OR Robot[3] OR Artificial Intelligence OR Autonomy) AND Team AND Trust in the titles and abstracts.[4] We only included studies that (a) investigated teams of humans and AI technologies, such as AI agents (i.e., a team consisting of one or more humans and one or more AI agents), (b) investigated adult participants during human-AI collaboration working on a mutual task, and (c) measured trust (on an individual and/or team level), not an associated variable such as reliability or confidence. In the final step, we (d) only included studies that considered mutual or team trust in the human-AI team rather than trust from the human user towards the AI agent and vice versa.

The initial search resulted in 186 articles ($n = 43$ articles were published in the field of psychology, $n = 143$ were published in the field of computer science). Overall, only a very small number of sources fit the search criteria and addressed trust towards the team as a whole (including all team members). Results confirmed that current literature on team trust in human-AI teams is still scarce (see Table 1 for an overview of articles included in the final set of sources).

All sources included in the final review were published between 2018 and 2020, indicating that the topic is still relatively new. Only one source (Ulfert & Georganta, 2020) based its definition of team trust on organizational team trust literature (Salas et al., 2005; Webber, 2002). Other sources referred to computer science and interpersonal trust literature (de Visser et al., 2020; McNeese et al., 2019), or did not define team trust at all (Correia et al., 2018). The two experimental studies captured team trust by measuring how a single human team member perceived its interpersonal trust towards another team member (McNeese et al., 2019) or towards the team that it belonged (Correia et al., 2018).

Although trust in human-AI teams is a well-established concept (see e.g., Huang et al., 2021; Schelble et al., 2022), the results of the scoping review indicate that the sources lack a clear definition and conceptualization of team trust in human-AI teams that considers the different entities and levels involved. This further highlights the need for theory building and integration of the different disciplines and literature streams so that it can be applied across organizational, HCI, and computer science research.

Towards a multidisciplinary and multilevel framework of team trust in human-AI teams

As a starting point, we propose a theoretical framework that describes how humans and AI team members form trust relationships in human-AI teams (see Figure 3). At the same time, we differentiate between human and AI team members with

regard to trust perceptions and beliefs. In our theoretical framework, we synthesize existing theories of trust in human teams (e.g., Colquitt et al., 2007; Costa et al., 2018) and models of trust in AI agents (e.g., Centeio Jorge et al., 2021) to define team trust in human-AI teams. Specifically, we acknowledge team trust from a multidisciplinary perspective, integrating research from organizational psychology, HCI, and computer science. Thereby, we recognize similarities and differences between disciplines and highlight how we should shape the existing knowledge about trustor-trustee and the different trust relationships to this new team context. At the same time, we recognize human-AI teams as complex systems that incorporate different entities and levels, shaping team trust. The multidisciplinary and multilevel approach reflects a first step to defining and conceptualizing team trust in human-AI teams. This framework is specifically focused on initial team trust formation, as we recognize that trust can change over time (Grossman & Feitosa, 2018). Further, this contributes to a better understanding of how AI team members can be successfully integrated into social structures, such as organizational teams (de Visser et al., 2020).

Similar to other efforts in psychological literature (Kozlowski et al., 2013), as part of our framework, we utilize formalized descriptions of the different trustworthiness beliefs and trust relationships (e.g., TW(entity)). This is in line with recent calls for a more widespread use of formalized theory to study complex, dynamic systems (Salmon et al., 2022). Emergent states, such as team trust, are inherently complex and dynamic in nature. In order to fully understand how trust evolves in a human-AI team, it is essential to understand all underlying dynamics in detail (i.e., who is perceiving whose trustworthiness?). First, this clear description is beneficial, as it is a way of expressing theory that is understood by a variety of disciplines (e.g., psychology and computer science). Second, in an organizational context, it can help to better analyse potential problems and develop directed solutions. Lastly, a clear definition and the respective formalized descriptions can help in developing more suitable and acceptable AI agents as teammates, as the formalized theory may be integrated into the system.

Human-AI teams may strongly differ with regard to their team composition (Onnasch & Roesler, 2021). Therefore, and for reasons of simplification, when describing our framework and presenting our propositions, we will focus on a human-AI team that consists of two human (h) and two AI (a) team members (see Figure 1).

Team trust in human-AI teams describes a shared perception of trust towards the team as a whole, which includes all team members, both humans and AI agents. At the same time, team trust is shaped by the different trust relationships that exist within the human-AI team (de Laat, 2016; Fan et al., 2021; Huang et al., 2021; Ulfert & Georganta, 2020; van Wissen et al., 2012). Given the different trust relationships that exist at different levels within teams (Breuer et al., 2020; Costa et al., 2018; Feitosa et al., 2020; Fulmer & Ostroff, 2021), we suggest differentiating between three entities towards which trust can develop: (1) the individual, referring to each team member (human or AI agent), (2) the dyad, referring to all possible dyadic relationships (human-human, human-AI, AI-AI), and (3) the team as a whole. Specifically, we argue that trust in a human-AI team consists of trust towards individual team

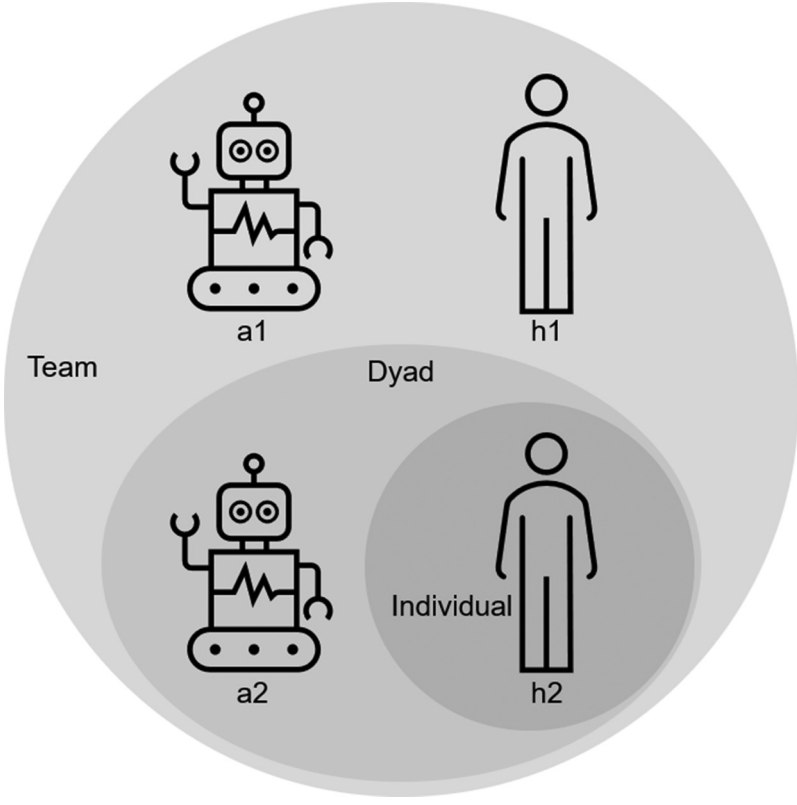

Figure 1. Exemplary human-AI team consisting of two AI agents (a) and two humans (h) and three types of entities (individual, dyad, and team).

members, trust towards dyads (i.e., interpersonal trust), and trust towards the team (i.e., team trust). Thus, we suggest the following:

Proposition 1: In a human-AI team, trust develops towards three entities: individual team members, dyads, and the team.

As previously noted, trust in human-AI teams, similar to trust in human teams, is not unidirectional (e.g., trust of the human only towards the AI team member and vice versa). Instead, trust is bidirectional so that a trust relationship between two individuals incorporates trust by each individual towards the other (Fulmer & Ostroff, 2021). Consequently, humans and AI team members act both as trustor and trustee, evaluating each other's trustworthiness (Azevedo-Sa et al., 2021). Building on this line of thinking, we argue that in human-AI teams, trust relationships are bidirectional and involve all team members, both humans and AI team members. Therefore, we propose the following:

Proposition 2: In a human-AI team, both human and AI team members act as trustors and trustees.

Trustworthiness, a central antecedent of trust in human-AI teams

Varied conceptualizations of trustworthiness exist in computer science and psychology (e.g., Castelfranchi & Falcone, 2000; Lee & See, 2004). Yet, the conceptualization of trustworthiness as described by Mayer's ABI Model (Mayer et al., 1995) has been equally utilized in both disciplines for humans and AI agents (see, e.g., Calhoun et al., 2019; Solberg et al., 2022; Toreini et al., 2020; Ulfert & Georganta, 2020). Specifically, the ABI model suggests that trustees have three characteristics that determine their trustworthiness – ability, integrity, and benevolence – and that the trustor forms evaluations about them (Breuer et al., 2020). Within the team setting, ability describes an individual team member's knowledge, competence, and interpersonal skills to collaborate effectively (Colquitt et al., 2007). Integrity describes a team member's consistency and whether it shares similar principles to the trustee or the overall team (Breuer et al., 2020; Fulmer & Gelfand, 2012). Benevolence reflects a team member's concern for the good of the team (Schoorman et al., 2007). A team member's level of trustworthiness – either a human's or AI agent's – impacts how trust towards that team member develops. Indeed, high trustworthiness comes with positive expectations about the team member's intentions, motivation, and behaviour, thus supporting collaboration within the team (Fulmer & Gelfand, 2012; Mayer et al., 1995). When collaborating with technologies, high trustworthiness has also been associated with positive expectations about the system's accuracy, capability, reliability, and functionality (Mcknight et al., 2011). For effective collaboration in a human-AI team, we argue that all team members (humans and AI agents) need to be able to evaluate each team member's trustworthiness. This argument builds on early work in the human-AI teaming literature suggesting that for AI agents to effectively interact with humans, they need to "be governed by the same principles that underlie human collaboration" (Rich & Sidner, 1997, p. 284). Thus, both human and AI team members require some degree of similarity in their capabilities for evaluating each other. In our team example (see Figure 1), this would mean that each human team member (e.g., h1) evaluates the trustworthiness (TW) of each human (h2) and of each AI team member (a1, a2). Similarly, each AI team member (e.g., a1) evaluates each human (h1, h2) and each AI (a2) team member's trustworthiness (TW). Consequently, each team member evaluates the trustworthiness of all other team members, while at the same time, all other team members evaluate its own trustworthiness. Thus, we propose that:

Proposition 3: In a human-AI team, both human and AI team members evaluate their team member's (human or AI agent) trustworthiness (TW) and are evaluated regarding their own trustworthiness.

Forming trust beliefs in human-AI teams

In a human-AI team, trust between two team members is shaped by the trustworthiness of each member – an objective property of the trustee (Castelfranchi & Falcone, 2000). In addition, trust between two team members is also shaped by how this trustworthiness is perceived by the other member – a subjective belief of the trustor. This so-called trust belief (e.g., how able I perceive a team member to be) can differ from the objective evaluation of a team member's trustworthiness (e.g., how able that team member indeed is), depending on the type of information available to the trustor (e.g., whether information about the trustee's abilities is available and whether the information is correct). At present, empirical evidence differentiating how humans and AI agents perceive the trustworthiness of others and how they form trust beliefs has limitations and it is unclear whether differences exist in how human and AI team members are perceived regarding the different trustworthiness dimensions. Further, research provides heterogenous findings on how trustworthiness of human and AI team members is perceived, suggesting under- and overestimation to be a problem (Langer et al., 2022). This further emphasizes the importance of differentiating subjective beliefs. Lastly, little is known about how humans and AI agents perceive trustworthiness based on others' behaviour (e.g., the nurse displaying trust or distrust behaviours towards another nurse).

We argue that both humans and AI team members perceive another human or AI team member's trustworthiness and consequently form trust beliefs. Specifically, in humans, the processes involved in perceiving someone's trustworthiness and forming trust beliefs are dynamic and impacted by a multitude of factors, such as context, individual differences, or even motivation (van der Werff et al., 2019). Further, humans are motivated by personal goals that make individuals foster certain relationships over others (van der Werff et al., 2019). The trust beliefs of human team members are also influenced by the

challenges experienced when evaluating and perceiving each other's trustworthiness. For instance, recent work suggests that humans may have a positively biased view of AI as more reliable, resulting in higher perceived ability (Schlicker et al., 2021).

When looking at trust beliefs in AI agents, they are formed based on the environment that the AI agents are embedded in. According to the Belief-Desire-Intention (BDI) architecture (Rao & Georgeff, 1995), a belief is the informative component of the system state, that is, an AI agent's perception about the world at that specific point in time (including itself and other entities involved), for example, "the human team member is able to perform its task".[5] These trust beliefs are also dynamic[6] (e.g., the trust belief regarding the human team member's ability can change to "my human team member is not able to perform their task") and can be updated, for instance, via active learning strategies (Zhong et al., 2015). For AI agents to be enabled to calculate trust beliefs, computer scientists often focus on aspects such as the characteristics of other entities (e.g., reputation; Sabater-Mir & Vercouter, 2013) and observed interactions (e.g., performance, helpfulness, Falcone et al., 2013). These beliefs determine the "degree of trust" towards another entity (e.g., a human team member) and shape the AI team members decision on whether or not to rely on that entity (Castelfranchi & Falcone, 2000). For example, the AI agent may choose whom to ask for help (Azevedo-Sa et al., 2021). However, AI agents often face problems in perceiving and evaluating a human's trustworthiness, or integrating trust expressed by humans and other AI agents (Azevedo-Sa et al., 2021). Furthermore, AI agents compared to humans require more coordination to communicate their behaviour and form trust beliefs (Castelfranchi, 1998). For an AI agent and consequently for an AI team member, trust is by definition an artificial construct (i.e., a set of variables e.g., numerical values, a probability distribution or a percentage; Falcone et al., 2003), namely a structure of artificial beliefs and goals (e.g., the agent having computed a "theory of mind" of the trustee). The computation of these trust beliefs and consequently the "degree of trust" of an AI team member is in the end a variable. Yet, to fulfil the same role that trust plays in human behaviour, AI agent literature typically builds on the human concept of trust.

Overall, we argue that both humans and AI team members form trust beliefs about another entity's trustworthiness ("the human team member is able to perform its task") and also beliefs about their trust towards that entity ("I trust the human team member"). Specifically, beliefs about trust towards a certain entity (individual, dyad, or team) are formed by each party involved in that entity, and those different beliefs reflect the overall trust relationship of that entity. For example, in a human-AI dyad, each team member perceives the trustworthiness of the other team member and the trust relationship of its dyad and thereby, forms the respective trust beliefs. These trust beliefs are formed by each team member of the dyad distinctively, and can thus differ. Together, they reflect the trust of the human-AI dyad. Overall, we propose the following:

Proposition 4: In a human-AI team, trust (T) towards another entity (individual, dyad, and team) consists of a belief (B) about an entity's trustworthiness (TW(entity)) and a belief (B) about the trust relationship towards that entity (TR(entity)).

Building a multidisciplinary framework of team trust in human-AI teams

Synthesizing work across disciplines, we have so far presented four propositions. First, given that trust is bidirectional (Azevedo-Sa et al., 2021; Fulmer & Ostroff, 2021), we argue that in a human-AI team, both humans and AI agents are trustors and trustees. Second, considering the multilevel nature of trust in human-AI teams, we suggest that within a human-AI team, trust is formed towards different entities (individual team member, dyad, and team). Third, in line with the ABI Model (Mayer et al., 1995) and conceptualizations of bidirectional trust (Fulmer & Ostroff, 2021), we argue that each team member evaluates the trustworthiness of every other team member (human and AI agent) and that its own trustworthiness is also being evaluated by everyone else. Fourth, utilizing the logic underlying the BDI architecture and conceptualizations of trust in AI agents and humans (e.g., Centeio Jorge et al., 2021, 2022; Rao & Georgeff, 1995), we argue that trust towards an entity is shaped by a belief about the entity's trustworthiness and by a belief about the trust relationships towards that respective entity.

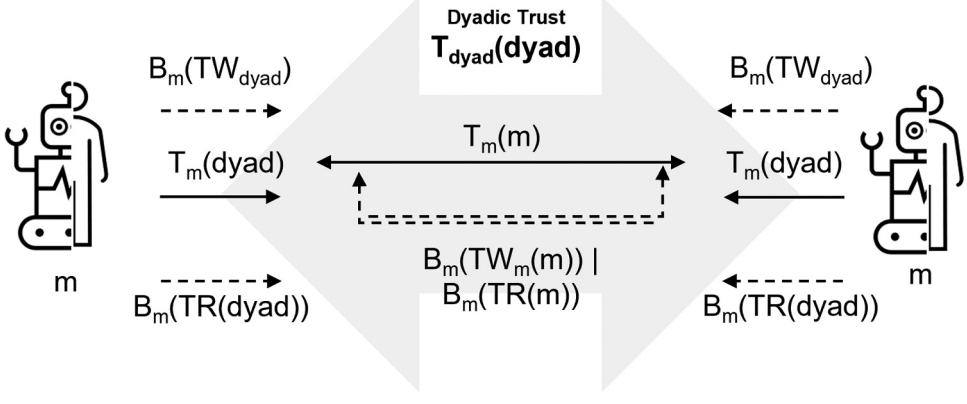

Figure 2. Trust towards two individual team members (m) and a dyad in human-agent teams.

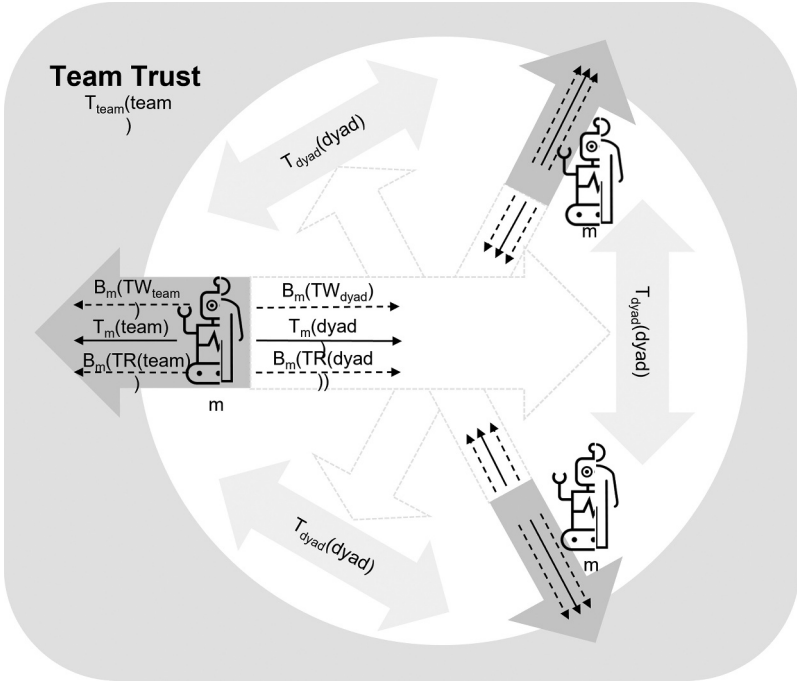

Figure 3. Multidisciplinary framework of team trust in human agent teams.

Building on these four propositions, Figure 2 illustrates trust towards an individual team member (human or AI agent), and trust towards a dyad (human-human, human-AI, or AI-AI). On the individual level, trust (T) consists of the belief formed through the evaluation of an individual team member (m) about the other team member's trustworthiness *($B_m(TW_m)$)* and of the belief formed about its trust relationship with this individual *(Bm(TR(m)))*. A trust belief reflects how an entity (i.e., individual, dyad, team) is subjectively perceived by another entity (e.g., another individual, dyad, or the whole team). This belief is important to build and promote appropriate trust between team members, particularly when the trustor is an AI agent (Centeio Jorge et al., 2021). Respectively, on the dyadic level, trust consists of the beliefs of both team members about the dyad's trustworthiness *($B_m(TW_{dyad})$* and their beliefs about the trust relationship of their dyad *($B_m(TR(dyad))$*.

To further specify the above propositions and trust towards all entities within a human-AI team, we further elaborate on interpersonal trust, that is the dyadic trust between team members. Lastly, we develop propositions with regard to team trust.

Dyadic trust between human team members

In human-AI teams, we assume that dyadic trust between two human team members develops the same way as in human-only teams. Specifically, dyadic trust between two humans is determined by the belief that each of the two human team members forms about the dyad's trustworthiness ($B_m(TW_{dyad})$. This evaluation is based on observed cues, such as behaviour and skills that reflect the dyad's reliability, integrity, and benevolence (Mayer et al., 1995). Furthermore, it is also shaped by how each one of them forms a belief about their dyadic trust relationship (*Bm(TR(dyad))*. Positive perceptions, beliefs, and expectations about the dyad's intentions, motivations, and behaviour lead to high levels of trust between the human team members (Fulmer & Gelfand, 2012). Sharing a common history and being similar to each other also creates positive beliefs and have been shown to support interpersonal relationships between human team members (Mayer et al., 1995). At the same time, the individual team members might be similar or asymmetric in their trust beliefs towards their dyad (Korsgaard et al., 2015), highlighting the importance of differentiating each team member's beliefs. Taken together, we propose the following:

Proposition 5: In a human-AI team, dyadic trust *($T_{dyad}(dyad)$)* between two human team members consists of the beliefs by each team member about their dyad's trustworthiness *($B_m(TW_{dyad})$* and of the beliefs of each team member about the trust relationships of their dyad *($B_m(TR(dyad))$*.

Dyadic trust between AI team members

As previously noted, conceptualizations of dyadic trust in AI agents are often derived from psychological literature, with many utilizing the ABI-model (Mayer et al., 1995) as a basis. Consequently, in a dyadic trust relationship between two AI team members, there will be similarities to dyadic trust between human team members (Smith & Hoffman, 2017). Specifically, beliefs about the dyad's trustworthiness (ability, integrity, and benevolence) are formed by both AI team members that shape their dyadic relationship. For this trustworthiness belief to form, similar to humans, AI agents need to be able to observe trustworthiness cues and evaluate them (Sabater-Mir & Vercouter, 2013). Specifically, AI team members

need to have information about the "ideal" level of the dyad's trustworthiness and compare this to the information they assess. To make this possible, human developers try to design AI team members with certain levels of functionality (ability), providing explanations for decision making in order to showcase integrity and benevolence. In this way, AI agents can use inferences or reasoning (i.e., an AI agent's judgement about a certain entity derived from rational reasoning, not involving direct experience with, recommendation of, or reputation about the entity; Falcone et al., 2013) to evaluate the dyad's trustworthiness. At the same time, AI team members form trust beliefs about the trust relationship of their dyad, which also impacts the dyadic trust between the two AI team members. To form such a belief, AI agents consider different sources (Falcone et al., 2013): direct experience (based on past interactions), recommendations (what particular team members say about the dyad), and reputation (the shared general opinion of the team regarding the dyad). Building on this previous work and our previous line of thinking, we propose the following:

Proposition 6: In a human-AI team, dyadic trust *($T_{dyad}(dyad)$)* between two AI team members consists of the beliefs formed by each team member about the dyad's trustworthiness ($B_m(TW_{dyad})$) and of the beliefs of each team member about the trust relationships of their dyad ($B_m(TR(dyad))$).

Dyadic trust between human and AI team members

For dyadic trust to develop between a human and an AI team member, both need to be able to form beliefs about the dyad's trustworthiness and about the trust relationship of their dyad. Simultaneously, all team members need to display how trustworthy they are in a way that the other team member can understand.

To form a belief through the evaluation of the ability, integrity, and benevolence of the human-AI dyad, human team members need to be able to perceive information and observe cues that reflect whether the human-AI dyad is trustworthy or not. Humans find it easier to estimate people's abilities than machines, even if they are trained to make better assessments (Nass & Moon, 2000). Therefore, "AI systems must be able to explain how and why they arrived at a particular conclusion so that a human can evaluate the system's rationale" (Banavar, 2016; para. 7). Providing such information will support human team members in understanding the characteristics of their AI team member and, in turn, how trustworthy their dyad is (Fulmer & Ostroff, 2021; Mayer et al., 1995). Consequently, this will shape how the human-AI trust relationship will develop. However, research on how AI agents can make their trustworthiness and displays of trust more understandable to humans is still emerging. Especially concerning benevolence, early findings suggest differences in how humans evaluate the trustworthiness of AI agents or humans (Hu et al., 2021).

The dyadic trust relationship between a human and an AI team member is further impacted by each team member's belief about the trust relationship of their dyad. With regards to the human team member, we argue that forming a belief about the human-AI trust relationship might differ from how beliefs are formed in human-human dyadic relationships. In line with the social categorization theory (Turner et al., 1987), previous research has suggested that differences in beliefs may depend on the degree of similarity between the human and the AI team member (Calhoun et al., 2019; Kuchenbrandt et al., 2013; Steain et al., 2019; Ulfert & Georganta, 2020). Specifically, the comparison of a human team member with other team members and with the rest of the team influences how its trust relationships are formed. When a team member is perceived as more similar to oneself and to the rest of the team, it allows to feel more comfortable with each other and, thus, interpersonal trust to develop (Turner et al., 1987). Yet, compared to a human, an AI team member will differ in appearance or communication and, as prior literature has suggested, an AI team member will be perceived as less similar compared to a human team member (Ulfert & Georganta, 2020). Specifically, depending on the AI agent's characteristics (e.g., anthropomorphism), an AI agent team member might be perceived as more (a team member) or less similar (a technology) to its human team member (Savela et al., 2021), influencing the human team member's belief about the trust relationship of the human-AI dyad.

For the AI team member to form trust beliefs about the human-AI dyad's trustworthiness and trust relationship, direct experience, recommendations, and reputation are used (Falcone et al., 2013), similar resources as for the AI-AI dyad. However, we also expect that the formation of trust beliefs will not be identical between the different types of dyads (human-AI or AI-AI). For instance, on the one hand, evaluating a human-AI dyad's ability may be similar to forming beliefs about an AI-AI dyad's ability, given that the AI team member can compare objective measures of performance for the AI agent and the human. On the other hand, forming beliefs about a human-AI dyad's benevolence may be more difficult for the AI agent, as it will require to understand affective reactions of the human team member, displayed through language or interaction behaviour. Future AI agents should be developed with capabilities to evaluate all dimensions of trustworthiness (ABI) in humans and thereby form beliefs about the trustworthiness and trust relationship of a human-agent dyad, something that requires further investigation.

Overall and in line with our previous propositions we propose the following:

Proposition 7: In a human-AI team, dyadic trust *($T_{dyad}(dyad)$)* between a human and an AI team member consists of the beliefs formed by each member about their dyad's trustworthiness ($B_m(TW_{dyad})$) and of the beliefs of each member about the trust relationships of their dyad ($B_m(TR(dyad))$).

Team trust in human-AI teams

In human teams, team trust describes the perceptual state that resides at the team level ("We trust one another") and the psychological state that resides within individuals of that entity ("I trust the team"; Fulmer & Ostroff, 2021). Similar to the individual and the dyadic level, we argue that team trust is shaped by beliefs about the team's trustworthiness. The

team's trustworthiness is composed of the team's overall level of ability, integrity, and benevolence (Breuer et al., 2020; Feitosa et al., 2020). In parallel, team trust is shaped by the beliefs about the trust relationships of the team as a whole as perceived by each team member (see Figure 3). Yet, neither clear definitions nor approaches exist that define how team trust can or should be best estimated (Feitosa et al., 2020). This becomes even more challenging when having diverse members, such as human and AI agents, because considerable differences may exist in how individual team members perceive team trustworthiness and the relationships within their team (B. de Jong et al., 2021). On the one hand, human team members may find it easier to perceive and estimate some trust factors over others (e.g., whether team members seem to act for the good of the team). On the other hand, AI team members may be better equipped to estimate other trust factors, such as a team's overall ability, due to a better and detailed understanding of the team task at a given point in time. Further, while team member's abilities may be additive, leading to higher team trust (Lim & Klein, 2006; Mathieu et al., 2008), differences in integrity or benevolence might have negative effects on team trust.

Additional factors may further impact the belief of each human and AI team member about the trust relationships of their team. As in human teams, the degree of similarity or asymmetry in team members' trust beliefs may impact how team trust is perceived (B. de Jong et al., 2021; Korsgaard et al., 2015). Specifically, human team members form their beliefs based on the degree of similarity across all team members and prior experiences in working with the team (Costa et al., 2018; Ulfert & Georganta, 2020). Relatedly, AI team members form their beliefs based on direct experience, recommendations, and reputation (Falcone et al., 2013). During early phases of interaction, where experience as a team is lacking, both human and AI team members will base their beliefs on initial interactions and preconceptions (de Visser et al., 2020). Building on this line of thinking and on the fact that there are currently neither computational models for implementing team trust estimation in AI agents nor appropriate frameworks representing shared perceptions, such as team trust (Breuer et al., 2020; Feitosa et al., 2020), in human-AI teams, we propose the following:

Proposition 8a: In a human-AI team, team trust *(T_{team}(team))* consists of the beliefs formed by each team member about the team's trustworthiness ($B_m(TW_{team})$ and of the beliefs of each member about the trust relationships of the team ($B_m(TR(team))$.

Team trust reflects an emergent state that evolves based on interpersonal interactions and group dynamics (Costa et al., 2018; Kiffin-Petersen, 2004; Shamir & Lapidot, 2003). Specifically, rich evidence suggests that trust at the individual and dyadic level as well as the cooperation among single team members impacts how collective team trust develops (Breuer et al., 2020; Feitosa et al., 2020). Thus, team trust is influenced by the trustworthiness and trust relationships of all entities and levels (all individuals and dyads), not only of the trust beliefs about the team as a whole. Consequently, we argue that all trust relationships that exist towards each individual and dyadic entity within a team also influence team trust. Specifically, we propose the following:

Proposition 8b: In a human-AI team, team trust *(T_{team}(team))* is impacted by trust towards the different entities within the team (i.e., individuals and dyads).

Thus far, we have described trust relationships in which team members are the trustor of the entity they evaluate. However, building on theories of team research (e.g., Cooke et al., 2013), team members evaluate and form beliefs not only about entities they belong to, but also about all the other entities that exist within the team (e.g., another dyad that the team member is not part of). These latter evaluations and beliefs also shape trust that develops within a team, especially trust towards the team as a whole (Costa et al., 2018). To be more specific, individual team members not only form a belief about their trust relationship with individual team members they directly interact, or with dyads that they are part of. Instead, they also form beliefs about dyads they are not part of. For example, an AI team member also forms a belief about a human-human dyad's trust relationship. In turn, this belief impacts how the AI team member forms beliefs about other trust relationships, such as the trust belief about its relationship with each of the two human team members as well as the trust belief towards the team as a whole. Thus, we propose the following:

Proposition 8c: In a human-AI team, team trust *(T_{team}(team))* is impacted by team members' trust towards entities they are not part.

Discussion

With the continuous development of highly capable AI systems, human-AI teaming is becoming increasingly prevalent in the workplace (Larson & DeChurch, 2020). In order to develop effective long-term collaboration in human-AI teams, developing team trust is essential. At the same time, in teams consisting of humans and AI team members, we need to consider that trust may develop differently compared to teams only consisting of humans. To develop a better understanding of team trust in human-AI teams, in the present paper, we reviewed and compared organizational, HCI, and computer science literature on human-AI team trust. A comparison of existing literature indicated that at this point, most studies understand trust in human-AI teams as trust of the human towards the AI agent as a technology, rather than a team member (e.g., McNeese et al., 2019). Further, definitions of team trust in the human-AI team context are lacking, with the multilevel nature of team trust, as described in organizational literature, often being disregarded. Building on initial empirical evidence and integrating computer science and organizational trust literature (e.g., Centeio Jorge et al., 2021; de Visser et al., 2020; Feitosa et al., 2020), we clarified that in human-AI teams, trust relationships develop towards three types of entities: individual team members, dyads (human-

human, AI-AI, human-AI), and the team as a whole. Specifically, we presented an initial multidisciplinary framework consisting of various propositions in an attempt to describe the complexity of team trust in human-AI teams.

With our framework, we advance current literature in both work and organizational psychology and computer science research in three ways. First, by reflecting on and integrating theories and research on trust in humans, trust in AI agents, and trust in work teams, discussing similarities and differences. Second, by highlighting the multilevel structure of team trust, explaining that trust can develop towards different entities (within- and cross-levels) and that individual as well as dyadic trust shape team trust. Third, by offering a first description of the complexity of team trust in human-AI teams that can act as a first steps towards understanding of human-AI teaming in organizations. These extensions of current psychological theories are essential as human-AI teams are increasingly being implemented at work. We hope that our framework can form an initial foundation for future empirical research and future developments in the field of human-AI teamwork. For instance, the framework can be used for detecting distrust in the team or untrustworthy team members at work. This is important as the framework highlights that distrust by one team member could negatively impact overall team trust. Further, our framework may inspire theory building in the domains of team formation, training, or AI-related organizational change processes.

Limitations and suggestions for future research

In order to advance the field of human-AI teamwork, there are still many additional challenges that need to be addressed, which are currently not considered in the proposed framework. Considering the diverse fields of literature that the framework is built on, further extensions building on organizational, HCI, and computer science theory and methods are needed. At the same time, many of the open questions that remain, also reflect central challenges in human team trust research.

First, while we recognize the impact of time on team trust development, this aspect was not included in the proposed framework. Team trust relationships are also shaped by the history that the human-AI team shares and the time working together. Our framework focused on initial trust formation and the diverse relationships that exist in human-AI teams. Yet, initial team trust in human-AI teams may strongly differ from team trust after some time of interaction (de Visser et al., 2020). Research highlights that it takes time for emergent constructs, such as team trust, to develop in order for team members to perceive it as a shared characteristic (Carter et al., 2018; Feitosa et al., 2020). Further, based on current literature it is still unclear how frequently humans update their perceptions about their team members or overall team's trustworthiness. In the development of an AI team member, this aspect would usually be considered, resulting in possible differences in how often human and AI team members update their trustworthiness beliefs. Thus, the proposed framework should be theoretically extended and empirically evaluated with regards to changes in trust beliefs and respectively in team trust over time.

Second, to study human-AI team trust, we will require new types of measures. When measuring human team trust, a common approach is to collect self-reports of individual team members that are later aggregated at the team level (e.g., Jarvenpaa et al., 1998). This approach has recently been criticized, as it may not represent the multilevel and dynamic nature of team trust and further often disregards different trust dimensions (Feitosa et al., 2020). Self-reports may further be susceptible to bias and may thus not correctly represent an individuals' beliefs (Krosnick, 1999; Podsakoff et al., 2003). At the same time, estimating latent constructs, such as team trust, may be more difficult for an AI team member compared to estimating observable behaviours. Further, it is still unclear how measures of trustworthiness in humans, mostly represented as a mean or sum score of rated items, and trustworthiness evaluations expressed by an AI team member, represented as a percentage, can be compared. This lack of comparable approaches also raises additional questions, for example, regarding the relative weight of each of these evaluations. New methodological approaches focusing more on behavioural indicators (e.g., sharing or not sharing information) seem to be necessary to quantify and measure team trust in human-AI teams, as similarly proposed for other multilevel constructs or for measuring trust in AI agents (Cole et al., 2011; Delice et al., 2019; Feitosa et al., 2020; Jorge et al., 2022).

Third, the presented framework does not focus on how characteristics of the human, the AI agent, or the context may impact team trust. Rich evidence suggest that individual differences, such as differences in an individual's competence beliefs or their propensity to trust other humans or technologies can impact team trust (Costa, 2003) as well as trust in technologies (Jessup et al., 2019; Mcknight et al., 2011; Schaefer et al., 2016). Meta-analyses on trust in AI agents or automated systems further suggest that characteristics of the system itself (e.g., embodiment, anthropomorphism, etc.) strongly impact trust (Kaplan et al., 2021; Schaefer et al., 2016). The findings further highlight the importance of contextual factors, such as task characteristics. Finally, organizational factors (e.g., organizational culture) can also impact team trust emergence (Fulmer & Gelfand, 2012). Future extensions of the proposed framework should include such individual-, system-, and context-level characteristics. This may assist in implementing human-AI teams into existing work processes but also foster the development of trustworthy AI agents and their ability to understand human behaviour.

Lastly, research highlights that team trust can be both cognitive and affective (Webber, 2008), aspects that currently not differentiated in the framework. Specifically, it has been suggested that it is easier for both humans and AI agents to develop cognitive trust towards each other than affective trust (Glikson & Woolley, 2020). Yet, current HCI research has thus far predominantly studied cognitive trust, often focusing on system capabilities, rather than addressing affective trust (Lynn et al., 2021). This one-sided approach to trust may be because affective trust develops over time (Webber, 2008). Yet,

most studies on human-AI collaboration focus on interaction during the implementation of the AI agent (Glikson & Woolley, 2020). Thus, to include the differentiation between cognitive and affective trust, we first need a better understanding of how cognitive and affective trust between humans and AI team members develops, considering the aspect of time as well.

Practical implications

Although the proposed framework predominantly contributes to theoretical and research efforts, it may also offer future implications for organizational practice. As human-AI teams will become more common in organizations, approaches will be needed to improve how these teams operate and how their members interact with each other. To build human-AI teams and guarantee long-term team effectiveness, we propose to focus on team trust. This perspective is relevant at several stages of implementing human-AI teams at work, but especially during initial interaction. With our framework, we would further like to make organizations aware of the complexity of trust in teams and the different trust relationships involved. This can help to improve the implementation of such teams and allow for developing suitable team training. Acknowledging the multilevel nature of team trust can support managers to differentiate between entities and identify where teams may need more support (e.g., training in correctly evaluating AI agent's ability, integrity, or benevolence). Distinguishing between different trust beliefs (about trustworthiness and trust relationships) can also help to gain an understanding why team members may differ in their perceptions and reactions towards the same team member (e.g., a doctor being unwilling to rely on an AI team member, while the nurse is willing to do so), and offer focused solutions to trust related challenges (e.g., providing training to the doctor that considers their specific role or tasks in the team). At the same time, having a team consisting of humans and AI agents may offer new opportunities for managing and developing teams. For AI agents to have an understanding about the relationship between human team members, they need to quantify this information. The fact that information on the different trust relationships and the team as a whole is going to be available to the AI agent in a quantified form, could enable the team to inform themselves about the trust levels within the team. This information could further be used to collectively reflect and discuss possible issues and solutions. Representing team trust in a formalized way (i.e., using formulas) can further serve as a bridge between different disciplines or departments, as it makes different beliefs explicit. For example, if human-AI team collaboration in an organization is dysfunctional due to a lack of trust, the formalized framework can help to pinpoint why this is the case and how the problem can be approached. Due to the formalized language, it becomes easier for AI developers to understand where and why the problem arise and thus, to react accordingly. It also allows for building better AI team members in the design phase, as the framework can be considered when, for instance, implementing agent behaviours.

Conclusion

Trust in teams is essential, especially when a team does not only consist of humans but also of AI agents. However, research thus far has missed to clearly define team trust in human-AI teams. We provide an overview and comparison of how current psychology and computer science literature have defined team trust in human-AI teams and further integrate the different research streams. In line with prior theoretical models of trust in teams, trust in technology, and trust in AI agents, we presented a multidisciplinary framework of team trust in human-AI teams. Our framework is intended to motivate researchers and practitioners to adopt multidisciplinary approaches for studying and implementing human-AI teams in organizations, considering the dynamic and multilevel nature of team trust.

Notes

1. In computer science, an environment refers to everything that surrounds the AI agent, but is not part of the agent itself (Russell & Norvig, 2016). Thus, for the AI agent, the environment includes other team members.
2. Trustworthiness is defined by the level of ability, integrity, and benevolence of a team member (Mayer et al., 1995). This definition has been previously applied to trustworthiness of humans as well as of agents (Langer et al., 2022).
3. Robotic systems that do not entail AI components (e.g., utilizing learning algorithms) were excluded based on exclusion criterium (a).
4. Our search was restricted to peer-reviewed publications between 2000 and 2021.
5. With BDI following the Bratman's theory of rational actions in humans (Bratman, 1987), AI agents can reason in abstract concepts (e.g., a belief) and also showcase their trustworthiness to other team members.
6. The detailed dynamics and formation of these beliefs have been discussed in computer science literature (see e.g., Bosse et al., 2007; Herzig et al., 2010) but are out of the scope of this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Anna-Sophie Ulfert Conceptualization, Methodology, Formalization, Visualization, Writing - Original Draft, Writing - Review & Editing

Eleni Georganta Conceptualization, Writing - Original Draft, Writing - Review & Editing

Carolina Centeio Jorge Conceptualization, Methodology, Formalization, Writing - Original Draft, Visualization

Siddarth Mehrotra: Conceptualization, Writing - Original Draft

Myrthe Tielman Conceptualization

ORCID

Anna-Sophie Ulfert http://orcid.org/0000-0001-6293-4173
Myrthe Tielman http://orcid.org/0000-0002-7826-5821

References

Azevedo-Sa, H., Yang, X. J., Robert, L. P., & Tilbury, D. (2021). A unified bi-directional model for natural and artificial trust in human–robot collaboration. *EEE Robotics and Automation Letters*, *3*(6), 5913–5920. https://doi.org/10.1109/LRA.2021.3088082

Banavar, G. (2016, November). What it will take for US to trust AI. *Harvard Business Review*. https://hbr.org/2016/11/what-it-will-take-for-us-to-trust-ai

Bayati, M., Braverman, M., Gillam, M., Mack, K. M., Ruiz, G., Smith, M. S., & Horvitz, E. (2014). Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *Plos One*, *9*(10), e109264. https://doi.org/10.1371/journal.pone.0109264

Bosse, T., Jonker, C. M., Treur, J., & Tykhonov, D. (2007). Formal analysis of trust dynamics in human and software agent experiments. In M. Klusch, K. V. Hindriks, M. P. Papazoglou, & L. Sterling (Eds.), *Cooperative information agents XI* (pp. 343–359). Springer. https://doi.org/10.1007/978-3-540-75119-9_24

Bratman, M. (1987). *Intention, plans, and practical reason*. https://doi.org/10.2307/2185304

Breuer, C., Hüffmeier, J., Hibben, F., & Hertel, G. (2020). Trust in teams: A taxonomy of perceived trustworthiness factors and risk-taking behaviors in face-to-face and virtual teams. *Human Relations*, *73*(1), 3–34. https://doi.org/10.1177/0018726718818721

Calhoun, C. S., Bobko, P., Gallimore, J. J., & Lyons, J. B. (2019). Linking precursors of interpersonal trust to human-automation trust: An expanded typology and exploratory experiment. *Journal of Trust Research*, *9*(1), 28–46. https://doi.org/10.1080/21515581.2019.1579730

Carter, N. T., Carter, D. R., & DeChurch, L. A. (2018). Implications of observability for the theory and measurement of emergent team phenomena. *Journal of Management*, *44*(4), 1398–1425. https://doi.org/10.1177/0149206315609402

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. https://doi.org/10.1145/2783258.2788613

Castelfranchi, C. (1998). Modelling social action for AI agents. *Artificial Intelligence*, *103*(1–2), 157–182. https://doi.org/10.1016/S0004-3702(98)00056-3

Castelfranchi, C., & Falcone, R. (2000). Trust is much more than subjective probability: Mental components and sources of trust. *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, vol.*1*, 10. https://doi.org/10.1109/HICSS.2000.926815

Centeio Jorge, C., Mehrotra, S., Tielman, M., & Jonker, C. M. (2021). Trust should correspond to trustworthiness: A formalization of appropriate mutual trust in human-agent teams. *22nd International Trust Workshop Co-Located with AAMAS 2021*. CEUR Workshop Proceedings.

Centeio Jorge, C., Tielman, M. L., & Jonker, C. M. (2022). Artificial trust as a tool in human-AI teams. *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 1155–1157). IEEE.

Cole, M. S., Bedeian, A. G., Hirschfeld, R. R., & Vogel, B. (2011). Dispersion-composition models in multilevel research: A data-analytic framework. *Organizational Research Methods*, *14*(4), 718–734. https://doi.org/10.1177/1094428110389078

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *The Journal of Applied Psychology*, *92*(4), 909–927. https://doi.org/10.1037/0021-9010.92.4.909

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive Science*, *37*(2), 255–285. https://doi.org/10.1111/cogs.12009

Correia, F., Mascarenhas, S., Prada, R., Melo, F. S., & Paiva, A. (2018). Group-based emotions in teams of humans and robots. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 261–269. https://doi.org/10.1145/3171221.3171252

Costa, A. C. (2003). Work team trust and effectiveness. *Personnel Review*, *32*(5), 605–622. https://doi.org/10.1108/00483480310488360

Costa, A. C., Fulmer, C. A., & Anderson, N. R. (2018). Trust in work teams: An integrative review, multilevel model, and future directions. *Journal of Organizational Behavior*, *39*(2), 169–184. https://doi.org/10.1002/job.2213

Cummings, P., Mullins, R., Moquete, M., & Schurr, N. (2021). *Hello world! I am charlie, an artificially intelligent conference panelist*. 380.

De Jong, B. A., Dirks, K. T., & Gillespie, N. (2016). Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *The Journal of Applied Psychology*, *101*(8), 1134–1150. https://doi.org/10.1037/apl0000110

de Jong, B., Gillespie, N., Williamson, I., & Gill, C. (2021). Trust consensus within culturally diverse teams: A multistudy investigation. *Journal of Management*, *47*(8), 2135–2168. https://doi.org/10.1177/0149206320943658

de Laat, P. B. (2016). Trusting the (Ro)botic other: By assumption? *SIGCAS Computers and Society*, *45*(3), 255–260. https://doi.org/10.1145/2874239.2874275

Delice, F., Rousseau, M., & Feitosa, J. (2019). Advancing teams research: What, when, and how to measure team dynamics over time. *Frontiers in Psychology*, *10*, 1324. https://doi.org/10.3389/fpsyg.2019.01324

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

Falcone, R., Pezzulo, G., & Castelfranchi, C. (2003). A fuzzy approach to a belief-based trust computation. In R. Falcone, S. Barber, L. Korba, & M. Singh (Eds.), *Trust, reputation, and security: Theories and practice* (pp. 73–86). Springer. https://doi.org/10.1007/3-540-36609-1_7

Falcone, R., Piunti, M., Venanzi, M., & Castelfranchi, C. (2013). From manifesta to krypta: The relevance of categories for trusting others. *ACM Transactions on Intelligent Systems and Technology*, *4*(2), 1–24. https://doi.org/10.1145/2438653.2438662

Fan, X., Liu, L., Zhang, R., Jing, Q., & Bi, J. (2021). Decentralized trust management: Risk analysis and trust aggregation. *ACM Computing Surveys*, *53*(1), 1–33. https://doi.org/10.1145/3362168

Feitosa, J., Grossman, R., Kramer, W. S., & Salas, E. (2020). Measuring team trust: A critical and meta-analytical review. *Journal of Organizational Behavior*, *41*(5), 479–501. https://doi.org/10.1002/job.2436

Fulmer, C. A., & Gelfand, M. J. (2012). At what level (and in whom) we trust: Trust across multiple organizational levels. *Journal of Management*, *38*(4), 1167–1230. https://doi.org/10.1177/0149206312439327

Fulmer, C. A., & Ostroff, C. (2021). Trust conceptualizations across levels of analysis. In N. Gillespie, A. C. Fulmer, & R. J. Lewicki (Eds.), *Understanding trust in organizations* (1st ed., pp. 14–41). Routledge.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *The Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Grossman, R., & Feitosa, J. (2018). Team trust over time: Modeling reciprocal and contextual influences in action teams. *Human Resource Management Review*, *28*(4), 395–410. https://doi.org/10.1016/j.hrmr.2017.03.006

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, *105*, 105–120. https://doi.org/10.1016/j.techfore.2015.12.014

Herzig, A., Lorini, E., Hubner, J. F., & Vercouter, L. (2010). A logic of trust and reputation. *Logic Journal of IGPL*, *18*(1), 214–244. https://doi.org/10.1093/jigpal/jzp077

Huang, L., Cooke, N. J., Gutzwiller, R. S., Berman, S., Chiou, E. K., Demir, M., & Zhang, W. (2021). Distributed dynamic team trust in human, artificial intelligence, and robot teaming. In C. S. Nam, E. P. Fitts, & J. B. Lyons (Eds.), *Trust in human-robot interaction* (pp. 301–319). Elsevier.

Hu, P., Lu, Y., & Gong, Y. (. (2021). Dual humanness and trust in conversational AI: A person-centered approach. *Computers in Human Behavior*, *119*, 106727. https://doi.org/10.1016/j.chb.2021.106727

Jarvenpaa, S. L., Knoll, K., & Leidner, D. E. (1998). Is anybody out there? Antecedents of trust in global virtual teams. *Journal of Management Information Systems*, *14*(4), 29–64. https://doi.org/10.1080/07421222.1998.11518185

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In J. Y. C. Chen & G. Fragomeni (Eds.), *Virtual, augmented and mixed reality. Applications and case studies* (pp. 476–489). Springer International Publishing. https://doi.org/10.1007/978-3-030-21565-1_32

Jorge, C. C., Tielman, M. L., & Jonker, C. M. (2022). Assessing artificial trust in human-agent teams: A conceptual model. *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, 1–3. https://doi.org/10.1145/3514197.3549696

Kaplan, A., & Haenlein, M. (2020). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons*, *63*(1), 37–50. https://doi.org/10.1016/j.bushor.2019.09.003

Kaplan, A., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2021). Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, *65*(2), 00187208211013988. https://doi.org/10.1177/00187208211013988

Kiffin-Petersen, S. (2004). Trust: A neglected variable in team effectiveness research. *Journal of Management & Organization*, *10*(1), 38–53. https://doi.org/10.1017/S1833367200004600

Korsgaard, M. A., Brower, H. H., & Lester, S. W. (2015). It isn't always mutual: A critical review of dyadic trust. *Journal of Management*, *41*(1), 47–70. https://doi.org/10.1177/0149206314547521

Kozlowski, S. W. J., Chao, G. T., Grand, J. A., Braun, M. T., & Kuljanin, G. (2013). Advancing multilevel research design: Capturing the dynamics of emergence. *Organizational Research Methods*, *16*(4), 581–615. https://doi.org/10.1177/1094428113493119

Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, *7*(3), 77–124. https://doi.org/10.1111/j.1529-1006.2006.00030.x

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*(1), 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Kuchenbrandt, D., Eyssel, F., Bobinger, S., & Neufeld, M. (2013). When a robot's group membership matters. *International Journal of Social Robotics*, *5*(3), 409–417. https://doi.org/10.1007/s12369-013-0197-8

Langer, M., König, C. J., Back, C., & Hemsing, V. (2022). Trust in artificial intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. *Journal of Business and Psychology*, 1–16. https://doi.org/10.1007/s10869-022-09829-9

Larson, L., & DeChurch, L. (2020). Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. *The Leadership Quarterly*, *31*(1), 1–18. https://doi.org/10.1016/j.leaqua.2019.101377

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Leicht Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2019). The challenges of algorithm-based HR decision-making for personal integrity. *Journal of Business Ethics*, *160*(2), 377–392. https://doi.org/10.1007/s10551-019-04204-w

Liao, P. -H., Hsu, P. -T., Chu, W., & Chu, W. -C. (2015). Applying artificial intelligence technology to support decision-making in nursing: A case study in Taiwan. *Health Informatics Journal*, *21*(2), 137–148. https://doi.org/10.1177/1460458213509806

Lim, B., & Klein, K. J. (2006). Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *27*(4), 403–418. https://doi.org/10.1002/job.387

Lynn, T., van der Werff, L., & Fox, G. (2021). Understanding trust and cloud computing: An integrated framework for assurance and accountability in the cloud. In T. Lynn, J. G. Mooney, L. van der Werff, & G. Fox (Eds.), *Data privacy and trust in cloud computing: Building trust in the cloud through assurance and accountability* (pp. 1–20). Springer International Publishing. https://doi.org/10.1007/978-3-030-54660-1_1

Mathieu, J. E., Maynard, M. T., Rapp, T., & Gilson, L. (2008). Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future. *Journal of Management*, *34*(3), 410–476. https://doi.org/10.1177/0149206308316061

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, *38*(1), 24–59. https://doi.org/10.2307/256727

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS)*, *2*(2), 12. https://doi.org/10.1145/1985347.1985353

McNeese, N. J., Demir, M., Chiou, E., Cooke, N., & Yanikian, G. (2019). *Understanding the role of trust in human-autonomy teaming*. *Proceedings of the 52nd Hawaii international conference on system sciences*. IEEE.

Mirowska, A. (2020). AI evaluation in selection: Effects on application and pursuit intentions. *Journal of Personnel Psychology*, *19*(3), 142–149. https://doi.org/10.1027/1866-5888/a000258

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *The Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors*, *64*(5), 904–938. https://doi.org/10.1177/0018720820960865

Onnasch, L., & Roesler, E. (2021). A taxonomy to structure and analyze human–robot interaction. *International Journal of Social Robotics*, *13*(4), 833–849. https://doi.org/10.1007/s12369-020-00666-5

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *The BMJ*, *372*, n160. https://doi.org/10.1136/bmj.n160

Podsakoff, P. M., MacKenzie, S. B., Lee, J. -Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *The Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Rao, A. S., & Georgeff, M. P. (1995). BDI agents: From theory to practice. *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)* (pp. 312–319). ICMAS.

Rich, C., & Sidner, C. L. (1997). COLLAGEN: When agents collaborate with people. *Proceedings of the First International Conference on Autonomous Agents - AGENTS '97*, 284–291. https://doi.org/10.1145/267658.267730

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach* (Third edition, Global ed.). Pearson.

Sabater-Mir, J., & Vercouter, L. (2013). Trust and reputation in multiagent systems. In G. Weiss (Ed.), *Multiagent systems* (pp. 381–419). MIT Press.

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a "big five" in teamwork? *Small Group Research*, *36*(5), 555–599. https://doi.org/10.1177/1046496405277134

Salmon, P. M., Read, G. J. M., Walker, G. H., Stevens, N. J., Hulme, A., McLean, S., & Stanton, N. A. (2022). Methodological issues in systems human factors and ergonomics: Perspectives on the research–practice gap, reliability and validity, and prediction. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *32*(1), 6–19. https://doi.org/10.1002/hfm.20873

Savela, N., Kaakinen, M., Ellonen, N., & Oksanen, A. (2021). Sharing a work team with robots: The negative effect of robot co-workers on in-group identification with the work team. *Computers in Human Behavior*, *115*, 106585. https://doi.org/10.1016/j.chb.2020.106585

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., & Freeman, G. (2022). Towards ethical AI: Empirically investigating dimensions of AI ethics, trust repair, and performance in human-AI teaming. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 001872082211169. https://doi.org/10.1177/00187208221116952

Schlicker, N., Langer, M., Ötting, S. K., Baum, K., König, C. J., & Wallach, D. (2021). What to expect from opening up 'black boxes'? Comparing perceptions of justice between human and automated agents. *Computers in Human Behavior*, *122*, 106837. https://doi.org/10.1016/j.chb.2021.106837

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). *An integrative model of organizational trust: Past, present, and future*. Academy of Management Briarcliff Manor.

Seeber, I., Waizenegger, L., Seidel, S., Morana, S., Benbasat, I., & Lowry, P. B. (2020). Collaborating with technology-based autonomous agents: Issues and research opportunities. *Internet Research*, *30*(1), 1–18. https://doi.org/10.1108/INTR-12-2019-0503

Shamir, B., & Lapidot, Y. (2003). Trust in organizational superiors: Systemic and collective considerations. *Organization Studies*, *24*(3), 463–491. https://doi.org/10.1177/0170840603024003912

Sheridan, T. B. (2019). Extending three existing models to analysis of trust in automation: Signal detection, statistical parameter estimation, and model-based control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *61*(7), 1162–1170. https://doi.org/10.1177/0018720819829951

Smith, P. J., & Hoffman, R. R. (Eds.). (2017). *Cognitive systems engineering: The future for a changing world* (1st ed.). CRC Press. https://doi.org/10.1201/9781315572529

Solberg, E., Kaarstad, M., Eitrheim, M. H. R., Bisio, R., Reegård, K., & Bloch, M. (2022). A conceptual model of trust, perceived risk, and reliance on AI decision aids. *Group & Organization Management*, *47*(2), 187–222. https://doi.org/10.1177/10596011221081238

Steain, A., Stanton, C. J., & Stevens, C. J. (2019). The black sheep effect: The case of the deviant ingroup robot. *Plos One*, *14*(10), e0222975. https://doi.org/10.1371/journal.pone.0222975

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 272–283. https://doi.org/10.1145/3351095.3372834

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Basil Blackwell.

Ulfert, A. -S., & Georganta, E. (2020). A model of team trust in human-agent teams. *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 171–176. https://doi.org/10.1145/3395035.3425959

van den Bosch, K., Schoonderwoerd, T., Blankendaal, R., & Neerincx, M. (2019). Six challenges for human-AI Co-learning. In R. Sottilare, & J. Schwarz (Eds.), *HCII 2019. Lecture Notes in Computer Science* (Vol. 11597, pp. 572–589). Springer. https://doi.org/10.1007/978-3-030-22341-0_45

van der Werff, L., Legood, A., Buckley, F., Weibel, A., & de Cremer, D. (2019). Trust motivation: The self-regulatory processes underlying trust decisions. *Organizational Psychology Review*, *9*(2–3), 99–123. https://doi.org/10.1177/2041386619873616

van Wissen, A., Gal, Y., Kamphorst, B. A., & Dignum, M. V. (2012). Human-agent teamwork in dynamic environments. *Computers in Human Behavior*, *28*(1), 23–33. https://doi.org/10.1016/j.chb.2011.08.006

Webber, S. S. (2002). Leadership and trust facilitating cross-functional team success. *Journal of Management Development*, *21*(3), 201–214. https://doi.org/10.1108/02621710210420273

Webber, S. S. (2008). Development of cognitive and affective trust in teams: A longitudinal study. *Small Group Research*, *39*(6), 746–769. https://doi.org/10.1177/1046496408323569

Zhong, Y., Bhargava, B., Lu, Y., & Angin, P. (2015). A computational dynamic trust model for user authorization. *IEEE Transactions on Dependable and Secure Computing*, *12*(1), 1–15. https://doi.org/10.1109/TDSC.2014.2309126