Multidisciplinary perspectives on human-AI team trust

Myrthe L. Tielman[1], Morgan Bailey[2], Francesco Frattolillo[3], Carolina Centeio Jorge[1], Anna-Sophie Ulfert[4] and André Meyer-Vitali[5]

[1] Delft University of Technology | [2] University of Glasgow | [3] Sapienza University of Rome | [4] Eindhoven University of Technology | [5] DFKI

Human-AI teamwork is no longer a topic of the future. Given the importance of trust in human teams, the question arises how trust functions in human-AI teams. Although trust has long been studied from a human-centred perspective (e.g. in psychology and philosophy), a computational perspective and from the perspective of human trust in AI (e.g. in human-computer interaction), the study of trust in human-AI interaction in a team setting is still a novel field. For this reason, the MULTITTRUST (Multidisciplinary perspectives on Human-AI Team Trust) workshop series was founded. In this paper, we present the main outcomes after three editions. Our contributions are: an overview of the shared language of concepts and definitions; an outline of the main open research challenges; and methodological guidelines for further studies in meaningful human-AI team trust. These three contributions form a foundational roadmap towards a better understanding of trust in human-AI team interactions.

Keywords: trust, human-AI teamwork, research challenges, trustworthy AI, human-AI collaboration, team trust

https://doi.org/10.1075/is.24048.tie
Interaction Studies 26:2 (2025), pp. 164–199. ISSN 1572-0373 | E-ISSN 1572-0381

1. Introduction

In the years 2023 and 2024, three editions of the multidisciplinary workshop on trust in hybrid human-AI[1] teams[2] took place (MULTITTRUST) (Jorge and Ulfert-Blank, 2023, Brandizzi et al., 2023, Tielman et al., 2024). In the course of the various keynote lectures, presentations and discussions, several converging concepts and definitions, as well as multiple research challenges for this field emerged. In this paper, we present the foundational concepts to the field, the main research challenges which were identified, and guidelines on methodology to start addressing these, which all arose from the discussions held during the workshops.

The study of trust dynamics in Human-AI Teams (HAT) is an increasingly relevant area of research as human-AI collaboration becomes more prevalent across industries such as healthcare (Huber et al., 2025), logistics (Kahr et al., 2025), and safety (Seraj and Gombolay, 2020). Human-AI teamwork is no longer a topic of the future, but a pressing issue in today's technological landscape (Brynjolfsson and Mitchell, 2017, Zhang et al., 2022). The types of AI systems in this context is diverse, from decision support algorithms to physical robots (Duan et al., 2025). As these teams evolve, addressing the unique challenges of trust between human and artificial agents becomes critical for ensuring effective collaboration and decision-making (Visser et al., 2020, Glikson and Woolley, 2020, Kaur et al., 2023). Trust has been a well-established area of inquiry across multiple disciplines, including human-computer interaction (HCI) (Bach et al., 2024, Hoff and Bashir, 2015, Riegelsberger et al., 2005), and Philosophy (Baier, 1986, Pouryousefi and Tallant, 2023). Many slightly different definitions for trust exist, which (Lee and See, 2004) distilled into *"the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability"*.

In human teams, trust is fundamental for cooperative behaviour, as it underpins decisionmaking and the willingness of team members to rely on one another (Costa et al., 2018, Sapp et al., 2019). This same concept has been used within multi-agent systems (MAS),[3] where agents use mechanisms inspired by human

---

1. In this work, we will use the term "AI" to refer to artificially intelligent agents, which represent agency as the behaviour of and interaction among independent and concurrent AI systems. We will sometimes use the terms "AI" and (artificial) "agent" interchangeably.

2. We follow (Rix, 2022, Berretta et al., 2023) in their conceptualisation of a human-AI team as having at least one human and one AI agent working towards some shared goal, which both have unique roles/functions and some form of interdependence.

3. With the term 'multi-agent system', we specifically refer to a system with multiple artificial agents, but no humans in the loop

trust to assess and make decisions about the reliability of other agents (Burnett et al., 2011, Ramchurn et al., 2004). Additionally, recent research has shifted towards understanding how humans trust AI systems and, conversely, how these systems can be designed to be trustworthy (Kaur et al., 2023, Li et al., 2023). However, there is a gap in in understanding how trust dynamics function within HATs, where human and AI agents interact in complex, recurrent, and diverse team settings (Rix, 2022). Work on trust in teams exists both from a purely human perspective (e.g. Costa et al., 2018, Sapp et al., 2019), which doesn't account for the presence of a non-human agent; or from a AI agent perspective (e.g. Burnett et al., 2011, Ramchurn et al., 2004), without a human in the loop. Many studies into trust in teamwork do include both a human and artificial agent (e.g. Lee and See, 2004, Duan et al., 2025, Campagna and Rehm, 2025), however, these typically focus purely on human trust in AI systems, which doesn't fully account for the mutual nature of trust necessary in a teamwork setting (Ulfert et al., 2024). We additionally identify that these three different research perspectives originate in distinct research fields (organisational psychology, multi-agent systems and human-computer interaction respectively) which often don't communicate with each-other. We aregue that an integrative approach could be highly beneficial, given that our focus lies on HAT trust: the mutual trust relationships and dynamics between a least one human and at least one artificial agent who are working together in a team towards a shared goal.

In summary, we identify a lack of integrative approaches that bridge the theoretical and practical insights on HAT trust from various disciplines. While some reviews related to human-AI team trust exist, these are often focused exclusively on human trust in AI rather than all agents in the team (e.g. Campagna and Rehm, 2025, Zerilli et al., 2022), focused on specific types of AI (e.g. robots) only (e.g. Campagna and Rehm, 2025), focused on trustworthiness rather than trust (e.g. Ramchurn et al., 2021), or focused on factors in teamwork, but more than just trust (e.g. Schmutz et al., 2024, Kolomaznik et al., 2024, Kumar et al., 2021). Moreover, their focus typically lies on understanding current literature rather than on identifying research challenges. Developing comprehensive models of trust in HATs is essential to inform both the design and deployment of these systems, ensuring their effectiveness and reliability across real-world applications. Therefore, this paper presents a research agenda highlighting central challenges for research on trust in human-AI teams based on the multi-disciplinary perspectives gained through MULTITTRUST workshops.

The MULTITTRUST workshop series arose from this need to create a multidisciplinary research community focused on studying the different perspectives and layers of trust dynamics in human-AI teams. Therefore, one of the primary objectives of the workshop was to ignite constructive discussions and debates

among participants from different backgrounds to leverage diverse information and insights. Contributions and presentations from students and researchers with diverse disciplinary backgrounds, such as Computer Science, Artificial Intelligence, Sociology, Philosophy, Psychology, and Human-Computer Interaction, were welcomed. During the workshops, a combination of lightning talks, keynotes, and interactive discussion sessions were organized. More details on the organisation and structure of the workshops can be found in the workshop proceedings (Jorge and Ulfert-Blank, 2023, Brandizzi et al., 2023, Tielman et al., 2024).

The interdisciplinary setting of the workshops allowed for productive discussions on the topic, but also necessitated a strong focus on shared language. Initial discussions highlighted this need via recurring discussions around definitions and meanings of concepts like *trust*. Through the course of three workshops, we moved closer to an understanding of the concepts and definitions involved, and what they mean for the different disciplines. Therefore, our first contribution in this paper is an overview of these concepts.

During the workshops, the focus lay on interactive sessions. This process involved dividing the participants into several groups, around different discussion topics, as further detailed in (Jorge and Ulfert-Blank, 2023, Brandizzi et al., 2023, Tielman et al., 2024). Aside from discussions on concepts and definitions, research challenges for human-AI team trust were identified. After the group discussions, plenary discussions helped to clarify these open challenges. The second contribution of this paper is to present these core challenges in human-AI team trust research.

Finally, the third edition of the workshop included a discussion group on the important elements of meaningful human-AI teamwork studies. To start addressing the important challenges in the field, we do not only need to understand the concepts involved, but also the teamwork settings in which trust becomes relevant. Therefore, the final contribution of this paper is a discussion on the important elements of meaningful studies on human-AI team trust.

After the 2024 workshop, notes from all discussion groups were saved, which served as a foundation for this paper. All discussion groups had at least one author present, who contributed to the related sections in this paper. Additionally, organizers involved with the first and second edition of the workshop were invited to contribute to fill potential gaps in perspectives from those editions.

In the remainder of this paper, we will address the three main contributions listed above. In Section 2 we will start with an overview of the important concepts and definitions in the field of human-AI team trust, and the insights gained from the workshop discussions. Next, we will present the main open challenges in the

field in Section 3. Finally, in Section 4 we will present the important components of meaningful human-AI team trust studies aimed at addressing the challenges.

2. Concepts and definitions

The need for convergence on key concepts and their definitions quickly became clear from the workshop discussions. Given the interdisciplinary nature of the topic of human-AI team trust, people with diverse backgrounds are actively involved. This is crucial for the field, but also creates a challenge around shared language. Even the core concept of *trust* is perceived very differently by a psychologist, philosopher, computer scientist and others. The goal of the discussions was not convergence to a single shared definition, but rather a shared understanding of the different nuances and perspectives. In this section, we present this understanding on the key concepts which arose from the discussions namely: Trust, Trustworthiness, Trust Calibration, Artificial Trust and Team Trust. In Figure 1 we represent these concepts and how they relate to each-other. Broadly, we identify three possible trustor and trustees: a human, an agent, or a team. For each, we have their trust as a trustor, and their trustworthiness as a trustee. Trustworthiness of a trustee influences trust of the trustor, and when human (natural) trust in an agent and agent trustworthiness are aligned, we talk about trust calibration.

2.1 Trust

We start with trust where a human is the trustor, or *natural trust*, as this is present in all research communities and domains. We can contrast this with trust where an artificial agent is a trustor, which we will call *artificial trust*, and return to in Section 3.2.

Trust as a concept has been evading a single shared definition for decades (McKnight and Chervany, 1996). However, there are some common factors that we want to outline. Firstly, trust exists between two entities, namely one that trusts another one (trustor) and the one that is being trusted (trustee) (Mayer et al., 1995). Naturally, trust may exist between two human beings, but also between a human trustor and a synthetic or artificial trusted entity (software agent or robot). The entity involved is often an individual, but can also be a social entity consisting of multiple agents itself such as an institution or a team. Secondly, in order for trust to even be considered, there needs to be a certain risk that the trustor perceives from the behaviour of the trustee. The trustor should be vulnerable to the trustee's decisions and actions (Jacovi et al., 2021, Duarte et al., 2023, Rousseau et al., 1998, Mayer et al., 1995). Related to this, trust is argued to only be relevant

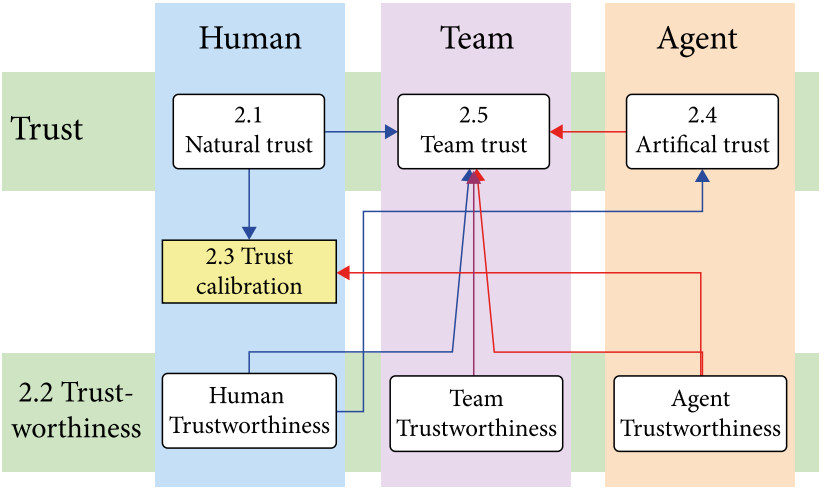

Figure 1. A schematic representation of the core concepts described in this section. We identify three possible trustor and trustees: A human (blue), an agent (red), or a team (purple). For each, we have their trust as a trustor, and their trustworthiness as a trustee. Trustworthiness of a trustee influences trust of the trustor, and when human (natural) trust in an agent and agent trustworthiness are aligned, we talk about trust calibration. The numbers represent the corresponding sections

in situations with uncertainty (Lascaux, 2008). In case of perfect knowledge of the other, one can act based on this knowledge, and does not need trust. Trust only comes into play when we do not know exactly what will happen (Lewis and Weigert, 1985). Thirdly, trust is typically characterised as a mental attitude, or a belief (Mayer et al., 1995), although some also allow trust to be a behaviour (Lee and See, 2004). However, in most conceptualisations trust and reliance are separated, where trust is the mental attitude or intention, and reliance the behavioural outcome. Although reliance is largely determined by trust, trust is not the only factor determining reliance. Finally, the mutual expectations and dependencies between trustor and trustee can be formalised in contracts. Contractual trust is achieved when a trustor has a belief that the trustee will stick to a specific contract (Jacovi et al., 2021, Duarte et al., 2023). For example, mutually agreed requirements can be used to specify the subject-matter of a contract, such as the various definitions of factors in the European Guidelines for Trustworthy AI Models (Directorate-General for Communications Networks, 2019). Contracts for different factors can be agreed upon, reflecting various dimensions of trust. Consequently, contracts specify the behaviour of the trustee to be anticipated and trust is the belief that a set of contracts will be upheld. In the case of an AI system, the

contracts imply an obligation by the AI developer to carry out a prior or expected agreement.

In cases of uncertainty and risk, a trustor needs to decide on whether they can rely on the trustee and their trustworthiness (Degli-Esposti and Arroyo, 2021). From this perspective, trust can be considered as the trustor's perception of the trustee's trustworthiness. There are different characterisations of how this perception is shaped, but one common perspective is that it can be divided into a belief of a trustee's ability, integrity, and benevolence (see ABI model in Figure 2 Mayer et al., 1995, Jorge et al., 2021). This conceptualisation of trustworthiness goes beyond the ability of the trustor to perform a task. In automation, many studies on trust in the past have focused purely on whether the system is trustworthy in the sense that it is able to do a job. However, a common notion of trust is that to trust means that a trustor must believe that the trustee has the intention to act in her best interest (Mayer et al., 1995). Intentions and interest go beyond simple capability, a trustee's personal choices also matter.

On the opposite, distrust is trust in the negative sense. A trustor A distrusts a trustee B, if A does not accept the vulnerability to B's actions, because A believes that B may not act in A's best interest (Jacovi et al., 2021). Distrust is not equal to the absence of trust, which might occur if the trust belief is not very strong, or not fully formed at all. In this case, a trustee might simply not know whether to trust or distrust.

In order to have the ability to anticipate the impact of the trustee's actions, determine their trustworthiness and form trust or distrust, a trustor must have experience with the trustee's behaviour or understand its functioning and motivation. In the case of a human's trust in an AI system, transparency and explainability are often mentioned as prerequisites. Another contributing factor to building trust involves an agent's reputation, i.e. what a social entity says about an agent regarding their behaviour. (Mui et al., 2003, Fullam et al., 2005, Herzig et al., 2009, Sabater-Mir and Vercouter, 2013, Pinyol and Sabater-Mir, 2013). Reputation is relevant particularly when direct experiences with a trustee are lacking.

Trust is not a homogeneous concept, but has multiple dimensions. As Mayer et al. (Mayer et al., 1995) explain, it is necessary to differentiate between trust itself, factors that contribute to trust and outcomes of trust. In dyadic trust among two specific parties, there are a few factors to consider, and multiple different theoretical models. Firstly, there are characteristics of the trustor itself, such as the trustor's propensity to trust (Mayer et al., 1995). Secondly, there are characteristics of the trustworthiness of the trustee as perceived by a trustor. Different perspectives on which characteristics exist, but one common model identifies ability, benevolence and integrity (ABI model, see Figure 2, Mayer et al., 1995). Ability is the set of skills and competencies in a specific domain. Benevolence is the level at

which a trustee is believed to benefit the trustor. Integrity refers the trustor's perception whether a trustee adheres to a set of personal or moral principles that the trustor identifies with.

Although there are different models of trustor's characteristics beyond the ABI model, most share a distinction between more functional aspects and more intentional aspects. In line with this, Malle and Ullman (Malle and Ullman, 2023), identify two main classes of trust, each with their characteristics:

- Performance
  - Reliability: are outputs coherent and consistent?
  - Competence: what is the quality and appropriateness of the output?
- Morality
  - Transparency: are functions and results understandable and reproducible?
  - Ethics: are values and norms respected?
  - Benevolence: does the trustee act in good faith and for the benefit of the trustor?

2.1.1 *Trust across disciplines*

As already introduced, the concept of trust has been studied in different disciplines. The general purpose of trust converges as: a way of facilitating collaboration in environments of uncertainty, allowing individuals to decide whether or not to rely on others. Despite that, there are also differences, mainly related to how trust is studied. From a sociological point of view, trust is viewed as a mechanism that reduces social complexity by means of shared social norms, institutions, and roles (Luhmann, 2018). Psychology distinguishes between cognitive trust, which is based on rational assessment of competence and reliability, and affective trust, which is grounded on emotional bonds and subjective factors such as personal connections (McAllister, 1995). Philosophy is primarily concerned with the ontological and normative conditions that trust must meet to be meaningful, which also include moral responsibilities and ethical obligations (Baier, 1986). Finally, disciplines such as AI, robotics, and computer science view trust as a variable that could be modelled and calibrated through features of the system (Castelfranchi and Falcone, 2010).

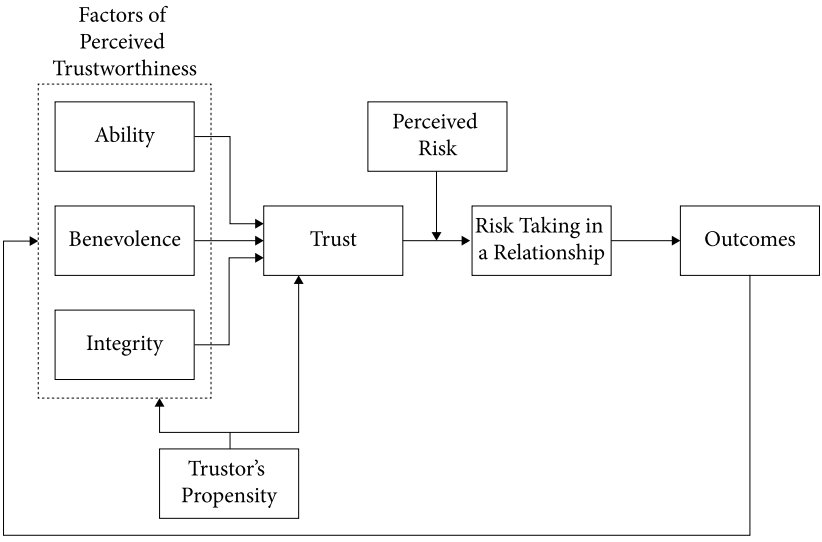

Figure 2. ABI model of trust (Mayer et al., 1995)

2.2 Trustworthiness

Trust comes from a perception by one entity of how worthy another entity is of being trusted. This worthiness, i.e. trustworthiness, can be seen as the property of an agent which determines whether they will uphold the trust contract (Jacovi et al., 2021). However, often definitions of trustworthiness will also attempt to define what it means to be worthy, or what contracts we should be able to expect a trustee to uphold. This means that definitions of trustworthiness of AI are often different from those of humans, for instance involving security, accountability and privacy, along with more common human factors such as reliability and fairness (Directorate-General for Communications Networks, 2019, Mattioli et al., 2023).

Beyond the fundamental requirements of being lawful, ethical and robust, the "European Guidelines for Trustworthy AI Models" (Directorate-General for Communications Networks, 2019) require AI systems to comply with a variety of principles:

1. Human agency and oversight.
2. Technical robustness and safety.
3. Privacy and data governance.
4. Transparency.
5. Diversity, non-discrimination and fairness.
6. Societal and environmental well-being.
7. Accountability.

While some of these principles reflect quantitatively measurable characteristics of an AI system,[4] many are of a qualitative nature. This is similar to the distinction between performative and moral trustworthiness above.

In a technical sense, an AI system is worthy of trust to a certain degree when it operates according to its specifications and requirements. Accordingly, actual trustworthiness can be measured by its performance characteristics, such as accuracy, resilience, robustness, etc. with regard to formal and functional requirements specifications. However, other elements are more difficult to objectively measure, such as fairness, societal well-being and accountability. Moreover, trustworthiness also depends on a given context, which includes the trustor, a physical or virtual environment and purpose. Given that trust can be seen as someone's belief on someone else's trustworthiness (Jorge et al., 2021), trustworthiness of the trustee can also depend on who the trustor is, making it more than an objective property of an agent. For example, benevolence of a trustee depends on shared values and familiarity of acquaintances, and will be different depending on the trustor.

2.3 Trust calibration

Trust calibration is the process of adjusting and aligning actual and perceived trustworthiness (see Figure 3 Okamura and Yamada, 2020, Visser et al., 2020, Visser et al., 2023, Chi and Malle, 2023). Actual trustworthiness (*Trustworthiness*) is the degree to which a system or actor complies with its required or promised expectations and performance, as defined above in Section 2.2. Perceived trustworthiness is the main antecedent to *Trust* as experienced by the trustor, as defined above in Section 2.1, and what we wish to adjust to match actual trustworthiness.

The level of perceived trustworthiness of an AI system is not necessarily equal to the actual objective trustworthiness. If the perceived trustworthiness is higher than the actual trustworthiness, the system will be over-trusted. This can lead to misuse of the system, because it can be used in cases where it does not perform sufficiently, according to exaggerated expectations (Parasuraman and Riley, 1997). Unjustified trust should be dampened to adjust the high perception to the actual level of trustworthiness. On the other hand, if the perceived trustworthiness is lower than the actual trustworthiness, the system will be under-trusted. This can lead to disuse of the system, because it will not be used in cases where it does perform well according to its specifications (Parasuraman and Riley, 1997).

4. https://oecd.ai/en/catalogue/metrics

Both trust repair and trust dampening refer to strategies to adjust human trust, typically to ensure better trust calibration. Trust repair occurs after a trust failure, and attempts to regain some of the trust which was lost (Kox et al., 2021, Tolmeijer et al., 2020, Tomlinson and Mayer, 2009, Visser et al., 2018). Trust dampening refers to the conscious lowering of trust in cases where trust is too high (Visser et al., 2020, Jensen and Khan, 2022).

Many different terms exist in the literature to refer to this concept of calibrated trust. Appropriate trust, warranted trust, responsible trust, etc. Following a literature review, Mehrotra et al. (2024) identify the most common terms, and suggest a conceptual mapping which defines whether the term is addressing calibrated beliefs, intentions or actions. In the case of calibrated trust, there is an implicit understanding that this is not just a state, but rather the result of a trust calibration process (Mehrotra et al., 2024). In Figure 1 we conceptualize this process of trust calibration as attempting to align natural trust with agent trustworthiness.

2.4 Artificial trust

We define *artificial trust* as trust by an artificial entity, such as an AI agent. This term is agnostic about how this is implemented, but this could for instance be as an internal variable representing a trust belief which is updated based on information about a trustee's trustworthiness and influences the agent's behaviour.

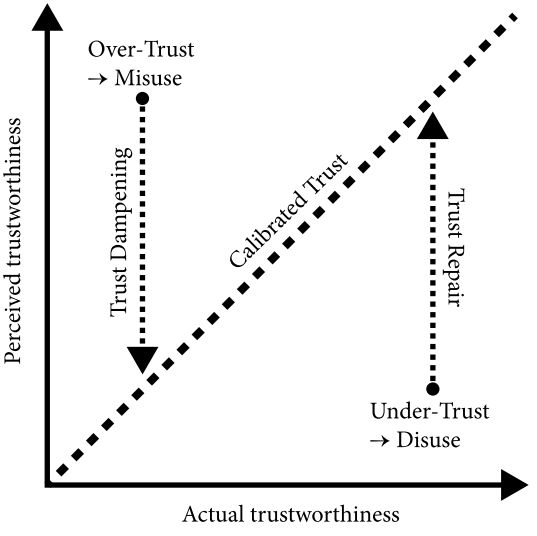

Figure 3. Trust Calibration (Visser et al., 2020, Visser et al., 2023)

Although the concept of having artificial agents with the ability to trust has been around for a while in multiagent systems (e.g. Falcone and Castelfranchi, 2004, Falcone et al., 2002, Sabater-Mir and Vercouter, 2013), only recently has there been an interest for an artificial agent to trust a human (e.g. Surendran and Wagner, 2019, Vinanzi et al., 2018). This interest raises as artificial agents become teammates, as they need to make decisions taking into account their teammates' ability and willingness for certain tasks. Before this recent interest, most works suggesting trust constructs for artificial agents to be trustors considered the trustee to be an artificial agent, too. Works such as (Jorge et al., 2021) suggest that *trust* can be a belief in the human teammate's *trustworthiness* to perform a task successfully. Thus, this belief can be used for an artificial teammate to calculate the teammates' reliability, similar to the way it was used in multiagent systems. However, saying that an artificial agent is enabled to trust a human is controversial for several reasons, including the fact that it worries people that they may be classified as untrustworthy. Some researchers also defend that trust is a human-only phenomenon and, as such, artificial agents could never trust. These researchers believe that trust cannot truly be modelled, and that artificial agent developers should instead focus on terms such as expectation, reliability, etc, which are already accepted in humanagent interaction. Potentially alleviating these worries, Azevedo-Sa et al. (2021) introduced the term *artificial trust* as a trust relationship where the trustor is an artificial agent (Azevedo-Sa et al., 2021). The authors distinguish this concept from *natural trust*, i.e., a trusting relationship where the trustor is a human, and open the road for an independent investigation of artificial trust, where models and definitions used on artificial agents do not have to align with the existing theories from social sciences. Artificial trust follows the definition of Kok and Soh (2020), which states that "given a trustor agent A and a trustee agent B, A's trust in B is a multidimensional latent variable that mediates the relationship between events in the past and A's subsequent choice of relying on B in an uncertain environment" (Kok and Soh, 2020). Departing from this notion, several works have advanced research on artificial trust for human-AI teams, such as exploring how it can be modelled (Jorge et al., 2024), how it can be used for decision-making (Jorge et al., 2022) or task allocation (Ali et al., 2022).

We distinguish *artificial trust*, i.e. the trust of an artificial trustor, from *computational trust*. Computational trust can be defined as the formal modelling of trust beliefs and dynamics into algorithms, representing either natural or artificial trust (Urbano et al., 2011). Computational trust models often collect information such as direct experiences, reputation, or recommendations to assess the trust or trustworthiness of agents, whether they represent individuals, organizations, or artificial entities (Sabater et al., 2005, Braga et al., 2018, Youssef et al., 2015). As seen in the previous paragraphs, when these beliefs are used to translate trust,

and the trustor is artificial, then we are talking about artificial trust, otherwise it is the computation of natural trust, for instance for the purpose of simulation or understanding humans. The computation of natural trust is the basis for calibrating trust in teams, and assessing team trust.

2.5 Team trust

The term team trust stems originally from social sciences (e.g., organisational psychology) and is generally defined as the shared belief among human team members that they can rely on each other to fulfil their roles and responsibilities within the team and that team members will act in a dependable, honest, and cooperative manner (Fulmer and Gelfand, 2012, Mayer et al., 1995). It encompasses confidence in teammates' intentions and competencies, facilitating collaboration under conditions of interdependence and uncertainty (Costa et al., 2018). Team trust plays a critical role in team effectiveness by fostering open communication, coordination, and resilience (Jong and Elfring, 2010). At the team level, trust is recognised as a complex construct consisting of multiple dimensions and layers (Castaldo et al., 2010). Conceptually, team trust integrates both individual and team-level perspectives, assuming a similarity in how trust beliefs of individual team members are shaped (Fulmer and Ostroff, 2021). Recently, these definitions have been further extended, suggesting team trust to be an emergent and dynamic state at the team level (Feitosa et al., 2020).

In human-AI teaming research, some initial work highlights that team trust may also have a substantial impact on team performance in human-AI teams (Georganta and Ulfert, 2024). Yet, conceptualisations of team trust in human-AI teams often do not follow the definitions proposed within organisational psychology but focus more on individual trust beliefs (Ulfert et al., 2024).

3. Research challenges

The previous section gives an overview of the core concepts in HAT trust. Based on this shared understanding, it is possible to now discuss the main open research challenges to achieving appropriate human-AI team trust. We can divide these challenges into three main groups. Firstly, we have natural trust calibration. This challenge is about ensuring calibrated trust from a human in the team and (AI) teammates. Secondly, we have artificial trust. This challenge is about ensuring appropriate trust from an AI agent in the team and the (human) teammates. Thirdly, we have team trust. This challenge is about re-defining what team trust means, how we achieve it and what its goals should be in a context where AI

becomes a team mate. These research challenges should be seen as first steps. Challenges around the influence of specific team settings still need more exploring.

3.1 Trust calibration

In Section 2.3 we describe the concept of Trust Calibration: the notion that trust needs to evolve and adjust to the trustee's trustworthiness to avoid under and over trust. There is a broad agreement that human trust in AI needs to be calibrated in this way (Lee and See, 2004, Winikoff, 2017, Bobko et al., 2022, Mehrotra et al., 2024). However, how to actually achieve this calibrated natural trust in human-AI teams is an open research question. In achieving calibrated trust in AI and human-AI teams, we focus on how to develop the AI with this goal in mind. This focus stems from the fact that these are the team members which we are actively designing and building, whereas we can only try to understand the human. Consequently, we identify two sub-challenges.

Challenge. How can AI understand a human's trust?

Firstly, we have the challenge of ensuring that the AI agent understands a human's trust. In order to calibrate trust, it is important to know whether there is over or under trust in the first place. An AI agent could only identify this, if it has some perception of the human's trust. Initial work in this direction has been done (Guo and Yang, 2020), but given the subjective nature of trust, automatic recognition of trust remains an open challenge.

Challenge. How can AI influence a human's trust?

Secondly, once under or over trust is recognised, the challenge is to influence the human's trust to combat this. What can the AI system do, in order to either dampen trust (in the case of over trust), or improve it (in the case of under trust)? Although there is quite a bit of work on improving trust (e.g. Kox et al., 2021, Tolmeijer et al., 2020, Tomlinson and Mayer, 2009), trust dampening is still an under-studied field (Visser et al., 2020, Jensen and Khan, 2022, Aroyo et al., 2021).

To tackle the first challenge, we can build on research into what influences trust of humans and AI. Studies show that trust in AI systems emerges from insight into and experience from using the systems, where commonly found characteristics that contribute to the emergence of trust are capability, explainability, integrity, reliability and anthropomorphism (Cabiddu et al., 2022, Glikson and Woolley, 2020). However, some of these characteristics can be difficult to quantify, and will affect different people in different ways (Küper and Krämer, 2023, Riedl,

2022). Therefore, just knowing how trustworthy these systems are with respect to those characteristics will not be enough.

Challenge. How do we deal with the subjective and dynamic nature of trust when trying to observe it?

Some efforts have been made to estimate human trust based on what they observe about agent performance (Guo and Yang, 2020). However, a remaining issue is that reliance behaviour isn't always sufficient to estimate trust, as hypothesised by theoretical models and supported by deviating measures of trust and reliance in experimental work (Mayer et al., 1995, Lee and See, 2004, Zhang et al., 2022). With constant trust questionnaires also not being a viable option, the best way to flexibly and dynamically estimate human trust remains an open challenge.

Challenge. Can we separate trust in an agent from trust in the humans behind it?

An additional difficulty when estimating trust in AI systems specifically, is that these systems usually do not operate in isolation. As much as an employee represents a company, so do most AI systems represent the company which builds or provides them. With the increasing power and concerns over big-tech, the influence of their reputation on the trust people have in their AI systems should not be under-estimated. However, the influence of trust in the developer on trust in the AI system is still an under-studied topic. Moreover, when AI is integrated in professional teams, the person behind the AI's introduction will also play a big role. Does trust in the manager influence trust in the AI team mate that they introduce? These additional players play a role which should be taken into account by any AI system estimating human trust (Cameron et al., 2024).

Challenge. Can we quantify how much trust is appropriate in a given situation?

Much of the current work on trust has focused on increasing trust (e.g. Nam and Lyons, 2020, Tucci et al., 2022, Kox et al., 2021), but as we want to avoid over-trust in teams, just increasing trust endlessly is not the goal. But how can an AI agent determine when trust is 'too high'? The first part of the answer lies in the first challenge described: knowing what a human's trust level is. The second part of the answer lies in determining what the trust level should be in a given situation, for a given agent, in order to be able to compare the two. However, selecting this level is not straightforward. Some situational aspects might play a role, for instance the interdependencies involved might influence this (Johnson and Bradshaw, 2021, Verhagen et al., 2024). Some studies look at situations in which over-reliance is easily established, for instance when monitoring fails or a choice needs to be made

which is clearly correct or not (Vasconcelos et al., 2023). However, there is also an argument that the trust level should depend on the agent's inherent trustworthiness, and not only on human behaviour, as it has been shown that these do not always align (Lee and Chew, 2023, Zhang et al., 2020).

Challenge. How can we compare human trust to an agent's trustworthiness in a meaningful way?

Following this argument, when trust is calibrated, the actual and perceived trustworthiness should align (Visser et al., 2020, Visser et al., 2023, Verhagen et al., 2022). This means that some form of comparison between a human's trust and the agent's trustworthiness should occur. This brings it's own challenges, firstly how to quantify an agent's trustworthiness. This seems inherently context-dependent and, some have even argued, dependent on the trustee (Jorge et al., 2021). However, even if a solid representation of trustworthiness can be achieved, there is still an open question of how to compare this to a human's trust. If trust is measured subjectively, for instance, does it truly make sense to compare a 'trustworthiness' number with questionnaire outcomes? How to make such a comparison between human trust and the agent's trustworthiness is very much still an open question.

Challenge. What strategies should AI systems use to increase or dampen trust?

When under- or over-trust has been established, the second challenge is to influence trust to either decrease or increase. Most work has focused on trying to increase trust, for instance after trust violations have reduced it (Tolmeijer et al., 2020, Kox et al., 2021). Less work has been done on trying to dampen trust, despite work showing the danger of over-trust (Robinette et al., 2016, Vasconcelos et al., 2023). In cases of over-trust, the AI should try to lower expectations of a human (Wagner et al., 2018). Despite the large body of work studying antecedents of trust, a lot more is still necessary to determine the correct strategies in cases of under- and over-trust.

Challenge. Are there trade-offs to more calibrated trust, which might not be worth it?

More work on calibrated trust might also shed light on the trade-offs of trying to achieve it. Teamwork often has some goals for efficiency and speed. Spending a lot of time on trust calibration might not always be worth it, given the mental effort which can be involved (Vasconcelos et al., 2023). Which settings would allow for less calibration is still an open question, as is more research into what the exact trade-offs would be.

Challenge. How can trust calibration take into account changing contexts as AI itself evolves?

Finally, AI is moving increasingly from a simple tool, to a more sophisticated teammate. When designing for trust calibration in human-AI teams, it is important to realise that at least one of the teammates is constantly changing and evolving, and probably becoming increasingly autonomous. This might, in turn, influence both human trust in the AI and the amount of risk involved in teamwork. How this evolution will influence trust calibration is still very much an open question.

3.2 Artificial trust

The counterpoint to calibrated trust, i.e. appropriate natural trust of a human in their team members, is artificial trust, i.e. appropriate trust of an artificial agent in their team members. In human-human teams, mutual trust is an antecedent of good team work (Salas et al., 2005), but for human-AI teams this is not as commonly accepted. Firstly, there is the justified question whether agents can trust as humans do, and if not, whether it makes sense to talk about mutual trust at all. Secondly, there is a concern that if agents are *allowed* to distrust humans, they might ignore or overrule humans in situations where this is not appropriate. Although discourse on both these questions is on-going, in this paper we follow the following premises. Firstly, as described in Section 3.2, we use *artificial trust* for agents using internal models of another's trustworthiness in their decision making on whether to rely on them in uncertain circumstances. We argue that this term makes it clear that we are referring to mechanisms inspired by human trust, yet which aren't necessarily the same. Secondly, we note that artificial trust should always aim to be appropriate, i.e. there should be no under trust, but also no over trust in a human. The boundaries for these might be different in situations with an agent trusting a human than vice versa, as we might want to err on the side of trusting the human too much, and the agent too little, especially in morally sensitive situations (Sio and Hoven, 2018, Waa et al., 2020, Verhagen et al., 2024). However, in many situations, it does make sense for an agent team mate to have a realistic estimation of human team mates' trustworthiness, even if this results in distrust. A simple example would be a task where an agent can ask two different humans for help. In this situation, distrusting the one who is busy and far away makes sense, as does trusting the one who is close by and idle. This also relates to work on "disobedient AI", where the case is made that in some cases, it is in a human's best interest to not do as they say, for instance if they don't have all information (Briggs et al., 2024, Briggs et al., 2022).

Challenge. What can we observe to base artificial trust on?

At its core, artificial trust is an internal model or variable of the agent. This can be based on previous knowledge and observations in the environment. These observations are crucial if we want artificial trust to adapt to the situation and specific team mates the agent is interacting with. Ideally, these observations say something about the trustworthiness of the trustee. This means that a central challenge in artificial trust is about how an AI can observe cues of trustworthiness (Jorge et al., 2024). This is challenging for several reasons. Firstly, what constitutes as trustworthy behaviour will differ per situation. In some situations, it is more important to do something quickly, and in others precision matters more. Understanding this situational context is crucial. Secondly, agents rarely exist in an environment with perfect observability. Some actions might not be observable at all, while others might not be interpretable until later. Examples are tasks that happen out of sight, where it might only be seen later whether they were completed; or a message of which the truthfulness cannot be determined until later. Finally, common antecedents of trust in humans are often not observable from a single action at all. Integrity and benevolence for instance, while well understood in human trust literature (Mayer et al., 1995, Breakey et al., 2015, Colquitt and Salam, 2009), are difficult to reduce to single actions. What an AI agent can actively observe which is meaningful to trust, and how to interpret those cues into larger constructs is a big open question.

Challenge. How do we deal with the lack of ground truth?

Artificial trust is essentially a model of trustworthiness, in the case of HAT often of a human. Ideally, such a model is evaluated, where we see how well it matches its real world counterpart. However, trustworthiness of a human is not something we can easily measure, or even define. This means that evaluating artificial trust on the basis of how well it represents actual trustworthiness is nearly impossible to do. There is no easy ground truth to compare to, which leads to the question of how we should be evaluating artificial trust instead?

Challenge. How do we determine the right goal of artificial trust?

The lack of ground truth to evaluate artificial trust in humans is a hurdle, but also serves to re-evaluate the goal of artificial trust. Often, the goal is not truly to perfectly model how humans would trust. Rather, the goal is to improve teamwork by allowing the agent to make better decisions using artificial trust. This means that there are potential alternative ways to evaluate which look at the result of an agent's behaviour, rather than its internal state. However, which evaluation metrics to use is still an open question. Some models look at improving appropri-

ate reliance (Vasconcelos et al., 2023, Schemmer et al., 2023), and another option would be to consider the result of task allocation. This does also raise the question of whether it always makes sense to equate good behaviour with a good trust model, if it is even possible to define good behavioural outcomes. How to properly and consistently evaluate artificial trust in a way which allows for comparison between models is an open question.

Challenge. How do we make an agent transparent about artificial trust?

Finally, a related question is how humans will respond to agents which can trust and distrust them. There is a broad push for transparent and explainable AI (Felzmann et al., 2020, Miller, 2019, IEEE, 2018), and ideally human team mates will know that agents can have a trust model which applies to them. However, how an agent should express such models is still very much under-studied. Despite the large body on explainable AI (Anjomshoae et al., 2019, Adadi and Berrada, 2018, Arrieta et al., 2020), very rarely does this involve an AI explaining what it thinks about the one it is explaining to. One can imagine that an agent stating 'I don't trust you' might have averse effects. How to balance the need for transparency with the need for a good working relationship in teams is very much unexplored.

3.3 Team trust

Depending on whether trust is conceptualised as a belief that only humans possess and how it is precisely defined, new challenges also arise in terms of understanding human-AI team trust.

Challenge. How do human trust and artificial trust differ, and how can they be compared and integrated in team trust?

Although humans and agents are likely to evaluate their team members in different ways, following the propositions by Ulfert and colleagues (Ulfert et al., 2024), it is important to note that all team members will form some form of belief about their team members and the team as a whole. The difficulty will lie in comparing these beliefs within team settings. For example, is it comparable when, within a team, a human or an AI agent makes an evaluation of a new team member based on information from a newspaper article? While the evaluation is mainly based on the same information (i.e., the newspaper article) and might result in the same outcome (e.g., a specific assessment of the new team member's ability), prior knowledge, biases, or the way the information is assessed may fundamentally differ between team members. For instance, the human team member may connect the information included in the article to prior knowledge or beliefs of individuals resembling the descriptions, whereas the AI agent's knowledge base

may be fundamentally different. At present, we lack methods to make a statement about how these team member evaluations could be compared or even combined.

Challenge. Do beliefs exist that are shared between humans and AI?

Although first approaches have been proposed to conceptually integrate artificial and human trust at the team level (e.g., Ulfert et al., 2024), it is unclear how and to what extent trust by the different team members would and should contribute to a shared belief. This is also related to the question of the extent to which concepts, such as trust, should be emulated in AI agents.

Challenge. Should we strive to model human-AI team trust after team trust in humans, or is it something completely new?

Even though modelling concepts after theories from other disciplines such as psychology has been a successful approach for the development of AI, it remains unclear to what extent human behaviour should be emulated. The same question should be posed in the context of trust in human-AI teaming. When would a shared trust belief be helpful for collaboration and when would it be detrimental?

4. Measures and methods

The previous section identifies some of the main research challenges in the field of human-AI team trust. When starting to address these challenges, we pose that there are important methodological questions to consider. From the workshop discussions, two main themes arose: measures and research environments. In this section, we present a reflection on potential measures for trust in human-AI teams, as well as requirements for valid research environments to study trust in teamwork.

4.1 Trust measures

Most studies on trust in HAT are experimental in nature; they investigate how human trust is influenced by different settings or agent behaviour (e.g. Verhagen et al., 2022, Esterwood, 2023, Hannibal et al., 2022). This setting necessarily means that natural trust needs to be observed or measured in some form or way. As seen in Section 2.1, trust is conceptualised in different ways, although commonalities exist. We pose that in experimental work on natural trust, the measurements of trust should match with the definitions that are used in the paper.

One important distinction to make here is between measures of behaviour and measures of an internal state. Most often, trust is conceptualised as an attitude

or belief (Mehrotra et al., 2024, Lee and See, 2004, Castelfranchi and Falcone, 2010). This means that the measures for trust need to be inherently subjective. The participants should be asked about their attitudes and beliefs. Most studies, therefore, take the route of questionnaires to measure trust. However, often some behavioural measures are included, which observe the actions of the individual, for instance the reliance behaviour. Such measures can be valuable, however, it is important to not equate reliance behaviour with trust unless one explicitly defines trust as a behaviour. There are studies which talk about trust as an attitude while measuring behaviour, which we would advice against.

Many different questionnaires on trust have been designed and validated, both originally for trust in other humans, and for trust in different types of systems (e.g. Lewicki and Brinsfield, 2015, Malle and Ullman, 2023, Gulati et al., 2019, Adams et al., 2008, Spain et al., 2008). Although one could wonder if we really require so many different scales, it makes sense when one considers that these questionnaires usually define what it means to be trustworthy in a given setting. Rather than asking directly about trust, they ask whether the trustor believes the trustee is e.g. capable, honest, etc. What defines trustworthiness in an interaction is, however, dependent on both contextual factors and the nature of the trustee. This legitimises the existence of many different questionnaires.

Most questionnaires for natural trust adopt the notion of different dimensions of trust. Different dimensions exist, for instance Ability, Benevolence, Integrity, following (Mayer et al., 1995), Competence and Willingness, following (Falcone and Castelfranchi, 2004), or Reliable, Competent, Ethical, Sincere (Ullman and Malle, 2018). These last factors appear in the Multi-Dimensional Measure of Trust, which was developed specifically for robots (Ullman and Malle, 2018, Malle and Ullman, 2023). Based on words which in some way describe trustworthiness, this questionnaire identified these four categories bottom-up, through an analysis of the relationships between the words. Although it might not always make sense to talk about an AI agent as 'sincere' or 'ethical', this questionnaire shows that people do often think about artificial agents in these terms. With the increase of AI systems which make decisions with moral impact, this trust in the ethical aspects of AI might be increasingly important. In general, despite their differences, trust scales distinguish between more performance/capability/reliability based trust (i.e. can the system do the thing), and more benevolence/moral/willingness related trust (i.e. will the system want to do the right thing). Which exact questionnaire suits HAT studies is dependent on which dimensions make sense in the specific teamwork settings and application domain.

Although less common than questionnaires, studies also attempt to measure or approximate trust through looking at user behaviour (Mehrotra et al., 2024). How exactly this is done typically depends on the type of task and in which ways

the user can rely on the system. Agreement percentage, which defines how often the user agrees with an AI's prediction is one method, but which only works in tasks where a clear single choice needs to be made. A related measure is switch percentage, how often people switch to an agent's choice for a final decision. Both these measures only work in tasks where there is a binary choice to rely on the system or not, and a clear binary judgement on whether a task (or choice) is done correctly or not. This means there is currently a quite limited view on measuring reliance behaviour related to trust.

4.2 Research setting

During the workshop's group discussions, some of the crucial components for meaningful studies into human-AI team trust were identified. In this section, we will describe these components. These *research settings* are the choices we make when we design the teamwork between humans and AI in studies, i.e. how the team members can work together and what the settings of the task are.

Risk: Trust is critical in all circumstances where people are in any way dependent on others' actions, and thus, it is more relevant in high-risk situations. More specifically, more trust is required when the perceived risk of relying on someone or something else is higher (Mayer et al., 1995), and one could argue that when there is no risk, trust is fully irrelevant (Stuck et al., 2021). For this reason, there needs to be some risk of the trustor relying on the trustee in order to meaningfully talk about trust. For example, an AI support system for logistics decisions would have risk associated with wrong decisions, e.g. a loss in revenue. In studies this could be translated as points lost on wrong decisions. Or a manufacturing human-robot team would have risks in the form of efficiency loss, which could be translated into penalty scores in an experiment. Experimental settings need to balance between the necessity to keep risks for human participants to a minimum, and the necessity of some risk in order to meaningfully study trust.

Dyadic 1-1 or 1-team: Much of the literature on human-AI team trust revolves around dyadic trust, i.e. trust between two parties. An example would be a single image classification algorithm working together with a medical professional. However, this may be an oversimplification, especially in bigger teams in which there are more agents (artificial or human). In such cases, trust could be computed with respect to a team of agents instead of a single individual. For example, when a search-rescue ground robot operates in a larger team with multiple professionals and a drone. In any case, the content of the team should always be well-described, and there is a current lack in studies considering teams larger than two agents.

Goal of the team: Within a specific team, the goals of different members must be aligned (Hoff and Bashir, 2015). Misaligned goals between different participants may lead to distrust since trust is also described as the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability (Lee and See, 2004). In designing studies on team-trust, care should be taken that goals are explicitly defined on a team level, and aligned among team members. An example would be a waiter robot working with a team of human waiters. It should be clear beforehand whether the robot is there to assist a specific human with their tables, or whether the robot has it's own tables to take care of.

Interdependencies: Interdependence refers to the mutual reliance between humans and AI systems in completing a task. Tasks with high level of interdependence requires human and AI agents to work closely and rely on each other. For example, a firefighting human-robot team where the robot relies on the human for decision making, and the human on the robot for going into unstable buildings. Factors influencing interdependence deal with the design of shared goals, division of labour and coordination/collaboration. Coactive Design (Johnson et al., 2014) is a tool to help designers identify interdependence relationships in a joint activity, so they can design systems that support these relationships, thus enabling designers to achieve the objectives of coordination, collaboration, and teamwork (Johnson et al., 2018). In studies on team trust, interdependence relationships should be realistic, and well-described.

Tool vs teammate: If an agent is seen merely as a tool, trust may be transactional and only dependent on performance. For instance, using an LLM to summarize a paper and speed up academic reading. However, if the agent takes on a more teammate-like role, deeper trust may be required. For instance, if the LLM becomes an academic assistant used for mutual brainstorming on hypotheses. Further, as artificial agents become more complex and go beyond tools, supporters (augmenting human capability), or performers (autonomously executing tasks), the importance of trust increases (Kox et al., 2021). A good understanding of the role of AI as tool or teammate is important for interpreting the results of trust studies.

Trust Calibration and Experiment: The length and continuity of interactions plays a significant role in trust development. Trust in human-AI teams is not static; it evolves based on the frequency and duration of collaborative experiences. Lengthier, ongoing engagements often allow for the gradual calibration of trust as users can evaluate the agent's performance over time. This continuous engagement gives the human partner the opportunity to refine their expectations, adjust reliance levels, and assess the agent's reliability. For example, an app giving recommendations on healthy eating over the course of weeks and aligning with user

expectations would give users the opportunity to calibrate trust over many sessions, while an app used for a single recommendation for a meal would not. Therefore, experiments should take this into account, monitoring and evaluating the values of trust over time.

Human Control, Transparency and Adaptability: Discussions around the topic of human control are also relevant for trust in HAT. There is a common agreement that a system must always have the ability to be stopped by a human, even when the human's condition may be clouded by stress or anxiety (Sio and Hoven, 2018). This is particularly relevant in situations where morally relevant decisions are being made, such as in security (search-rescue, firefighting) or healthcare (clinical decision support, surgical robot). Furthermore, a trustworthy agent should be somewhat predictable or at least transparent, providing on-demand explanations for its choices or reasoning (IEEE, 2018). Depending on the context, explanations should be adapted to the level of expertise of the human teammate and allow for meaningful human control. These factors should be taken into account when designing good AI teammates for human-AI team studies.

Stakeholder Involvement and Realism: Including stakeholders from various levels, such as users, developers, and decision-makers, in the design and evaluation phases of human-AI teams is essential for designing ecologically valid studies. Many current studies are done in dummy environments, which makes it difficult to see how results translate to real-world settings. More realism in the types of tasks, interactions, risks and timelines would benefit research into HATs.

5. Conclusions

In this paper, we present the results of three editions of the MULTITTRUST workshop on trust in human-AI teams. The workshop took an multidisciplinary approach to this research field, which requires an understanding of mutual trust relationships in a teamwork context. In this paper, we present a shared understanding of the key concepts in this research field, namely Trust, Trustworthiness, Trust Calibration, Artificial Trust, and Team Trust. Furthermore, we identify core research challenges in the field, which can be broadly grouped into achieving calibrated natural trust from the human, achieving appropriate artificial trust from the agent, and understanding team trust in a human-AI team context. Finally, we present some of the key characteristics of meaningful human-AI teamwork studies, with a particular focus on measures and research settings. With this work, we present an overview of the state of the art in the field of human-AI team trust, and key open challenges still to be addressed.

Funding

This work is partially funded by the European Regional Development Fund (ERDF) and the Saarland within the scope of the project (To)CERTAIN. Morgan Bailey's work was supported by the UKRI Centre for Doctoral Training in Socially Intelligent Artificial Agents, Grant Number EP/S02266X/1. Francesco Frattolillo's work was supported by the Air Force Office of Scientific Research under award number FA8655-23-1-7257

Author's contributions

All authors contributed to writing and revising (sections of) the manuscript, and were involved in the organisation of editions of the MULTITTRUST workshop. C. Centeio Jorge and A.S. Ulfert organised the first edition of the workshop, conceptualising what the workshop is about. M.L. Tielman, M. Bailey, F. Frattolillo and A. Meyer-Vitali organised the third edition of the workshop, from which this paper directly followed. M.L. Tielman and A. Meyer-Vitali conceptualised the overall storyline of the paper. M. Bailey and M.L. Tielman wrote the introduction. A. Meyer-Vitali, M.L. Tielman, C. Centeio Jorge and A.S. Ulfert wrote the concepts section, M.L. Tielman wrote the challenges section with contributions by A.S. Ulfert, and M.L. Tielman & F. Frattolillo wrote the Measures & Methods section. M.L. Tielman edited the first full paper draft and led the revision process, all authors edited the full final draft and contributed to the revisions.

Acknowledgements

This work is partially supported by the Delft AI Initiative, as part of the AI*MAN lab.

This publication is part of the project 'Hybrid Intelligence: augmenting human intellect' (https://hybrid-intelligence-centre.nl,) with project number 024.004.022 of the research programme 'Gravitation' which is (partly) financed by the Dutch Research Council (NWO).

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), *IEEE access 6* 52138–52160. Publisher: IEEE.

Adams, B.D., Waldherr, S., & Sartori, J. (2008). Trust in Teams Scale, *Trust in Leaders Scale: Manual for Administration and Analyses*.

Ali, A., Azevedo-Sa, H., Tilbury, D.M., & Robert Jr, L.P. (2022). Heterogeneous human-robot task allocation based on artificial trust, *Scientific Reports 12* 15304.

Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable Agents and Robots: Results from a Systematic Literature Review, in: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, International Foundation for Autonomous Agents and Multiagent Systems*, pp. 1078–1088. Place: Richland, SC.

Aroyo, A. M., Bruyne, J. D., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., Jones, S., Lutz, C., Sætra, H., Solberg, M., & Tamò-Larrieux, A. (2021). Overtrusting robots: Setting a research agenda to mitigate overtrust in automation, Paladyn, *Journal of Behavioral Robotics 12* 423–436. https://www.degruyter.com/document/doi/10.1515/pjbr-2021-0029/htmlDe Gruyter Open Access Section: Paladyn.

Azevedo-Sa, H., Yang, X. J., Robert, L. P., & Tilbury, D. M. (2021). A Unified Bi-Directional Model for Natural and Artificial Trust in Human-Robot Collaboration, *IEEE Robotics and Automation Letters 6* 5913–5920. conference Name: IEEE Robotics and Automation Letters.

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2024). A systematic literature review of user trust in ai-enabled systems: An hci perspective, *International Journal of Human-Computer Interaction 40*.

Baier, A. (1986) Trust and antitrust, *Ethics 96* 231–260. https://www.jstor.org/stable/2381376

Barredo Arrieta, A., Díaz-Rodríguez, N., J. Del Ser, Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion 58* 82–115.

Berretta, S., Tausch, A., Ontrup, G., Gilles, B., Peifer, C., Kluge, A. (2023). Defining human-AI teaming the human-centered way: a scoping review and network analysis, *Frontiers in Artificial Intelligence 6*. Frontiers.

Bobko, P., Hirshfield, L., Eloy, L., Spencer, C., Doherty, E., Driscoll, J., & Obolsky, H. (2022). Human-agent teaming and trust calibration: a theoretical framework, configurable testbed, empirical illustration, and implications for the development of adaptive systems, *Theoretical Issues in Ergonomics Science*. 1–25. Taylor & Francis.

Braga, D. D. S., Niemann, M., Hellingrath, B., & Neto, F. B. D. L. (2018). Survey on computational trust and reputation models, *ACM Computing Surveys 51*.

Brandizzi, N., C. Centeio Jorge, Cipollone, R., Frattolillo, F., Iocchi, L., & A.-S. Ulfert-Blank (2023). Multittrust: 2nd workshop on multidisciplinary perspectives on human-ai team trust, in: *Proceedings of the 11th International Conference on Human-Agent Interaction*, pp. 496–497.

Breakey, H., Cadman, T., & Sampford, C. (2015). Conceptualizing Personal and Institutional Integrity: The Comprehensive Integrity Framework, volume 14 of *Research in Ethical Issues in Organizations*, Emerald Group Publishing Limited, pp. 1–40.

Briggs, G., Williams, T., Jackson, R. B., & Scheutz, M. (2022) Why and How Robots Should Say 'No', *International Journal of Social Robotics 14*. 323–339.

Briggs, G., Law, T., Mirsky, R., Rogers, K., & Rosero, A. (2024). Rebellion and Disobedience in Human-Robot Interaction (RaD-HRI), in: *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, HRI '24*, Association for Computing Machinery, New York, NY, USA, pp. 1308–1310.

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? workforce implications, *Science 358*.

Burnett, C., Norman, T. J., & Sycara, K. (2011). Trust decision-making in multi-agent systems, in: *IJCAI International Joint Conference on Artificial Intelligence*.

Cabiddu, F., Moi, L., Patriotta, G., & Allen, D. G. (2022) Why do users trust algorithms? A review and conceptualization of initial trust and trust over time, *European management journal 40*. 685–706. Elsevier.

Cameron, D., Collins, E. C., S. de Saille, Eimontaite, I., Greenwood, A., & Law, J. (2024) The Social Triad Model: Considering the Deployer in a Novel Approach to Trust in Human-Robot Interaction, *International Journal of Social Robotics 16*. 1405–1418.

Campagna, G., & Rehm, M. (2025) A Systematic Review of Trust Assessments in Human-Robot Interaction, *Journal of Human-Robot Interaction 14*. *30*:1–30:35.

Castaldo, S., Premazzi, K., & Zerbini, F. (2010) The meaning (s) of trust. a content analysis on the diverse conceptualizations of trust in scholarly research on business relationships, *Journal of business ethics 96*. 657–668.

Castelfranchi, C., & Falcone, R. (2010). Definitions of Trust: From Conceptual Components to the General Core, in: *Trust Theory: A Socio-Cognitive and Computational Model*, Wiley, pp. 7–33. https://ieeexplore.ieee.org/document/8041696. conference Name: Trust Theory: A Socio-Cognitive and Computational Model.

Castelfranchi, C., & Falcone, R. (2010). Trust theory: A socio-cognitive and computational model, *John Wiley & Sons*.

Chi, V. B., & Malle, B. F. (2023). Calibrated Human-Robot Teaching: What People Do When Teaching Norms to Robots*, in: *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1308–1314. https://ieeexplore.ieee.org/abstract/document/10309635. iSSN: 1944-9437.

Colquitt, J. A., & Salam, S. C. (2009). Foster trust through ability, benevolence, and integrity, *Handbook of principles of organizational behavior: Indispensable knowledge for evidence-based management*. 389–404. A John Wiley and Sons, Ltd, Publication.

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*.

Degli-Esposti, S., & Arroyo, D. (2021). Trustworthy humans and machines, in: *Trust and Transparency in an Age of Surveillance*, 1 ed., Routledge, London, pp. 201–220. https://www.taylorfrancis.com/books/9781003120827/chapters/10.4324/9781003120827-15.

Directorate-General for Communications Networks, Content and Technology (European Commission), Grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji, Ethics guidelines for trustworthy AI, Publications Office of the European Union. (2019). https://data.europa.eu/doi/10.2759/346720

Duan, W., Flathmann, C., McNeese, N., Scalia, M. J., Zhang, R., Gorman, J., Freeman, G., Zhou, S., Hauptman, A. I., & Yin, X. (2025). Trusting Autonomous Teammates in Human-AI Teams — A Literature Review, in: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25*, Association for Computing Machinery, New York, NY, USA, pp. 1–23.

Duarte, R. d. B., Correia, F., Arriaga, P., & Paiva, A. (2023). AI Trust: Can Explainable AI Enhance Warranted Trust? — de Brito Duarte — 2023 — Human Behavior and Emerging Technologies — Wiley Online Library, *Human behavior and Emerging technologies* (2023). https://onlinelibrary.wiley.com/doi/10.1155/2023/4637678

Esterwood, C., & Robert Jr., L. P. (2023). Three Strikes and you are out!: The impacts of multiple human-robot trust violations and repairs on robot trustworthiness, *Computers in Human Behavior 142*. 107658. https://www.sciencedirect.com/science/article/pii/S0747563223000092.

Falcone, R., & Castelfranchi, C. (2004). Trust dynamics: How trust is influenced by direct experiences and by trust itself, in: *3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004)*, 19–23 August, New York, NY, USA, IEEE Computer Society, 2004, pp. 740–747. https://doi.ieeecomputersociety.org/10.1109/AAMAS.2004.10084.

Falcone, R., Pezzulo, G., & Castelfranchi, C. (2002). A fuzzy approach to a belief-based trust computation, in: R. Falcone, K. S. Barber, L. Korba, M. P. Singh (Eds.), *Trust, Reputation, and Security: Theories and Practice, AAMAS 2002 International Workshop, Bologna, Italy, July 15, 2002, Selected and Invited Papers, volume 2631 of Lecture Notes in Computer Science*, Springer, pp. 73–86.

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

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamo-Larrieux, A. (2020). Towards Transparency by Design for Artificial Intelligence, *Science and Engineering Ethics 26*. 3333–3361.

Fullam, K. K., Klos, T. B., Muller, G., Sabater, J., Schlosser, A., Topol, Z., Barber, K. S., Rosenschein, J. S., Vercouter, L., & Voss, M. (2005). A specification of the agent reputation and trust (art) testbed: experimentation and competition for trust in agent societies, in: *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS '05, Association for Computing Machinery*, New York, NY, USA, p. 512–518.

Fulmer, C. A., & Gelfand, M. J. (2012). At what level (and in whom) we trust: Trust across multiple organizational levels, *Journal of management 38*. 1167–1230.

Fulmer, C. A., & Ostroff, C. (2021). Trust conceptualizations across levels of analysis, in: *Understanding trust in organizations*, Routledge, pp. 14–42.

Georganta, E., & Ulfert, A.-S. (2024). Would you trust an ai team member? team trust in human-ai teams, *Journal of Occupational and Organizational Psychology*.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research, *Academy of Management Annals 14*.

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research, *Academy of Management Annals*.

Gulati, S., Sousa, S., & Lamas, D. (2019). Design, Development and Evaluation of a Human-Computer Trust Scale, *Behaviour & Information Technology 38*. 1004–1015Taylor &; Francis.

Guo, Y., & Yang, X. J. (2020). Modeling and Predicting Trust Dynamics in Human-Robot Teaming: A Bayesian Inference Approach, *International Journal of Social Robotics*. Springer Science and Business Media B.V.

Hannibal, G., Dobrosovestnova, A., & Weiss, A. (2022). Tolerating Untrustworthy Robots: Studying Human Vulnerability Experience within a Privacy Scenario for Trust in Robots, in: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 821–828. https://ieeexplore.ieee.org/abstract/document/9900830. , iSSN: 1944-9437.

Herzig, A., Lorini, E., Hubner, J.F., & Vercouter, L. (2009). A logic of trust and reputation, *Logic Journal of the IGPL 18*. 214–244.

Hoff, K.A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust, *Human Factors 57*.

Huber, S., Weppert, L., Baumeister, L., Happel, O., & Grundgeiger, T. (2025). Team Roles of Artificial Intelligence in Anesthesiology — A Scoping Review, in: *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '25*, Association for Computing Machinery, New York, NY, USA, pp. 1–13.

Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems version 2, Technical Report, IEEE, 2018.

Jacovi, A., Marasović', A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI, in: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 624–635.

Jensen, T., & Khan, M.M.H. (2022). I'm Only Human: The Effects of Trust Dampening by Anthropomorphic Agents, in: J.Y.C. Chen, G. Fragomeni, H. Degen, S. Ntoa (Eds.). *HCI International 2022 — Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, Springer Nature Switzerland, Cham, pp. 285–306.

Johnson, M., & Bradshaw, J.M. (2021). The role of interdependence in trust, in: *Trust in Human-Robot Interaction*, Elsevier, pp. 379–403.

Johnson, M., Bradshaw, J.M., Feltovich, P.J., Jonker, C.M., M.B. van Riemsdijk, & Sierhuis, M. (2014). Coactive design: designing support for interdependence in joint activity, *J. Hum.-Robot Interact. 3*. 43–69.

Johnson, M., Bradshaw, J.M., & Feltovich, P.J. (2018). Tomorrow's human-machine design tools: From levels of automation to interdependencies, *Journal of Cognitive Engineering and Decision Making 12*. 77–82.

Jong, B.A. De, & Elfring, T. (2010). How does trust affect the performance of ongoing teams? the mediating role of reflexivity, monitoring, and effort, *Academy of Management journal 53*, 535–549.

Jorge, C. Centeio, & A.S. Ulfert-Blank (2023). Multittrust-multidisciplinary perspectives on human-ai team trust, in: *CEUR Workshop Proceedings*, volume *3456*, CEUR-WS, pp. 132–136.

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

Jorge, C. Centeio, Tielman, M.L., & Jonker, C.M. (2022). Artificial trust as a tool in human-ai teams, in: D. Sakamoto, A. Weiss, L.M. Hiatt, M. Shiomi (Eds.), *ACM/IEEE International Conference on Human-Robot Interaction, HRI 2022*, Sapporo, Hokkaido, Japan, March 7 — 10, 2022, IEEE / ACM, pp. 1155–1157.

C. Centeio Jorge, Jonker, C.M., & Tielman, M.L. (2024). How should an AI trust its human teammates? exploring possible cues of artificial trust, *ACM Transactions of Interactive Intelligent Systems 14*, 5:1–5:26.

Kahr, P., Rooks, G., Snijders, C., Willemsen, M.C. (2025). Good Performance Isn't Enough to Trust AI: Lessons from Logistics Experts on their Long-Term Collaboration with an AI Planning System, in: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25*, Association for Computing Machinery, New York, NY, USA, pp. 1–16.

Kaur, D., Uslu, S., Rittichier, K.J., & Durresi, A. (2023). Trustworthy artificial intelligence: A review, *ACM Computing Surveys 55*.

Kok, B.C., & Soh, H. (2020). Trust in robots: Challenges and opportunities, *Current Robotics Reports 1*, 297–309.

Kolomaznik, M., Petrik, V., Slama, M., & Jurik, V. (2024). The role of socio-emotional attributes in enhancing human-AI collaboration, *Frontiers in Psychology 15*. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1369957/fullFrontiers.

Kox, E.S., Kerstholt, J.H., Hueting, T.F., & P.W. de Vries (2021). Trust repair in human-agent teams: the effectiveness of explanations and expressing regret, *Autonomous Agents and Multi-Agent Systems 35*, 30. Publisher: Springer.

Kox, E., Kerstholt, J., Hueting, T., & P. de Vries (2021). Trust repair in human-agent teams: the effectiveness of explanations and expressing regret, *Autonomous agents and multi-agent systems 35*.

Kumar, S., Savur, C., & Sahin, F. (2021). Survey of Human-Robot Collaboration in Industrial Settings: Awareness, Intelligence, and Compliance, *IEEE Transactions on Systems, Man, and Cybernetics: Systems 51*, 280–297. https://ieeexplore.ieee.org/document/9302892.

Küper, A., & Krämer, N. (2023). Psychological Traits and Appropriate Reliance: Factors Shaping Trust in AI, *International Journal of Human-Computer Interaction 0*, 1–17. Taylor & Francis _eprint:

Lascaux, A. (2008). Trust and uncertainty: a critical re-assessment, *International Review of Sociology 18*, 1–18. URL:, publisher: Routledge _eprint:

Lee, M.H., & Chew, C.J. (2023). Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making, *Proceedings of ACM Human-Computer Interaction 7*. New York, NY: Association for Computing Machinery.

Lee, J.D., & See, K.A. (2004). Trust in automation: Designing for appropriate reliance, *Human Factors 46*, 50–80. pMID: 15151155.

Lewicki, R.J., & Brinsfield, C. (2015). Trust research: measuring trust beliefs and behaviours, in: *Handbook of research methods on trust*, Edward Elgar Publishing.

Lewis, J.D., & Weigert, A. (1985). Trust as a Social Reality, *Social Forces 63*, 967–985.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy ai: From principles to practices, *ACM Computing Surveys 55*.

Luhmann, N. (2018). *Trust and power*, John Wiley & Sons.

Malle, B.F., & Ullman, D. (2023). Measuring Human-Robot Trust with the MDMT (Multi-Dimensional Measure of Trust). arXiv:2311.14887 [cs].

Mattioli, J., Sohier, H., Delaborde, A., Pedroza, G., Amokrane, K., Awadid, A., Chihani, Z., & Khalfaoui, S. (2023). Towards a holistic approach for AI trustworthiness assessment based upon aids for multi-criteria aggregation, in: G. Pedroza, X. Huang, X. C. Chen, A. Theodorou (Eds.), *SafeAI 2023 — The AAAI's Workshop on Artificial Intelligence Safety, volume 3381*, Washington, D.C.: AAAI. https://hal.science/hal-04086455

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust, *Academy of management review 20*, 709–734. Publisher: Academy of Management Briarcliff Manor, NY 10510.

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations, *Academy of management journal 38*, 24–59.

McKnight, D., & Chervany, N. (1996). The Meanings of Trust.

doi Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C. M., & Tielman, M. L. (2024). A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, *Opportunities and Challenges, ACM Journal of Responsible Computing*. just Accepted.

doi Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence 267*, 1–38.

doi Mui, L., Halberstadt, A., & Mohtashemi, M. (2003). Evaluating Reputation in Multi-agents Systems, in: R. Falcone, S. Barber, L. Korba, M. Singh (Eds.), *Trust, Reputation, and Security: Theories and Practice*, Springer, Berlin, Heidelberg, pp. 123–137.

doi Nam, C. S., & Lyons, J. B. (Eds.), Trust in Human-Robot Interaction, *Elsevier*, 2020.

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration, *Plos one 15*) e0229132. Publisher: Public Library of Science San Francisco, CA USA.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse, *Human factors 39*, 230–253. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

doi Pinyol, I., & Sabater-Mir, J. (2013). Computational trust and reputation models for open multi-agent systems: a review, *Artificial Intelligence Review 40*, 1–25.

doi Pouryousefi, S., & Tallant, J. (2023). Empirical and philosophical reflections on trust, *Journal of the American Philosophical Association 9*.

doi Ramchurn, S. D., Huynh, D., & Jennings, N. R. (2004). Trust in multi-agent systems, *Knowledge Engineering Review 19*.

doi Ramchurn, S. D., Stein, S., & Jennings, N. R. (2021). Trustworthy human-AI partnerships, *iScience 24*, 102891. https://www.sciencedirect.com/science/article/pii/S2589004221008592.

doi Riedl, R. (2022). Is trust in artificial intelligence systems related to user personality? *Review of empirical evidence and future research directions, Electronic Markets 32*, 2021–2051.

doi Riegelsberger, J., Sasse, M. A., & McCarthy, J. D. (2005). The mechanics of trust: A framework for research and design, *International Journal of Human Computer Studies 62*.

doi Rix, J. (2022). From tools to teammates: Conceptualizing humans' perception of machines as teammates with a systematic literature review, in: Proceedings of the 55th Hawaii International Conference on System Sciences.

doi Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios, in: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, pp. 101–108.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A crossdiscipline view of trust, *Academy of management review 23*, 393–404. https://www.jstor.org/stable/259285, publisher: Academy of Management Briarcliff Manor, NY 10510.

Sabater, J., & Sierra, C. (2005). Review on computational trust and reputation models, *Artificial Intelligence Review 24*, 33–60.

Sabater-Mir, J., & Vercouter, L. (2013). Trust and reputation in multiagent systems, *Multiagent systems*) 381. Publisher: MIT Press.

Sabater-Mir, J., & Vercouter, L. (2013). Trust and reputation in multiagent systems, *Multiagent systems*) 381.

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a "big five" in teamwork?, *Small group research 36*, 555–599.

Sapp, J. E., Torre, D. M., Larsen, K. L., Holmboe, E. S., & Durning, S. J. (2019). Trust in group decisions: A scoping review, *BMC Medical Education 19*.

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations, in: *Proceedings of the 28th International Conference on Intelligent User Interfaces, Iui '23*, Association for Computing Machinery, New York, NY, USA, pp. 410–422. event-place: Sydney, NSW, Australia.

Schmutz, J. B., Outland, N., Kerstan, S., Georganta, E., & Ulfert, A.-S. (2024). AI-teaming: Redefining collaboration in the digital era, *Current Opinion in Psychology 58*, 101837. https://www.sciencedirect.com/science/article/pii/S2352250X24000502.

Seraj, E., & Gombolay, M. (2020). Coordinated Control of UAVs for Human-Centered Active Sensing of Wildfires, in: *2020 American Control Conference (ACC)*, pp. 1845–1852. https://ieeexplore.ieee.org/document/9147613. iSSN: 2378-5861.

F. Santoni de Sio, & J. Van den Hoven (2018). Meaningful human control over autonomous systems: A philosophical account, *Frontiers in Robotics and AI*) 15. Publisher: Frontiers.

Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). Towards an empirically developed scale for system trust: Take two, in: *Proceedings of the human factors and ergonomics society annual meeting, volume 52*, SAGE Publications Sage CA: Los Angeles, CA, pp. 1335–1339. Issue: 19.

Stuck, R. E., Holthausen, B. E., & Walker, B. N. (2021). The role of risk in human-robot trust, in: *Trust in human-robot interaction*, Elsevier, pp. 179–194.

Surendran, V., & Wagner, A. R. (2019). Your robot is watching: Using surface cues to evaluate the trustworthiness of human actions, in: *28th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2019*, New Delhi, India, October 14–18, IEEE, pp. 1–8.

Tielman, M. L., Meyer-Vitali, A., Bailey, M., & Frattolillo, F. (2024). Multittrust: 3rd workshop on multidisciplinary perspectives on human-ai team trust, in: *Proceedings of HHAI 2024 Workshops, CEUR*. https://ceur-ws.org/Vol-3825/prefaceW5.pdf

Tolmeijer, S., Weiss, A., Hanheide, M., Lindner, F., Powers, T. M., Dixon, C., & Tielman, M. L. (2020). Taxonomy of Trust-Relevant Failures and Mitigation Strategies, in: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 3–12.

Tomlinson, E. C., & Mayer, R. C. (2009). The Role of Causal Attribution Dimensions in Trust Repair, *The Academy of Management Review 34*, 85–104. https://www.jstor.org/stable/27759987, publisher: Academy of Management.

Tucci, V., Saary, J., & Doyle, T. E. (2022). Factors influencing trust in medical artificial intelligence for healthcare professionals: a narrative review, *Journal of Medical Artificial Intelligence 5*. https://jmai.amegroups.org/article/view/6664. number: 0 Publisher: AME Publishing Company.

Ulfert, A.-S., Georganta, E., C. Centeio Jorge, Mehrotra, S., & Tielman, M. L. (2024). Shaping a multidisciplinary understanding of team trust in human-ai teams: a theoretical framework, *European Journal of Work and Organizational Psychology 33*. 158–171.

Ullman, D., & Malle, B. F. (2018). What Does it Mean to Trust a Robot? Steps Toward a Multidimensional Measure of Trust, in: *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI '18*, Association for Computing Machinery, New York, NY, USA, pp. 263–264.

Urbano, J., Rocha, A. P., & Oliveira, E. (2011) Computational trust: A review, *ACM Computing Surveys 43* 1–36.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations Can Reduce Overreliance on AI Systems During Decision-Making, *Proc. ACM Human-Computer Interaction 7*. place: New York, NY, USA Publisher: Association for Computing Machinery. ,

Verhagen, R. S., Neerincx, M. A., & Tielman, M. L. (2022). The influence of interdependence and a transparent or explainable communication style on human-robot teamwork, *Frontiers in Robotics and AI 9* 243. Publisher: Frontiers.

Verhagen, R. S., Neerincx, M. A., & Tielman, M. L. (2024). Meaningful human control and variable autonomy in human-robot teams for firefighting, *Frontiers in Robotics and AI 11*Frontiers. .

Verhagen, R. S., Marcu, A., Neerincx, M. A., & Tielman, M. L. (2024). The Influence of Interdependence on Trust Calibration in Human-Machine Teams, in: *HHAI 2024: Hybrid Human AI Systems for the Social Good*, IOS Press, pp. 300–314.

Vinanzi, S., Patacchiola, M., Chella, A., & Cangelosi, A. (2018). Would a robot trust you? developmental robotics model of trust and theory of mind, in: A. Chella, I. Infantino, A. Lieto (Eds.), *Proceedings of the 6th International Workshop on Artificial Intelligence and Cognition*, Palermo, Italy, July 2–4, 2018, volume *2418* of *CEUR Workshop Proceedings*, CEUR-WS.org p. 74. https://ceur-ws.org/Vol-2418/paper7.pdf

Visser, E. J. de, Pak, R., & Shaw, T. H. (2018) From automation to autonomy: the importance of trust repair in human-machine interaction, *Ergonomics 61* 1409–1427Taylor & Francis _eprint:

Visser, E. J. de, Marieke, M. M. Peeters, Malte, F. Jung, Kohn, S., Tyler, H. Shaw, Pak, R., & Neerincx, M. A. (2020) Towards a Theory of Longitudinal Trust Calibration in Human-Robot Teams, *International Journal of Social Robotics 12* 459–478. iSBN: 5,98,108,117,1.

Visser, E. J. de, Momen, A., Walliser, J. C., Kohn, S. C., Shaw, T. H., & Tossell, C. C. (2023). Mutually Adaptive Trust Calibration in Human-AI Teams. https://ceur-ws.org/Vol-3456/short4-8.pdf

Waa, J. van der, Diggelen, J. van, Siebert, L. Cavalcante, Neerincx, M., Jonker, & C. (2020). Allocation of Moral Decision-Making in Human-Agent Teams: A Pattern Approach, in: D. Harris, W.-C. Li (Eds.), *Engineering Psychology and Cognitive Ergonomics*. Cognition and Design, Springer International Publishing, Cham, pp. 203–220.

Wagner, A. R., Borenstein, J., & Howard, A. (2018). Overtrust in the robotic age, *Communications of the ACM 61* 22–24ACM New York, NY, USA.

Winikoff, M. (2017). Towards Trusting Autonomous Systems, in: *International Workshop on Engineering Multi-Agent Systems*, Springer, pp. 3–20.

Youssef, M., E.-N. Abdeslam, & Mohamed, D. (2015). A jade based testbed for evaluating computational trust models, in: *2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA)*, pp. 1–7.

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence, *Patterns 3* 100455.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, '20*, Association for Computing Machinery, New York, NY, USA, pp. 295–305. event-place: Barcelona, Spain.

Zhang, Q., Lee, M. L., & Carter, S. (2022). You complete me: Human-ai teams and complementary expertise, in: *Conference on Human Factors in Computing Systems — Proceedings*.

Zhang, Q., Lee, M. L., & Carter, S. (2022). You Complete Me: Human-AI Teams and Complementary Expertise, in: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22*, Association for Computing Machinery, New York, NY, USA, pp. 1–28.

Address for correspondence

Myrthe L. Tielman
Delft University of Technology
Van MourikBroekmanweg 6
2628 XE Delft
The Netherlands
m.l.tielman@tudelft.nl

Biographical notes

Dr. **Myrthe Tielman** is an assistant professor at Delft University of Technology. Her work revolves around the theme of responsible and trustworthy AI. In particular, she is interested in how we can help AI systems understand people better. For instance, to understand people's values, how much a person trusts an AI system and whether a person understands the AI system back. In my ideal world, we will have AI team mates which inspire appropriate trust, which take into account people's values, which are explainable in their decision making and which understand the capabilities and needs of the humans they work with.

https://orcid.org/0000-0002-7826-5821

Dr. **Morgan Bailey** got her Ph.D. at the Centre for Doctoral Training in Socially Intelligent Artificial Agents at the University of Glasgow. Her research focuses on advancing our compre-

hension of human-AI teams in more intricate scenarios, specifically those involving multiple humans and agents. Within these diverse team dynamics, she explores the impact of various AI presentations on dyadic trust among team members and the perceived performance of distinct AI types, such as anthropomorphic versus robotic entities. The primary objective of her work is to gain insights into these dynamics, to ensure appropriate trust calibration in the future of Human-AI Teams.

https://orcid.org/0009-0006-2626-1323

m.bailey.1@research.gla.ac.uk

Dr. **Francesco Frattolillo** received his Ph.D. in Computer Engineering from Sapienza University of Rome. His research focuses on reinforcement learning (RL) in multi-agent systems, with an emphasis on improving sample efficiency and coordination. He has developed hierarchical and model-based solutions to enhance multi-agent RL performance. Additionally, he explores RL applications in hybrid human-agent teams, where he is particularly interested in formalizing trust dynamics and designing trust-aware RL algorithms to improve collaboration and decision-making. Before, he received his MSc in Artificial Intelligence and Robotics from Sapienza.

https://orcid.org/0000-0002-2040-3355

frattolillo@diag.uniroma1.it

Carolina Centeio Jorge is a Ph.D. candidate at Delft University of Technology. Her Ph.D. focuses on mental models in the context of human-AI teams, such as enabling artificial agents to understand and predict human teammates and effectively act accordingly. In particular, she has been focusing on artificial trust, i.e., how an AI should trust a human teammate. Before, Carolina received her MSc from University of Porto, with exchange periods in Polytechnic University of Catalonia and University of Twente. After graduating, Carolina was a trainee in Computer Vision and Deep Learning at Omron Sinic X Corp., in Tokyo, Japan.

https://orcid.org/0000-0002-6937-5359

C.Jorge@tudelft.nl

Dr. **Anna-Sophie Ulfert** is an Assistant Professor of Organizational Behavior and Artificial Intelligence in the Human Performance Management Group, Department of Industrial Engineering and Innovation Sciences at Eindhoven University of Technology. Her research focuses on the design of future workplaces in which humans and Artificial Intelligence collaborate effectively. Her work focuses on trust, adaptation, and socio-technical work design in human-AI teams. Currently, she leads several international interdisciplinary projects on responsible AI implementation in organizations.

https://orcid.org/0000-0001-6293-4173

a.s.ulfert.blank@tue.nl

Dr. **André Meyer-Vitali** got his Ph.D. in software engineering, ubiquitous computing, and distributed AI from the University of Zurich. Currently, he is a senior researcher at DFKI (Saarbrücken, Germany) focused on Trusted AI and is active in the AI networks Adra, the

European Trustworthy AI Association and CAIRNE. He is principal investigator of the CERTAIN. His research interests include Software and Knowledge Engineering, Design Patterns, Neuro-Symbolic AI, Causality, Human-Agent Interaction, and Agent-based Social Simulation with the aim of creating "Trust by Design". He is also contributing to the development of standards, tools, and roadmaps for trustworthy AI.

https://orcid.org/0000-0002-5242-1443

andre.meyer-vitali@dfki.de

Publication history

Date received: 9 October 2024
Date accepted: 17 August 2025