4

c0020

Appropriate context-dependent artificial trust in human-machine teamwork*

Carolina Centeio-Jorge[a], Emma M. van Zoelen[a,b], Ruben Verhagen[a], Siddharth Mehrotra[a], Catholijn M. Jonker[a,c], and Myrthe L. Tielman[a]

[a]DELFT UNIVERSITY OF TECHNOLOGY, DELFT, THE NETHERLANDS [b]TNO, SOESTERBERG, THE NETHERLANDS [c]LEIDEN UNIVERSITY, LEIDEN, THE NETHERLANDS

s0005

1 Introduction

p0010 Artificial agents are becoming more intelligent and able to execute relevant tasks for our daily lives, including tasks in work environments, home assistance, on the battlefield, and crisis response [1]. For some of these tasks, humans and artificial agents should learn to cooperate, coordinate, and collaborate, forming *human-machine teams*. (These teams have alternative names, such as human-AI teams, human-agent teams, and human-automation teams. We use *human-machine team* in this chapter.) A key driver for achieving effective teamwork is *mutual trust* [2], that is, teammates should trust each other. In particular, we consider that *appropriate* mutual trust is a fundamental property in effective human-machine teamwork. When there is appropriate trust, there is no undertrust (leading to underreliance) or overtrust (leading to overcompliance) [3], which minimizes negative performance outcomes [4]. As such, we take appropriate to mean that a human's trust in an agent (natural trust) should correspond to that agent's trustworthiness, and an agent's trust in a human (artificial trust) should correspond to the human's trustworthiness. In fact, assessing a teammate's trustworthiness is one of the decisive factors when a person considers whether to engage in an interdependent relationship with that teammate [5]. To achieve appropriate mutual trust, we first need to understand trust and how to implement and measure it in the artificial teammates. This chapter is meant to explore how artificial agents can appropriately trust their human teammates. We explore what trust and trustworthiness mean in the context of human-machine teams, which construct the artificial agent needs to understand to reason about trust, and how these constructs can be estimated from interaction. Q1

np0015 *This document is the result of the research project funded by AI*MAN lab from TU Delft AI Initiative.

Putting AI in the Critical Loop. https://doi.org/10.1016/B978-0-443-15988-6.00007-8

p0015 Artificial agents (referred to as the cognitive part of the machine) need to be able to observe, direct, and predict teammates [6] in order to make decisions and ensure effective human-machine teamwork. We argue that between observing and predicting a human teammate there is a process of assessing the human trustworthiness, which we call artificial trust. In this process, artificial agents model the krypta of the teammates (model of their internal characteristics) through the accessible manifesta (behavioral cues of the teammates) [7, 8]. The krypta can then be used to form their beliefs of artificial trust, that is, *competence* belief and *willingness* belief [9]. What the krypta should be, how it is present in the manifesta, and then how it transforms into formal beliefs for the artificial agent are not trivial, but we can explore human-human models of trust as a first step [10]. In the literature, we can find trust models in human-human teams, such as the ABI model [11], which may be suitable for krypta of human teammates in human-machine teams. Once we know what the krypta should be, we can work on manifesta to learn the dimensions of the krypta by interaction (e.g., prior task performance as manifesta of the krypta's ability). Depending on the situation, however, it may not be possible to build the krypta over extensive and frequent interactions. In this case, we can consider different ways of assessing trust, for example, with *swift trust* [12], which relies mostly on first interaction. By assessing trust, artificial agents can then decide on whether or not to trust a human for a certain task and act accordingly (by helping the human, e.g., mitigating risks and ensuring the team's goal). Engaging (or not) in a trusting action involves risks and it is also an important part of the decision-making, that is, after knowing how much I trust someone, I still have to decide whether I should engage in a trusting action. This decision, as well as the trust assessment itself, depends on the context.

p0020 Trust is then context-dependent. In human-machine teamwork, this context can be composed of task and team configuration. We followed the taxonomy presented by Parashar et al. [13] to reflect on the different characteristics of context that can affect trust (we particularly look at the teammate's krypta that the agent should reason about). This taxonomy aims at characterizing human-robot interactions in a teamwork setting and is illustrated with the examples of the urban search and rescue (USAR) domain as well as the assembly line manufacturing setting. Navigating from one to the other, we can also explore how trust models are sensitive to the context. Both manifesta and krypta are highly dependent on this situation characterization, that is, what is important to observe and reason about in USAR and in manufacturing setting may differ based on the characteristics of the context. In a USAR setting, for example, the task may require integrity from the trustee given that moral decisions may be required, whereas in an assembly line perhaps ability may be the only important aspect to consider when trusting a teammate. Moreover, much of the existing work on human-machine trust has the goal of defining one model of trust that fits any situation.

p0025 In this chapter, we argue that it could be useful not to take a "one-model-fits-all" approach, and instead see (1) how different trust models might accommodate different contexts, (2) how within one model some dimensions may be more relevant than others, and (3) how we can start formalizing trust as a belief of context-dependent trustworthiness. Thus the chapter is structured as follows: we start by presenting the definition of trust for this chapter in Section 2, and then we go through the related work and important concepts required to understand the rest of the chapter in Section 3; we present our taxonomy of context-dependent trust in Section 4, and explore a possible formalization of the beliefs of trust in Section 5; we finally discuss the main findings in Section 6 and conclude in Section 7.

s0010 2 Trust definition

p0030 Trust is a dyadic attitude or behavior between a trustor (the one who trusts) and a trustee (the entity being trusted) and it can be defined as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (p. 712) [11]. In a team composed of both humans and machines, we need to write this definition in a more formal way in order to implement and measure it. Thus we approach trust from a functional perspective, in which trust is a relational construct between the trustor x, the trustee y, about a defined (more or less specialized) task (τ), as in Falcone et al. [7]. Particularly, we propose that trust is one agent's *perception of the trustworthiness* of another, meaning that how much x trusts y depends on how trustworthy x believes y is. This means x appropriately trusts y when x's belief in y's trustworthiness actually corresponds to y's trustworthiness. For example, if an agent x trusts another agent y to execute a task (e.g., driving a car) that requires skills that y does not have, agent x overtrusts agent y and the consequences can be negative and even disastrous (e.g., car accident). On the other hand, if agent x does not trust agent y to execute a task (e.g., driving a car) and agent y is perfectly capable of successfully executing the task, agent x is undertrusting agent y, which can also negatively affect team effectiveness (e.g., walking instead). In particular, when x is a human and y is an artificial agent, and trust is not appropriate, this will lead to disuse or misuse of technology [1]. Thus a dyadic relationship between a human and an artificial agent in a human-machine team should be designed in such a way that it supports (1) appropriate trust from the human toward the agent and (2) appropriate trust of the agent toward the human. As such, we need the artificial agent to understand trust, how to form these beliefs and how others form these beliefs. In particular, artificial agents need to have models of trustworthiness of their teammates.

s0015 3 Trust models, Krypta and Manifesta

s0020 3.1 Models

p0035 Trust has been vastly explored in the context of human-human interaction, with well-known contributions such as the ABI model [11] (which suggests trustworthiness is based Q2 on ability, benevolence, and integrity), for organizational behavior. In particular, trust in human teams has been recently explored in contexts such as virtual teams [14], sports [15], and university group projects [16]. Furthermore, in multiagent systems (MAS) trust has been used as a security and control mechanism, to protect agents from not knowing other agents' code of conduct [17]. Among others, we can find a formalization for trust and reputation (e.g., [18]), ways of categorizing agents to explain internal qualities (krypta) with their observable signs (manifesta) in order to promote trust (e.g., [7, 19]), and, more recently, models for assessing an agent's trust based on human values (e.g., [20, 21]). Similarly, trust in human-machine interaction has been gaining increasing attention. We can consider the most consensual model of human trust in technology as being Performance, Process, and Purpose [22]. Moreover, there are works on the dynamics of human trust Q3 toward technology (e.g., [23, 24]), how agents can assess and promote appropriate trust in humans (e.g., [25–29]), and the role of (appropriate) trust in human-machine teams (e.g., [3, 30–32]). There are also contributions on artificial trust, such as how an artificial agent can detect that a situation requires trust [33, 34] and also how an artificial agent can detect whether a human is being trustworthy, based on episodic memory [35] and social cues [36]. Furthermore, Azevedo-Sa et al. [37] suggest a model of trust prediction in human-machine teams, based on capabilities and task requirements. Essentially, we can find in the literature: (1) how humans trust humans, (2) how agents can trust other agents, (3) how humans trust artificial agents, and (4) how artificial agents can calibrate this trust with certain actions. Nonetheless, we found the literature on trust from the perspective of an agent toward a human to be scarce. Making artificial agents able to detect under which situations they could use trust and when they can trust a human, based on social cues, memory, and capabilities, is of utmost importance. Enabling them to understand human trustworthiness and its dimensions can lead to a better human-machine understanding and team effectiveness.

s0025 3.2 Manifesta and Krypta

p0040 We assess trustworthiness through available cues (manifesta) [38]. Often, the manifesta are cues of certain internal qualities (krypta), which are dimensions of trustworthiness. Frequently, it is suggested that human trustworthiness has as dimensions the krypta of ability, benevolence, and integrity (ABI model [11]). In the literature, we can find instruments that follow the ABI model and measure, through questionnaires, propensity to trust [39], and perceived trustworthiness (of teammates) in military teams [40, 41]. Although we can find instruments to measure trust subjectively, it is in our interest to measure trust using objective measures that can be used in interaction. So far, studies have measured trust objectively through physiological signals, such as electroencephalography (EEG) and electrocardiography (ECG) and, sometimes, audio and electrooculography (EOG) [42], which are not ecologically valid in human-machine teamwork. Breuer et al. [14] present a taxonomy of behaviors that affect how teammates perceive each other's trustworthiness in virtual human teams, although it is not clear how this taxonomy can transfer to human-machine teams. Finally, some works have presented how the krypta can be computed from the manifesta (once we know which manifesta and krypta suit our domain), such as POMDPs [29], dynamic-Bayesian networks [43], machine theory of mind [44], and instance-based learning [45]. None of these methods have been successfully tested when interacting with humans in order to estimate their trustworthiness.

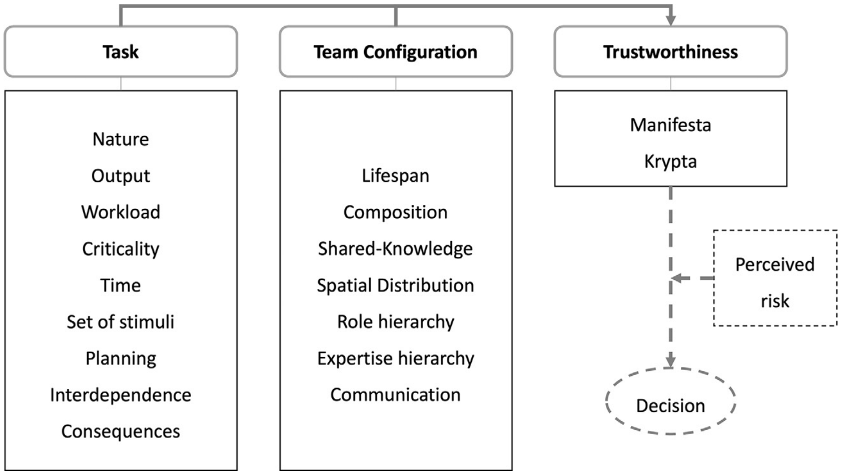
s0030 3.3 Context-dependent models and their dimensions

p0045 Most of the existing models represent trust as something that develops between people over time and is built over a series of shared experiences and interactions [12, 46]. The ABI model [11] considers trust to be, besides the trustor's propensity to trust and other external factors, the perception of trustworthiness as ability, benevolence, and integrity. Ability comprises the set of skills and competences of the trustee; benevolence has to do with the relationship between the trustor and the trustee and whether the trustee is believed to want the trustor's good; finally, integrity deals with the set of principles and moral values that the trustee adopts and whether the trustor finds them acceptable. However, there are certain situations in life, including in human-machine teams, in which time is not a given nor are some of the dynamics that allow trustors to interact and share experiences sufficiently to build their trust in such detail. Certain situations require *swift trust*. Swift trust usually happens in situations that are temporary and that may require some level of urgency [12]. This type of trust model does not build after an extensive observation, but is rather built at first (based, e.g., on imported information, propensity to trust, and surface-level cues [47]) and fine-tuned later, through interaction and observation. Although these are not the only two models of human trust in organizational settings, they show that, depending on the context, the relevant internal characteristics of teammates (krypta) may differ, as well as how they will show through behavioral cues (manifesta). Similarly, although we do not know if these models can be used for artificial trust, that is, an artificial agent trusting a human, this is the closest we have so far. In order to reason about these models and build them from interaction, an agent needs to understand both the relevant *krypta* (which would be the dimensions of trustworthiness in this case) and the *manifesta*, that is, the cues that an artificial agent can perceive in order to build the teammates' krypta. Which krypta and manifesta are important for each situation is not trivial, but we suggest we need to study them within a well-characterized context.

s0035 4 Trust as a context-dependent model

p0050 Artificial agents need to build beliefs regarding their teammates' and their own trustworthiness. We have seen before that there are models presented in the literature that represent trust in different situations and contexts. In particular, there are models of how humans trust other humans, such as the ABI model and Swift trust model, as well as how humans trust machines, such as Performance, Process, and Purpose. When we consider these models in human-machine settings, it is not trivial which of these models fits the situation best. In particular, these models are constructs, that is, they contain several different conceptual dimensions, and we might wonder whether all of these dimensions are equally important. Taking the ABI model as an example and its three antecedents as dimensions, namely, ability, benevolence, and integrity, one can wonder whether integrity is relevant (or as relevant as other antecedents) in tasks such as assembly lines in manufacturing. Moreover, the literature still does not provide trust models that artificial agents can use to assess their teammates' and their own trustworthiness in the context of teamwork. We do not know to which krypta are important for each situation. Consequently, we also do not know which manifesta the artificial agent should pay attention to build its mental models. Nevertheless, we suggest that we may not be able to find a general krypta and manifesta for all human-machine interactions. This being said, we should start by characterizing well the context in which the agent needs to assess trust in their teammates. In this section, we reflect on which characteristics of the context, including task and team configuration, may affect the krypta that the artificial agent should build to assess trustworthiness.

p0055 In the literature, we can find a taxonomy of the interactions of human-robot teams by Parashar et al. [13], which comprehends characteristics of tasks and team configuration. Departing from this chapter, and making use of the illustrative examples it provides (USAR and assembly line), we have built a taxonomy that can be used to describe a situation when an artificial agent needs to trust a human during human-machine teamwork, which can be found in Fig. 4.1. We have also included certain concepts from inspirations of other papers, such as *set of stimuli* and *time* from Farina et al. [48], *workload* from Neerincx et al. [49], *lifespan* from Haring et al. [12], and *nature* and *output* from Farina et al. [48],



f0005 **FIG. 4.1** Taxonomy to characterize situation for which an artificial agent needs to assess trust in human-machine teams. Characteristics of task and team configuration influence trustworthiness's components of Krypta and Manifesta. The assessed trustworthiness will contribute to the decision on whether to engage in a trusting action (trust decision) after a risk evaluation (perceived risk).

Wildman et al. [50], and McGrath [51]. According to our interpretation, task characteristics comprise the basic information required to distinguish one task from the other, such as type of output required, the expected time, etc. On the other hand, team configuration consists of the information regarding the team that will execute the task or the set of tasks and their dynamics, for example, the *lifespan* of the team can be 2 months for a certain project, irrespective of the tasks and their *time* that will be involved in the same project.

s0040 4.1 Task

p0060 We start by reflecting on the **Nature** of the task and its impact when assessing a teammate's trustworthiness. Although nature is quite generic, we would like to make the distinction between *cognitive* and *physical* tasks. The choice of the model of trust may depend on its nature, in particular on the manifesta expected from one another. This means that the visual cues that an artificial agent can use to, for example, conclude whether a teammate is able and/or benevolent, may vary considerably depending on the nature of the task. As nature is still a broad concept, we find it important to include the task's **Output**. In Parashar et al. [13], we can find *focus*, with examples of transit, area coverage, management, etc. Output is in that sense similar to focus but has the intention of being more measurable and implementable. For example, instead of *management* as a possibility, we intend to have a specification such as "Allocation of three tasks among the team for today." The complexity of the task as well as some of the necessary skills to be successful at it should be expressed in the output. For example, verbs of Bloom's taxonomy for educational purposes [52] could be used to better express the type of task (e.g., build, rate, choose). **Workload** classifies a task in terms of the amount of work it requires. Certainly workload varies from person to person, but it can still be useful to characterize different tasks, for example, in terms of cognitive load [49, 53].

p0065 The trust decision is highly related to the perceived risk [11]. As such, we consider **Criticality** (*low, medium, high, severe*) to be important for task characterization. This criticality concerns the risks involved with the performance of the task or the failure of such task, for example, choosing who to help in a USAR situation. In such scenarios, some constructs of trust may be more important, for example, integrity or high ability, whereas in an assembly line, if there is a task of low criticality, perhaps the ability and integrity required may be lower. When talking of criticality, we automatically think of *urgency*, which is criticality in terms of **Time**. The timeframe set for a task, that is, when it has to be done, is also important when assessing trust. We speculate that more important than assessing trust, a task's timeframe may play a major role when deciding whether to engage in the risk-taking relationship (trust decision) [5]. This means that, for example, we may decide to trust someone for an urgent task (e.g., carry a victim) that we would not trust if there was more time (perhaps we would do it ourselves instead). This characteristic also plays with the risk of not trusting being higher than trusting. Finally, we included **Set of stimuli** so that we can have a measure of motivation to successfully do a certain task or engagement while doing the task. This characteristic may be a hard one to quantify, as the present stimuli may not be obvious. However, aligned with output and workload, it can give us a better sense of the difficulty and complexity of the task. We chose stimuli to include the features of a task that can instinctively create some reaction on the agent, in particular the human. These can range from the objects involved in the task (e.g., lights, screws, robots) to social stimuli (e.g., as other people involved, a baby, and even a social robot), music, etc. [54]. We speculate that the set of stimuli can be very important to assessing overall competence and willingness of a teammate to perform a certain task, given that it may influence their engagement. Furthermore, we can even reflect on whether certain types of stimuli, for example, social stimuli, can be important to understand the relevance of benevolence in a task.

p0070 **Planning** (*online*, *offline*, *hybrid*) has to do with how the task is prepared and how much is ad hoc or improvised. The assessment of trust may be different when this characteristic changes. For example, one's assessment of trust in a USAR situation where we do not know what to expect and, as such, are not able to plan it beforehand, may change considerably in an assembly line situation where everything is planned from start to end. Additionally, it is the joint nature of key tasks that defines teamwork, which is only possible through the effective management of **Interdependence** [6] (*none*, *soft*, *hard*), that is, "the set of complementary relationships that two or more parties rely on to manage required (hard) or opportunistic (soft) dependencies in joint activity" (p. 3). We are unsure how and whether interdependence affects the trust assessment, but we can imagine that it affects the engagement in risk-taking relationships. For example, by knowing that someone's action will affect mine, that may lead me to engage in the risk-taking relationship even though I would not trust them if our actions were independent, or the opposite. Last but not least, we speculate that the **Consequence** of the task influences the trust decision. Consequence differs from output in the sense that it entails what happens to the system after the task is completed (or failed), that is, it is the direct consequence of the output. In video games, we can illustrate consequences in more measurable ways, such as rewards, levels, etc. In human-machine teams, these consequences may take more complex rewards, such as saving lives (in USAR), items to sell (in assembly lines), and so on. However, the consequences of a task may also be social, such as getting closer to someone by the means of helping them, or feeling we did what was right. These consequences can influence the weight of dimensions such as benevolence and/or integrity, respectively. Ultimately, consequences of a task are related to personal risk and reward. Although consequences may not be easy to quantify or measure, we believe we can categorize tasks as *high-risk-high-reward*, *high-risk-low-reward*, *low-risk-high-reward*, and *low-risk-low-reward.*

s0045 4.2 Team configuration

p0075 The **Lifespan** of a team is very important when choosing the trust model to use. It is important to note that this is not the **Time** of the task, for example, if I am in a group project at university for a semester, the lifespan of the team is a semester, but probably we will have tasks that take a different time, for example, preparing a presentation that takes us 2 days. In a team with a short lifespan, trust forms in a very different way than when one has time to actually get to know their teammates. Take the example of the USAR context; one may be in the field with people they have not met before (e.g., they came to reinforce help after a terrible catastrophe), or it could be daily life at a manufacturing assembly line, where they see their colleagues every day during their shift, over a few months. We do not expect trust to develop in the same way in both situations and one reason is the lifespan, since trust does not have the same time to develop. For example, in the case of USAR, one will probably not have enough time to form beliefs regarding another's benevolence and/or integrity. Or perhaps those dimensions of trustworthiness are just not important in this case, since the team will dissolve once the mission is accomplished. Swift trust models may be more suitable to teams with low lifespan, whereas ABI can suit long-term teams better.

p0080 On the other hand, team **Composition** (*single human to single machine*, *multihuman to multimachine*, etc.) may affect the type of model we use to assess a teammate's trustworthiness, given its nature and overall context of the team. Moreover, it will affect the overall team trust model, that is, the trust we have in the team, despite (not necessarily independent of) the trust we have in the teammates individually, see, for example, Ulfert et al. [30]. In human-machine teams, it is also important to consider the **Shared-Knowledge** (*independent*, *partially independent*, *overlap*), which may affect the way we assess trust regarding a certain teammate. For example, if I know everything my teammate knows by default, that is, we both have access to the same information, then the way I build my trust in them will be different than in a situation where we each have access to different information. It is also important to characterize how much, what, and among whom knowledge in a team should be shared [55]. This knowledge may comprise information regarding ontologies (e.g., domain, team member, and organization), world state (e.g., map and task sequence), and team member's models (e.g., their availability, capabilities, etc.) [56].

p0085 The **Spatial Distribution** (*proximal*, *remote*, *hybrid*) of the team can also change the way we perceive trust. As humans we know that having in-person meetings is simply different than having online meetings, whatever you may prefer. With remote work being more and more part of our lives, some people prefer working remotely, while others prefer to be physically present in the office, while there is yet another group of people that prefers a hybrid setting, which includes both spatial distributions (remote and in person) [57]. In fact, spatial distribution affects people's satisfaction, performance, and productivity [58], and it can also affect trust [59, 60]. Proximal distribution allows team members to perceive each other's characteristics differently. In particular, in human-robot teams anthropomorphic features of robots, for example, also change the way we perceive them and, consequently, how we trust them [61]. Furthermore, when we trust someone in our team, the **Role Hierarchy** (*supervisor*, *peer*, *mentor*) and **Expertise Hierarchy** (*fixed*, *fluid*) are important to consider. In terms of roles, we wonder whether one assesses the trustworthiness of their subordinates in the same way one would do for a superior. Or whether, if in different roles we simply expect different constructs and our trust model is bound to that. There are recent contributions suggesting structured roles in human-machine teams, including the roles of coordinator, creator, perfectionist, and doer [62]. It may be that even without a necessary hierarchy we construct our trust toward teammates in these different roles differently, for example, it may be that one expects higher integrity of a coordinator, given their authority, than a doer. It can also happen that the dynamics of trust change when there is hierarchy in terms of role or expertise. For example, in a surgery situation where a machine is used, we can imagine that the way a nurse trusts the machine may depend on how the doctor trusts the machine, or vice versa. We can argue that in that case there is trust transitivity [63] and that constructs such as integrity may play a bigger role than, for example, benevolence or ability. Finally, **Communication** (*environment-based*, *sensing-based*, *direct-partial*, *direct-full)* is highly related with trust in the sense that we depend on it to build our trust models [2]. The way we communicate may affect not only how we perceive the krypta but also which krypta we end up building. Communication is related to shared-knowledge as well [2], meaning that we can share knowledge through communication and build shared knowledge from communication.

s0050 4.3 Summary: A taxonomy

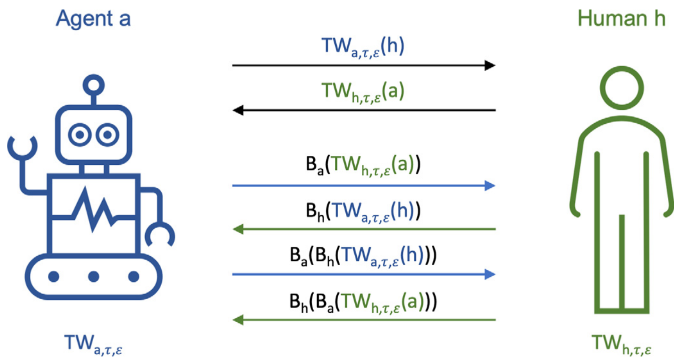
p0090 In this section, we have explored several characteristics of the context that may affect how an artificial agent should assess trust. The resulting taxonomy (in Fig. 4.1) proposes a set of important characteristics that influence the choice of trustworthiness model. In particular, we looked at task and team configuration and reflected on which characteristics may influence the choice of krypta (model containing the internal characteristics of the teammate that are relevant to assess trustworthiness) and the manifesta (behavioral cues of the teammate that hint to their krypta) to appropriately estimate a teammate's trustworthiness.

p0095 Certain task and team configuration characteristics may not only impact the estimation of trustworthiness but also the decision to trust, that is, engage in a trusting action. The decision on whether to engage in a trusting relationship is dependent on the risk that decision represents. It is important to note that the decision on whether to engage in a trusting relationship may have risks for both positive and negative decisions. This means that sometimes it may be riskier not to trust than to trust.

p0100 The taxonomy is proposed as a tool to choose krypta and manifesta, in order to assess trustworthiness appropriately. Once krypta and manifesta are chosen, it is important to formalize trust so that it can be implemented in the artificial agents. Trust must be formalized as a belief of trustworthiness, which is a construct dependent on task, team configuration, trustor, and trustee. In the next section, we propose a general formalization of these beliefs and reflect on what it takes to make them appropriate.

s0055 5 Trust as a belief of trustworthiness

p0105 We have seen that assessing trust is not trivial. In particular, trustworthiness is a complex concept, and following the literature it can consist of a set of dimensions that range from the trustee's competence to their intentions [64]. In the previous section,


Agent a
Human h
$TW_{a,\tau,\varepsilon}(h)$
$TW_{h,\tau,\varepsilon}(a)$
$B_a(TW_{h,\tau,\varepsilon}(a))$
$B_h(TW_{a,\tau,\varepsilon}(h))$
$B_a(B_h(TW_{a,\tau,\varepsilon}(h)))$
$B_h(B_a(TW_{h,\tau,\varepsilon}(a)))$
$TW_{a,\tau,\varepsilon}$
$TW_{h,\tau,\varepsilon}$

f0010 **FIG. 4.2** Human-machine dyadic trust. *An adaptation from C. Centeio Jorge, S. Mehrotra, C.M. Jonker, M.L. Tielman, Trust should correspond to trustworthiness: a formalization of appropriate mutual trust in human-agent teams, in: D. Wang, R. Falcone, J. Zhang (Eds.), Proceedings of the 22nd International Workshop on Trust in Agent Societies (TRUST 2021) Co-located with the 20th International Conferences on Autonomous Agents and Multiagent Systems (AAMAS 2021), London, UK, May 3–7, 2021, CEUR Workshop Proceedings, vol. 3022, CEUR-WS.org, 2021, http://ceur-ws.org/Vol-3022/paper4.pdf.*

we reflected on how trustworthiness can be context-dependent, including the manifesta (behavioral cues that show trustworthiness) and krypta (the construct that defines trustworthiness in a situation). Examples of krypta are *ability*, *benevolence*, and *integrity* (ABI) model [11] or *willingness*, *competence* [65]. It is important to formalize trustworthiness as beliefs of the artificial agent so that it can use them to make decisions. The artificial agent can then form beliefs of trustworthiness regarding the adequate (in terms of context) krypta from the manifesta and other possible indirect sources of information, such as overall reputation of another teammate [17]. In this section, we propose a first step toward modeling the beliefs of trustworthiness, taking into account its possible krypta dimensions and context dependencies. As several of krypta's dimensions (e.g., Benevolence, Integrity), may relate to both trustor and trustee, we stipulate that we need to define the trust of an agent x in agent y as the belief $\mathcal{B}$ of agent x regarding the trustworthiness of y with respect to x:

$$T(x, y, \tau, \epsilon) = \mathcal{B}_x(\mathcal{TW}_{y,\tau,\epsilon}(x)) \tag{4.1}$$

where τ is the *task* and ϵ is the *team configuration* as explored in the previous section. Q4 Fig. 4.2 schematizes a dyadic human-machine relationship.

s0060 5.1 Forming (appropriate) artificial trust

p0115 As an example, let us consider the task of driving a car. Inspired by Mecacci and de Sio [66], let's imagine a dual-mode vehicle, which can be driven both by an artificial agent or by a human. The default setting is the human driving according to the agent's instructions, but the agent takes over when it recognizes dangerous situations. Although it may be counter-intuitive, we need the agent to *trust* the human to drive safely (their joint goal), while complying with societal ethics, so that it knows when to take over. In this example, we will have the trustworthiness of the agent a, given a human h, $\mathcal{TW}_{a,\tau,\epsilon}(h)$, and the trustworthiness of the human h given an artificial agent a, $\mathcal{TW}_{h,\tau,\epsilon}(a)$ in a certain context, that is, task τ and team configuration ϵ. In practical terms, this means that the way the human is going to follow the agent's instructions may vary according to the agent that is helping (e.g., depending on whether the human relies on this particular agent's knowledge/intelligence). Moreover, we have the trust of the artificial agent in the human, meaning the agent's belief in the human's trustworthiness, $T(a,h,\tau,\epsilon)=\mathcal{B}_a(\mathcal{TW}_{h,\tau,\epsilon}(a))$ (from Expression 1), and the trust of the human in the agent, which is the human's belief on agent's Q5 trustworthiness $T(h,a,\tau,\epsilon)=\mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h))$. The trust of the artificial agent in the human $(T(a, h, \tau, \epsilon))$ is what the agent believes that the human will do if the agent gives the human a certain instruction.

p0120 In order to estimate $\mathcal{B}_a(\mathcal{TW}_{h,\tau,\epsilon}(a))$, we may also need the agent's belief in the human's trust in the agent, that is, $\mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h)))$, since some dimensions of trust (such as benevolence) depend both on the trustor and trustee, by definition. Following the example, for the agent to trust the human to follow an instruction, the agent needs to believe that the human trusts the agent (e.g., the human relies on this particular agent's knowledge/intelligence). Finally, when estimating whether it can trust its human teammate to follow an instruction, the *agent*'s trust in the human should correspond to the actual human's trustworthiness (e.g., to what actually the human can and/or wants to do), that is,

$$T(a,h,\tau,\epsilon)\equiv\mathcal{B}_a(\mathcal{TW}_{h,\tau,\epsilon}(a))\equiv\mathcal{TW}_{h,\tau,\epsilon}(a) \tag{4.2}$$

which requires that the agent also accurately estimates the human's trust in the agent, $T(h,a,\tau,\epsilon)\equiv\mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h)))$. The *human*'s trust in the agent, on the other hand, is the belief of the human in the agent's trustworthiness, $\mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h))$, and should correspond to the agent's actual trustworthiness $(\mathcal{TW}_{a,\tau,\epsilon}(h))$, that is,

$$T(h,a,\tau,\epsilon)\equiv\mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h)))\equiv\mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h))\equiv\mathcal{TW}_{a,\tau,\epsilon}(h) \tag{4.3}$$

What's more, we argue that appropriateness of trust depends on the task and team configuration, meaning that one may trust another in a certain context but not in another, for example, a may appropriately trust h to drive a car (a believes that h can drive a car) but not to pilot a plane (e.g., a believes that h can pilot a plane but h actually cannot). We do not illustrate possible different team configurations for the sake of simplification, but that is definitely to consider as well.

p0125 Lastly, since the nested concepts presented in Expression 3 are based on $\mathcal{TW}_{a,\tau,\epsilon}(h)$, this Q6 means that we may be able to calibrate the human's trust in the agent $(T(h, a, \tau, \epsilon))$, by manipulating $\mathcal{TW}_{a,\tau,\epsilon}(h)$ through the accurate belief of the agent's own trustworthiness. This means that if the agent is aware of its own trustworthiness, meaning that if the agent's belief in the agent's trustworthiness corresponds to the actual agent's trustworthiness, that is,

$$B_a(\mathcal{TW}_{a,\tau,\epsilon}(h))\equiv\mathcal{TW}_{a,\tau,\epsilon}(h) \tag{4.4}$$

the agent may be able to alter its own trustworthiness (or simply how it lets the human perceive it) and, consequently, calibrate human's trust. In our example, the agent might understand that it is not being perceived as intelligent, and start justify its instructions, possibly leading the human to trust it more.

p0130 We want to apply this formalization to existing trustworthiness models, by modeling and learning their dimensions. As discussed before, with the information regarding τ and ϵ, we can choose the krypta that best fits the situation, as well as which dimensions should have more impact when forming a belief. For example, we can consider the ABI trustworthiness model once more. We consider the trustworthiness of the human to be the weighted sum of their ability, integrity, and benevolence toward a specified task, in a certain environment. Similarly, we consider the agent's trust in the human to be the weighted sum of the agent's beliefs in the ability, benevolence, and integrity of the human, toward a specified task (τ), in a certain team configuration (ϵ). The belief in *ability* (Ab) of the human takes into account the task τ and team configuration ϵ. The belief in human's *benevolence* (Ben), however, besides the task and team configuration, also takes into account the agent, given that, by definition, benevolence has to do with the relationship between both [11]. Benevolence may also, among other things, implicitly use the belief of the human's trust in the agent, $\mathcal{B}_a(T(h,a))$, which, as previously discussed, can be expressed as $\mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_a(h)))$, since it comprises how the trustor perceives the trustee's willingness to do good to them (trustor) [11]. Finally, the belief in the human's *integrity* (I) depends on the agent, task, and environment. By definition, perception of integrity deals with how the trustor finds the trustee's values and moral principles acceptable. Thus Q7

$$\mathcal{B}_a(\mathcal{TW}_{h,\tau,\epsilon}(a)) = W \cdot [\mathcal{B}_a(Ab_{h,\tau,\epsilon}),\ \mathcal{B}_a(Ben_{h,\tau,\epsilon}(a)),\ \mathcal{B}_a(I_{h,\tau,\epsilon}(a))] \tag{4.6}$$

where W is a weight vector. This weight vector is once more dependent on context, as discussed in the previous section. Q8

s0065 5.2 Calibrating natural trust

p0135 Although we focus mostly on how to make artificial trust appropriate, that is, how an artificial agent can trust their human teammates appropriately, it is also important that the human's trust in the agent (natural trust) is also appropriate. By giving the necessary tools to the agent to reason about trust, we argue that it can also affect natural trust, once these beliefs may be nested (as illustrated in Fig. 4.2). As such, leveraging on the idea that agents reflect about their own trustworthiness, we may be able to influence humans to appropriately fine-tune their trust in them. For example, let us again consider the task of driving a car. Considering that the agent reflects about its own trustworthiness regarding its ability and willingness to drive the vehicle, the agent may then influence the human teammate to adapt to the agent's strengths and weaknesses (fine-tuning the human's trust in the agent). As such, it is important that the agent not only reflects on its own trustworthiness but that it does so considering the context, that is, the task and team configuration, since the context may influence the relevance of certain internal characteristics.

p0140 We posit that how trustworthy an agent is for a human and how a human trusts the agent (human's belief in agent's trustworthiness) in a certain context (task and team configuration) should be similar to get appropriate trust. If the belief of an agent in their own trustworthiness toward the human is different from their belief of the human's trustworthiness toward them in a certain context, then we come closer to undertrust $T(a, h, \tau, \epsilon)\downarrow$ or overtrust $T(a, h, \tau, \epsilon)\uparrow$, that is

$$\mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h)) > \mathcal{B}_a(\mathcal{TW}_{a,\tau,\epsilon}(h)) \rightarrow T(a,h,\tau,\epsilon)\uparrow \tag{4.5}$$

$$\mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h)) < \mathcal{B}_a(\mathcal{TW}_{a,\tau,\epsilon}(h)) \rightarrow T(a,h,\tau,\epsilon)\downarrow \tag{4.6}$$

Therefore, to avoid such situations, the agent's belief in their own trustworthiness should match with their belief about the belief of the human's trustworthiness in them. This will result in eliciting appropriate trust in a human from an agent's perspective. Most literature sees appropriate trust as the alignment between the perceived and actual performance of the agent by the human in terms of the agent's abilities [67] looking at "ability" as the core factor of estimating trust [68–70]—that is, *focusing upon the engineering aspect of trust*. However, as seen before, we propose to view trustworthiness as more than just ability, including *psychological aspects* [71] such as benevolence and integrity. What's more, we argue that to find this alignment, we need to first define the context as a task and team configuration. In this work, we aim to propose a first attempt on how several context-dependent dimensions of the krypta can be modeled (so that they can be learned) by an artificial agent, both to appropriately trust (artificial trust) and be trusted by a human teammate (natural trust).

s0070 6 Discussion and future work

p0145 In this chapter, we have argued that not only the way in which we trust our team members is context-dependent, but also whether we decide to engage in trusting actions is context-dependent as well. This is especially important when we talk about human-machine teams and more specifically artificial trust, that is, how artificial team members (such as machines) should estimate the trustworthiness of their human partners, making use of krypta and respective manifesta. We have based our reasoning on experience with past human-machine trust research, in which it proved difficult to determine exactly why people chose to trust or to be trustworthy at a given moment. Moreover, we have also experienced that it is not always obvious how and in which contexts human trust models (such as ABI) can be imported to human-machine scenarios. Our hypothesis is that analyzing task characteristics and team configuration can help to assess how trust models should be built up for a given scenario.

p0150 We have used existing task models and taxonomies to create a new human-machine task and team configuration taxonomy that can support such analysis. Given that many human-machine team tasks are used in the literature, but that there are few taxonomies to describe or categorize these tasks, our taxonomy provides a tool for task and team configuration design and analysis in the context of human-machine teaming research. While our aim is to provide pointers for what trust model can be used given a task and team configuration, we have currently only provided some suggestions of how different task and team configuration factors might influence the choice of model. We believe that, in order to draw clear conclusions on how task and team configuration characteristics can influence trust models, systematic empirical research is necessary, both in human-human as well as human-machine team settings. Additionally, our taxonomy presents a tool that may help in the process of designing experiments in human-machine teams. This taxonomy can also open the possibility of better describing experimental scenarios, tasks and teams, as to make research more reusable and organized. As such, our taxonomy can be used as a starting point for research, by

o0005 **1.** creating tasks that vary on specific factors in the taxonomy, such as Nature (cognitive vs. physical) or Criticality (high, medium, or low);

o0010 **2.** creating team configurations that vary on specific factors in the taxonomy, such as Lifespan (short [e.g., an hour] vs. long [e.g., a month]) or Shared-Knowledge (full shared knowledge vs. partial shared knowledge); and

o0015 **3.** setting up experiments that use different sets of krypta to model trust (and as a consequence different manifesta) in different contexts, as defined by task characteristics and team configurations.

p0175 To maintain consistency throughout such experiments, the formalization we described in Section 5 can be a starting point. Once we know what trustworthiness should look like (in terms of krypta and manifesta) in a certain human-machine teamwork scenario, we aim at implementing it on the artificial agents. The ultimate goal of exploring artificial trust is that an artificial agent can form beliefs regarding a teammate's trustworthiness and reason about these beliefs toward a better team performance, through decision-making. With such beliefs, an artificial agent can predict and direct its human teammates toward the team goal, while avoiding risks, that is, by helping the human or allocating tasks differently within a team. It is also important that the agent can reflect on its own trustworthiness in a certain context, so that it can calibrate its teammates' trust. By ensuring there is no undertrust, we aim at minimizing risks in human-machine interaction. With a set of expressions we show how this formalization can be used to implement appropriate (both artificial and natural) trust.

p0180 Our formalization provides a first step toward making it possible to implement artificial trust beliefs, which take into account the trustor, the trustee, and the context, that is, task and team configuration. In this chapter, we propose that the alignment of beliefs of trustworthiness and, consequently, the appropriateness of trust depend on the task and team configuration. This means that one may trust appropriately another in certain contexts but not in others. It is important to study how these beliefs can actually form from manifesta and how they can represent each different krypta adequately. When illustrating how this formalization can be used with a certain krypta, we have also presented an array of undefined weights. Although these weights are ultimately defined by the context, we argue that more complex structures may be appropriate to consider the different dimensions of krypta. We do not expect the learning of these beliefs, as well as their implementations, to be trivial and recognize these beliefs require further studying through simulation and human-machine experiments. The update of the beliefs of artificial trust is also something that should be addressed in future work. We find it important to mention that manifesta are not the only source of information to build the beliefs of krypta, that is, indirect information such as reputation may also have its role. However, we focus mainly on the interaction itself and we can infer krypta from observable behavior.

s0075 7 Conclusion

p0185 In this chapter, we present the concept of appropriate context-dependent artificial trust in human-machine teamwork. The goal of this chapter is twofold. First, we propose a taxonomy based on existing literature that can be used to choose the most appropriate model of human trustworthiness, when assessing artificial trust. Through this taxonomy we reflect on how, depending on several task and team configuration characteristics, the internal characteristics of a teammate (krypta) and how they show them (manifesta) can vary. We argue that we may not find one trustworthiness model that fits all the situations, but instead, a taxonomy that helps in characterizing the context and choosing the right model. This taxonomy contributes to the field of human-machine teamwork by proposing a set of characteristics that can define a certain context, which can facilitate the experimental design and definition of research questions. Second, we propose how we can formalize artificial trust as a belief of context-dependent trustworthiness. Our work provides a departure point for the implementation of artificial trust in artificial agents, which will make machines more adaptable and useful to their human teammates.

References

[1] M. Lewis, K. Sycara, P. Walker, The role of trust in human-robot interaction, in: Studies in Systems, Decision and Control, Springer International Publishing, 2018, pp. 135–159, https://doi.org/10.1007/978-3-319-64816-3_8.

[2] E. Salas, D.E. Sims, C. Burke, Is there a “Big Five” in teamwork? Small Group Res. 36 (2005) 555–599.

[3] M. Lewis, H. Li, K. Sycara, Deep learning, transparency, and trust in human robot teamwork, in: Trust in Human-Robot Interaction, Elsevier, 2020, pp. 321–352.

[4] S. Ososky, D. Schuster, E. Phillips, F. Jentsch, Building appropriate trust in human-robot teams, in: AAAI Spring Symposium: Trust and Autonomous Systems, 2013.

[5] M. Johnson, J.M. Bradshaw, Chapter 16—The role of interdependence in trust, in: C.S. Nam, J.B. Lyons (Eds.), Trust in Human-Robot Interaction, Academic Press, 2021, pp. 379–403, https://doi.org/10.1016/B978-0-12-819472-0.00016-2.

[6] M. Johnson, Coactive design: designing support for interdependence in human-robot teamwork,, 2014. Q9

[7] R. Falcone, M. Piunti, M. Venanzi, C. Castelfranchi, From manifesta to krypta: the relevance of categories for trusting others, ACM Trans. Intell. Syst. Technol. 4 (2013), https://doi.org/10.1145/2438653.2438662.

[8] M. Bacharach, D. Gambetta, Trust as type detection, in: C. Castelfranchi, Y.-H. Tan (Eds.), Trust and Deception in Virtual Societies, Springer Netherlands, Dordrecht, 2001, pp. 1–26, https://doi.org/10.1007/978-94-017-3614-5_1.

[9] R. Falcone, C. Castelfranchi, Trust dynamics: how trust is influenced by direct experiences and by trust itself, in: AAMAS, IEEE Computer Society, 2004, pp. 740–747, https://doi.org/10.1109/AAMAS.2004.10084.

[10] C. Centeio Jorge, M.L. Tielman, C.M. Jonker, Assessing artificial trust in human-agent teams: a conceptual model, in: C. Martinho, J. Dias, J. Campos, D. Heylen (Eds.), IVA'22: ACM International Conference on Intelligent Virtual Agents, Faro, Portugal, September 6–9, 2022, ACM, 2022, pp. 24:1–24:3, https://doi.org/10.1145/3514197.3549696.

[11] R.C. Mayer, J.H. Davis, F.D. Schoorman, An integrative model of organizational trust, Acad. Manag. Rev. 20 (1995) 709–734.

[12] K.S. Haring, E. Phillips, E.H. Lazzara, D. Ullman, A.L. Baker, J.R. Keebler, Applying the swift trust model to human-robot teaming, in: Trust in Human-Robot Interaction, Elsevier, 2021, pp. 407–427.

[13] P. Parashar, L.M. Sanneman, J.A. Shah, H.I. Christensen, A taxonomy for characterizing modes of interactions in goal-driven, human-robot teams, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2019, Macau, SAR, China, November 3–8, 2019, IEEE, 2019, pp. 2213–2220, https://doi.org/10.1109/IROS40897.2019.8967974.

[14] C. Breuer, J. Hüffmeier, F. Hibben, G. Hertel, Trust in teams: a taxonomy of perceived trustworthiness factors and risk-taking behaviors in face-to-face and virtual teams, Hum. Relat. 73 (2020) 3–34.

[15] H. Huynh, C.E. Johnson, H.S. Wehe, Humble coaches and their influence on players and teams: the mediating role of affect-based (but not cognition-based) trust, Psychol. Rep. 123 (2019) 1297–1315.

[16] A.M. Naber, S.C. Payne, S.S. Webber, The relative influence of trustor and trustee individual differences on peer assessments of trust, Pers. Individ. Differ. 128 (2018) 62–68, https://doi.org/10.1016/j.paid.2018.02.022.

[17] J. Sabater-Mir, L. Vercouter, Trust and reputation in multiagent systems, in: Multiagent Systems, MIT Press, 2013, p. 381.

[18] A. Herzig, E. Lorini, J.F. Hübner, L. Vercouter, A logic of trust and reputation, Logic J. IGPL 18 (2009) 214–244, https://doi.org/10.1093/jigpal/jzp077.

[19] C. Burnett, T.J. Norman, K. Sycara, Stereotypical trust and bias in dynamic multiagent systems, ACM Trans. Intell. Syst. Technol. 4 (2013), https://doi.org/10.1145/2438653.2438661.

[20] K. Chhogyal, A.C. Nayak, A. Ghose, K.H. Dam, A value-based trust assessment model for multi-agent systems, in: 28th International Joint Conference on Artificial Intelligence (IJCAI-19), 2019.

[21] C. Cruciani, A. Moretti, P. Pellizzari, Dynamic patterns in similarity-based cooperation: an agent-based investigation, J. Econ. Interact. Coord. 12 (1) (2017). Q10

[22] J.D. Lee, K.A. See, Trust in automation: designing for appropriate reliance, Hum. Factors 46 (2004) 50–80.

[23] M. Winikoff, Towards trusting autonomous systems, Lect. Notes Comput. Sci. 10738 (2018) 3–20,- https://doi.org/10.1007/978-3-319-91899-0_1.

[24] C. Nam, P. Walker, H. Li, M. Lewis, K. Sycara, Models of trust in human control of swarms with varied levels of autonomy, IEEE Trans. Hum.-Mach. Syst. 50 (2020) 194–204, https://doi.org/10.1109/THMS.2019.2896845.

[25] M.W. Floyd, M. Drinkwater, D.W. Aha, Learning trustworthy behaviors using an inverse trust metric, in: Robust Intelligence and Trust in Autonomous Systems, Springer, 2016.

[26] I.B. Ajenaghughrure, S.C. Sousa, I.J. Kosunen, D. Lamas, Predictive model to assess user trust: a psycho-physiological approach, in: Proceedings of the 10th Indian Conference on Human-Computer Interaction, 2019, pp. 1–10.

[27] Y. Guo, X.J. Yang, Modeling and predicting trust dynamics in human-robot teaming: a Bayesian inference approach, Int. J. Soc. Robot. (2020), https://doi.org/10.1007/s12369-020-00703-3.

[28] C. Neubauer, G. Gremillion, B.S. Perelman, C.L. Fleur, J.S. Metcalfe, K.E. Schaefer, Analysis of facial expressions explain affective state and trust-based decisions during interaction with autonomy, in: Advances in Intelligent Systems and Computing, Proceedings of the 3rd International Conference on Integrating People and Intelligent Systems, February 19–21, 2020, Modena, Italy, vol. 1131, Springer, 2020, pp. 999–1006, https://doi.org/10.1007/978-3-030-39512-4_152.

[29] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, S.S. Srinivasa, Planning with trust for human-robot collaboration, in: T. Kanda, S. Sabanovic, G. Hoffman, A. Tapus (Eds.), Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2018, Chicago, IL, USA, March 5–8, 2018, ACM, 2018, pp. 307–315, https://doi.org/10.1145/3171221.3171264.

[30] A.-S. Ulfert, E. Georganta, A model of team trust in human-agent teams, in: Companion Publication of the 2020 International Conference on Multimodal Interaction, Association for Computing Machinery, New York, NY, 2020, pp. 171–176, https://doi.org/10.1145/3395035.3425959.

[31] K.E. Schaefer, B.S. Perelman, G.M. Gremillion, A.R. Marathe, J.S. Metcalfe, A roadmap for developing team trust metrics for human-autonomy teams, in: Trust in Human-Robot Interaction, Academic Press, 2021, https://doi.org/10.1016/B978-0-12-819472-0.00012-5.

[32] E.J.D. Visser, M.M.M. Peeters, M.F. Jung, S. Kohn, T.H. Shaw, R. Pak, M.A. Neerincx, Towards a theory of longitudinal trust calibration in human-robot teams, Int. J. Soc. Robot. 12 (2020) 459–478, https://doi.org/10.1007/s12369-019-00596-x.

[33] A.R. Wagner, R.C. Arkin, Recognizing situations that demand trust, in: 2011 RO-MAN, IEEE, 2011, pp. 7–14.

[34] A.R. Wagner, P. Robinette, A. Howard, Modeling the human-robot trust phenomenon: a conceptual framework based on risk, ACM Trans. Interact. Intell. Syst. 8 (2018), https://doi.org/10.1145/3152890.

[35] S. Vinanzi, M. Patacchiola, A. Chella, A. Cangelosi, Would a robot trust you? Developmental robotics model of trust and theory of mind, Philos. Trans. R. Soc. B 374 (2019), https://doi.org/10.1098/rstb.2018.0032.

[36] V. Surendran, A. Wagner, Your robot is watching: using surface cues to evaluate the trustworthiness of human actions, in: 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2019, pp. 1–8.

[37] H. Azevedo-Sa, X.J. Yang, L.P. Robert, D.M. Tilbury, A unified bi-directional model for natural and artificial trust in human-robot collaboration, IEEE Robot. Autom. Lett. 6 (3) (2021) 5913–5920, https://doi.org/10.1109/LRA.2021.3088082.

[38] N. Schlicker, M. Langer, Towards warranted trust: a model on the relation between actual and perceived system trustworthiness, in: S. Schneegass, B. Pfleging, D. Kern (Eds.), MuC '21: Mensch und Computer 2021, Ingolstadt, Germany, September 5–8, 2021, ACM, 2021, pp. 325–329, https://doi.org/10.1145/3473856.3474018.

[39] R.C. Mayer, J.H. Davis, The effect of the performance appraisal system on trust for management: a field quasi-experiment, J. Appl. Psychol. 84 (1) (1999) 123.

[40] B.D. Adams, R. Webb, Trust in Small Military Teams, Command and Control Research Program, 2002.

[41] B.D. Adams, S. Waldherr, J. Sartori, Trust in Teams Scale, Trust in Leaders Scale: Manual for Administration and Analyses, 2008.

[42] I.B. Ajenaghughrure, S. Sousa, D.J.R. Lamas, Measuring trust with psychophysiological signals: a systematic mapping study of approaches used, Multimodal Technol. Interact. 4 (2020) 63.

[43] A. Xu, G. Dudek, OPTIMo: online probabilistic trust inference model for asymmetric human-robot collaborations, in: ACM/IEEE International Conference on Human-Robot Interaction, 2015, IEEE Computer Society, 2015, pp. 221–228, https://doi.org/10.1145/2696454.2696492. vol.

[44] N.C. Rabinowitz, F. Perbet, H.F. Song, C. Zhang, S.M.A. Eslami, M.M. Botvinick, Machine theory of mind, in: J.G. Dy, A. Krause (Eds.), Proceedings of Machine Learning Research, Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, vol. 80, PMLR, 2018, pp. 4215–4224.

[45] T.N. Nguyen, C. Gonzalez, Cognitive machine theory of mind, in: S. Denison, M. Mack, Y. Xu, B.C. Armstrong (Eds.), Proceedings of the 42th Annual Meeting of the Cognitive Science Society—Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual, July 29 to August 1, 2020.

[46] J.K. Rempel, J.G. Holmes, M.P. Zanna, Trust in close relationships, J. Pers. Soc. Psychol. 49 (1) (1985) 95.

[47] J.L. Wildman, M.L. Shuffler, E.H. Lazzara, S.M. Fiore, C.S. Burke, E. Salas, S. Garven, Trust development in swift starting action teams: a multilevel framework, Group Org. Manag. 37 (2) (2012) 137–170.

[48] A.J. Farina, G.R. Wheaton, E.A. Fleishman, Development of a taxonomy of human performance: the task characteristics approach to performance prediction,, 1971. Q11

[49] M.A. Neerincx, et al., Cognitive task load analysis: allocating tasks and designing support, in: Handbook of Cognitive Task Design, vol. 2003, Lawrence Erlbaum Associates, Mahwah, NJ, 2003, pp. 283–305.

[50] J.L. Wildman, A.L. Thayer, M.A. Rosen, E. Salas, J.E. Mathieu, S.R. Rayne, Task types and team-level attributes: synthesis of team classification literature, Hum. Resour. Dev. Rev. 11 (1) (2012) 97–129.

[51] J.E. McGrath, Groups: Interaction and Performance, vol. 14, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[52] B.S. Bloom, Committee of College and University Examiners, Taxonomy of Educational Objectives, vol. 2, Longmans, Greene, NY, 1964.

[53] J. Sweller, Cognitive load theory, in: Psychology of Learning and Motivation, vol. 55, Elsevier, 2011, pp. 37–76.

[54] J. Cohen-Mansfield, M.S. Marx, L.S. Freedman, H. Murad, N.G. Regier, K. Thein, M. Dakheel-Ali, The comprehensive process model of engagement, Am. J. Geriatr. Psychiatry 19 (10) (2011) 859–870.

[55] M. Harbers, C.M. Jonker, M.B. van Riemsdijk, Context-sensitive sharedness criteria for teamwork, in: A.L.C. Bazzan, M.N. Huhns, A. Lomuscio, P. Scerri (Eds.), International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS ’14, Paris, France, May 5–9, 2014, IFAAMAS/ACM, 2014, pp. 1507–1508.

[56] C.M. Jonker, M.B. van Riemsdijk, I. van de Kieft, M.L. Gini, Compositionality of team mental models in relation to sharedness and team performance, in: H. Jiang, W. Ding, M. Ali, X. Wu (Eds.), Lecture Notes in Computer Science, Advanced Research in Applied Artificial Intelligence—25th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2012, Dalian, China, June 9–12, 2012. Proceedings, vol. 7345, Springer, 2012, pp. 242–251, https://doi.org/10.1007/978-3-642-31087-4_26.

[57] Slack, Moving beyond remote: workplace transformation in the wake of Covid-19, 2020. https://slack.com/blog/collaboration/workplace-transformation-in-the-wake-of-covid-19.

[58] A. Alfaleh, A.N. Alkattan, A. Alageel, M. Salah, M.M. Almutairi, K. Sagor, K. Alabdulkareem, Onsite versus remote working: the impact on satisfaction, productivity, and performance of medical call center workers, Inquiry 58 (2021). Q12

[59] S.P. Mikawa, S.K. Cunnington, S.A. Gaskins, Removing barriers to trust in distributed teams: understanding cultural differences and strengthening social ties, in: S.R. Fussell, P.J. Hinds, T. Ishida (Eds.), Proceedings of the 2009 International Workshop on Intercultural Collaboration, IWIC ’09, Palo Alto, California, USA, February 20–21, 2009, ACM, 2009, pp. 273–276, https://doi.org/10.1145/1499224.1499275.

[60] D.S. Staples, P. Ratnasingham, Trust: the panacea of virtual management? in: J.I. DeGross, R. Hirschheim, M. Newman (Eds.), Proceedings of the Nineteenth International Conference on

Information Systems, ICIS 1998, Helsinki, Finland, December 13–16, 1998, Association for Information Systems, 1998, pp. 128–144.

[61] M. Natarajan, M.C. Gombolay, Effects of anthropomorphism and accountability on trust in human robot interaction, in: T. Belpaeme, J.E. Young, H. Gunes, L.D. Riek (Eds.), HRI '20: ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, United Kingdom, March 23–26, 2020, ACM, 2020, pp. 33–42, https://doi.org/10.1145/3319502.3374839.

[62] D. Siemon, Elaborating team roles for artificial intelligence-based teammates in human-AI collaboration, Group Decis. Negot. (2022) 1–42. Q13

[63] L. Huang, N.J. Cooke, R.S. Gutzwiller, S. Berman, E.K. Chiou, M. Demir, W. Zhang, Distributed dynamic team trust in human, artificial intelligence, and robot teaming, in: Trust in Human-Robot Interaction, Elsevier, 2021, pp. 301–319.

[64] N. Griffiths, Task delegation using experience-based multi-dimensional trust, in: AAMAS '05, 2005.

[65] C. Castelfranchi, R. Falcone, Trust & Self-Organising Socio-Technical Systems, Springer International Publishing, 2010, pp. 209–229, https://doi.org/10.1007/978-3-319-29201-4_8.

[66] G. Mecacci, F.S de Sio, Meaningful human control as reason-responsiveness: the case of dual-mode vehicles, Ethics Inf. Technol. 22 (2) (2020) 103–115, https://doi.org/10.1007/s10676-019-09519-w.

[67] J.M. McGuirl, N.B. Sarter, Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information, Hum. Factors 48 (4) (2006) 656–665.

[68] F. Yang, Z. Huang, J. Scholtz, D.L. Arendt, How do visual explanations foster end users' appropriate trust in machine learning? in: Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, pp. 189–201.

[69] F. Ekman, M. Johansson, J. Sochor, Creating appropriate trust in automated vehicle systems: a framework for HMI design, IEEE Trans. Hum.-Mach. Syst. 48 (1) (2017) 95–101.

[70] S.H. Huang, K. Bhatia, P. Abbeel, A.D. Dragan, Establishing appropriate trust via critical states, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 3929–3936.

[71] R.R. Hoffman, A taxonomy of emergent trusting in the human-machine relationship, in: Cognitive Systems Engineering: The Future for a Changing World, Talyor & Francis, Boca Raton, FL, 2017, pp. 137–163.