Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams

de Visser, Ewart J.; Peeters, Marieke M.M.; Jung, Malte F.; Kohn, Spencer; Shaw, Tyler H.; Pak, Richard; Neerincx, Mark A.

DOI
10.1007/s12369-019-00596-x

Publication date
2020

Document Version
Final published version

Published in
International Journal of Social Robotics

Citation (APA)
de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics*, *12*(2), 459-478. https://doi.org/10.1007/s12369-019-00596-x

International Journal of Social Robotics (2020) 12:459–478
https://doi.org/10.1007/s12369-019-00596-x

Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams

Ewart J. de Visser[1] · Marieke M. M. Peeters[4] · Malte F. Jung[6] · Spencer Kohn[3] · Tyler H. Shaw[3] · Richard Pak[2] · Mark A. Neerincx[4,5]

Accepted: 23 September 2019 / Published online: 23 November 2019

Abstract

The introduction of artificial teammates in the form of autonomous social robots, with fewer social abilities compared to humans, presents new challenges for human–robot team dynamics. A key characteristic of high performing human-only teams is their ability to establish, develop, and calibrate trust over long periods of time, making the establishment of longitudinal human–robot team trust calibration a crucial part of these challenges. This paper presents a novel integrative model that takes a longitudinal perspective on trust development and calibration in human–robot teams. A key new proposed factor in this model is the introduction of the concept *relationship equity*. Relationship equity is an emotional resource that predicts the degree of goodwill between two actors. Relationship equity can help predict the future health of a long-term relationship. Our model is descriptive of current trust dynamics, predictive of the impact on trust of interactions within a human–robot team, and prescriptive with respect to the types of interventions and transparency methods promoting trust calibration. We describe the interplay between team trust dynamics and the establishment of work agreements that guide and improve human–robot collaboration. Furthermore, we introduce methods for dampening (reducing overtrust) and repairing (reducing undertrust) mis-calibrated trust between team members as well as methods for transparency and explanation. We conclude with a description of the implications of our model and a research agenda to jump-start a new comprehensive research program in this area.

Keywords Relationship equity · Social autonomy · Trust repair · Trust calibration · Work agreements · Agents · Social abilities · Human–robot interaction · Collaboration · Team

1 Trust in Human–Robot Teams

The possibility of mature human–robot teams (HRTs) seems within reach with recent advances in unmanned systems, self-driving cars, and similar applications of artificial intelligence [50,132]. We define an HRT as a team consisting of at least one human and one robot, intelligent agent, and/or other AI or autonomous system. Strictly speaking, a robot is an intelligent system with a physical embodiment, yet in the context of this paper, we choose to use the term human–robot teaming to encompass a broader range of human-autonomy teaming constellations (see Table 1). Even as artificial intelligence and robotics mature to the point of ubiquitous use, the question remains how to create high performing HRTs [4,12,31,32,55,98,108,117,124,137,148,154]. A recent study showed that a key predictor of good teamwork is not about having good (technical) capabilities, but about having a way to allow for vulnerable communication in casual and non-work related interactions [37,39]. Such communication is a major facilitator of positive trust development within the team and ultimately, as the study showed, a major predictor of a team's success. On the flip side of promoting healthy trust relationships is avoiding unhealthy

✉ Ewart J. de Visser
ewartdevisser@gmail.com

[1] Warfighter Effectiveness Research Center, United States Air Force Academy, 2354 Fairchild Drive Suite 5L-50, Colorado Springs, CO 80840, USA

[2] Department of Psychology, Clemson University, 418 Brackett Hall, Clemson, SC 29634, USA

[3] George Mason University, 4400 University Drive 3F5, Fairfax, VA 22030, USA

[4] TNO, Perceptual and Cognitive Systems, Kampweg 55, 3769 DE Soesterberg, The Netherlands

[5] Delft University of Technology, Delft, The Netherlands

[6] Department of Information Science, Cornell University, Gates Hall, 220A, Ithaca, NY 14853, USA

trust relationships. Decades of industrial psychology, human factors and robotics research have shown that inappropriate or insufficient trust in another team member can have costly consequences [53,59,113,120,122,125]. Trusting too much ("overtrust") can condition operators into complacent states and misuse which can lead to costly disasters with the loss of human lives and destruction of expensive equipment [58,111,121,127]. Trusting too little ("undertrust") can cause inefficient monitoring and unbalanced workload, leading to disuse of a machine or the avoidance of a person (see Fig. 1).

Mutual trust is thus a fundamental property and predictor of high performing teams. During collaboration, team members continuously engage in a process of establishing and (re)calibrating trust among each other. We define trust 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" [96]. Since the establishment and maintenance of trust is crucial for team performance and given the projected dramatic increase in robots that support teamwork it is crucial to understand how the introduction of such systems affects team trust development and maintenance, and ultimately team performance.

1.1 The Research Challenge

While robotic systems that support teamwork have improved tremendously in the last decade, creating functioning social abilities in a robot is one of the most difficult remaining challenges [70]. A key question in the next decade will be how artificial team members can be tightly integrated into the social structure of hitherto human-only teams.

Previous research on trust in HRTs has primarily focused on identifying initial trust states and potential determinants [9,53,126]. A new approach is needed to identify what aspects of a robot's design and behavior determine the adjustment of overtrust and undertrust states over longer periods of time by analyzing the trust process within a broader perspective of longitudinal teaming. By examining trust as a calibration process between team members collaborating in various constellations throughout a range of tasks, we will understand the crucial role of trust calibration for the incremental refinement of task division, communication, and coordination among the team members.

With few exceptions (e.g. [59,87,157]), we have little understanding of the temporal dynamics of trust formation and maintenance, nor of how trust increases or decreases over time as a result of moment-to-moment interactions among HRT members. New approaches to understanding trust are therefore needed and especially those that are affectively grounded [73,128]. To understand how the introduction of social robots in a team might affect trust development and maintenance, we examine trust development in human–human teams [108,152]. Even though trust between humans and robots may not be tantamount to trust among humans [87], we may still draw insights from human–human trust development frameworks [29]. In this paper, we draw inspiration from the work by Gottman on identifying healthy and unhealthy relationship patterns [67,68]. Gottman's research charted the dynamics of trust in couples by analyzing moment-to-moment interactions over longer periods of time, and identified specific trust repair strategies to be used when trust was too low, and trust dampening strategies when trust was too high. We believe a similar approach could be fruitfully applied to HRTs when investigating how relationships with artificial teammates can evolve over longer periods of time. This is particularly relevant given recent evidence that points to similarities in how humans establish relationships with machines via the hormone oxytocin [27].

1.2 Paper Overview

1.2.1 Main Contributions

The conceptual and computational modeling of trust is a research topic that has been thoroughly researched in the past and is currently receiving much attention in various research communities. To distinguish our approach from others and to acknowledge some of the limitations of our approach we describe here what our paper does and does not address. First, and foremost, our model concerns the future autonomous *social capabilities* of robots. Several other approaches focus on task-specific trust or adaptive trust calibration approaches that measure trust passively and then adjust to the operator. Our approach is unique in the sense that it proposes a future where robots function with social human-like abilities. Second, our model serves as a of meta model to a number of isolated trust process models previously proposed. This has the advantage of allowing for a broader outlook of trust in human–robot teams and allows for a scalable approach that can describe longer-term human robot interactions. Third, although our model as it stands is not ready for computational implementation, it may be amenable to a variety of computational approaches that can be implemented. There are many models available that specify the formulas and computational models required from possible or existing implementation. Fourth, we deliberately did not specify the measurement of trust for this model. Trust is a multi-faceted concept and measurement approaches vary greatly between disciplines. The concepts proposed here are theoretical and could be interpreted and measured in a variety of ways. The strength of this approach is that our model allows for flexible and diverse application and implementation. Lastly, trust in robots involves a number of different variables. The work described in this paper focuses on relationship equity and those variables believed to affect longitudinal trust develop-

Table 1 Concept definition

Concept	Definition
Actor	An entity capable of acting, e.g. a human, robot, or agent
Team	An interdependent social group with a shared identity and common goal [123]
Human–robot team	A team of actors consisting of at least one human and one robot, intelligent agent, or autonomous system
Longitudinal teaming	Forming and developing a team over time (months to years), by engaging in continuous adaptations of the collaborative process to improve team performance
Relationship equity	An emotional resource that predicts the degree of goodwill between two actors. The resource is accumulated throughout the interaction history between the actors, referring to the cumulative positive or negative assessment of relationship acts performed by each actor
Relationship (regulation) act	An act, performed by an actor, that affects the relationship equity, either positively or negatively. A relationship *regulation* act is an act performed *deliberately* by an actor with the intent to affect the relationship equity in a certain manner, either positively or negatively
Formal work agreement	A formal and explicit agreement between two actors on how they collaborate
Informal collaboration	An informal and implicit agreement between two actors on how they collaborate
Trustworthiness	The extent to which an actor has the ability to execute relevant tasks, demonstrates integrity, and is benevolent towards fellow team members [96]
Reliance	The extent to which an actor delegates certain tasks -deemed important by said actor- to another actor expecting that actor to perform that action timely and effectively
Trust	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 [96]
Trust stance	An actor's attitude to (dis)trust another actor with regard to a particular task or situation
Trust violation	An (in)action by an actor representing a misalignment between the observed trustworthiness and current trust stance
Open trust stance	A trust stance that is open to the other actor performing a given task, representing a high level of trust. With this attitude, trustors may, for example, reveal intimate information about themselves, allow access to themselves, or comply with advice provided
Closed trust stance	A trust stance that is closed to the other actor performing a given task, representing a low level of trust. With this attitude, trustors may, for example, protect information about themselves, do not allow access to themselves, and do not comply with advice provided
Trust calibration	The process of updating the trust stance by aligning the perception of an actor's trustworthiness with its actual trustworthiness so that the prediction error is minimized [18,87]
Trust repair	Performing a behavior aimed at increasing trustworthiness and opening up another actor's trust stance toward oneself
Trust dampening	Performing a behavior aimed at reducing trustworthiness and closing down another actor's trust stance toward oneself
Mental model	An internal representation in the mind of one actor about the characteristics of another actor
Self-confidence	An actor's estimation of their own performance and capabilities
Theory of mind	An actor's (e.g. actor A) estimation of another actor's (e.g. actor B) mental model of that actor (e.g. actor A).
Miscalibration	(1) A prediction error between an actor's perceived trustworthiness of another team member and the other team member's actual trustworthiness. (2) A prediction error between an actor's self-confidence and an actor's actual trustworthiness

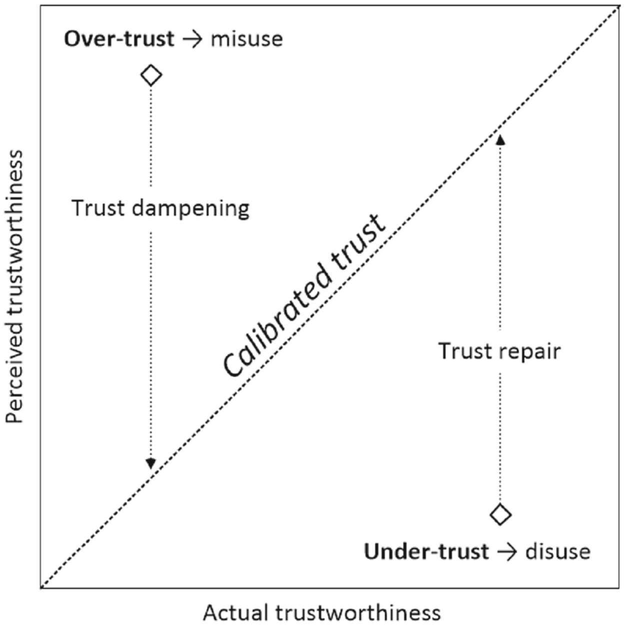

Fig. 1 Overtrust, undertrust, and calibrated trust as a function of perceived trustworthiness versus actual trustworthiness

ment. We have excluded a number of other variables that may be important in non-social situations.

1.2.2 Overview

We start by presenting the human–robot team (HRT) trust model, to describe how iterative collaboration helps team mates to incrementally construct accurate models of one another's ability, integrity, and benevolence, and how trust calibration can contribute to this process. Thereafter, throughout the rest of this paper, we go through the various elements in the HRT trust model to describe a thorough and integrated theory of how trust develops over time as a result of a series of one-shot interactions. During each one-shot interaction, the robot determines whether or not it might cause a trust violation, i.e. behave in a way that is not in line with its team member's trust stance. This allows the robot to engage in active trust calibration by using a social signal detection theory. Through the application of our presented design guidelines, the designer of the robot may determine what type of behaviour the robot should use to calibrate trust either in advance of the potential trust violation, or afterwards in case of a false detection or a miss. Long-term interaction is described as a repeated series of one-shot interactions, the outcomes of which are stored in a relationship equity bank that builds up (or breaks down) depending on whether trust is either violated or complied to. The theory presented in this paper allows for a range of propositions that can be tested and validated by implementing the proposed model and investigating the effects observed when humans team with trust-calibrating robots. We state each of these propositions at the end of each described theoretical component of the model.

2 A Model for Longitudinal Trust Development in Human–Robot Teams

Figure 2 presents a new model explaining the role and process of establishing *longitudinal social trust calibration* throughout the life cycle of an HRT. The HRT Trust Model describes the development and role of trust calibration in HRT collaboration. HRT consists of four parts including 1) Relationship Equity, 2) Social Collaborative Processes, 3) Perceptions of Team Partner, and 4) Perceptions of Self.

2.1 Relationship Equity (Light Blue)

Central to our model is the idea of *relationship equity* which describes the cumulative result of the cost and benefit relationship acts that are exchanged during repeated collaborative experiences (including social and/or emotional interactions) between two actors. The concept is somewhat similar to the notion of social capital [11] and goodwill [41]. It is also somewhat inspired by equity theory as part of social exchange and interdependence theory [61]. While our concept of relationship equity is primarily the difference between the cumulative costs and benefits between two partners, equity in this theory refers to whether the ratio of relationship outcomes and contributions is equal between partners. Unbalanced ratios cause relationship distress.

2.2 Social Collaborative Processes (Red and Green)

The middle part of the model describes the collaborative task performance between the teammates. Together, they perform a joint activity with the purpose of achieving a common goal. Collaboration is risky: actions may fail and circumstances may change. Therefore, the individual actors monitor the behavior and collaboration of themselves and their teammates. Based on their observations, they aim to establish appropriate trust stances towards one another, so as to mitigate the potential risks involved in accomplishing the joint task (also see Sect. 2.3). This trust stance allows actors (both human and robot) to decide on safe and effective ways to collaborate *on the current task with the current team constellation*. Based on the trust stance, a teammate may decide to rely on a combination of formal, explicit work agreements (especially in cases where relationship equity is low) and informal, implicit collaborative agreements (especially in cases where the relationship equity is high). Both types of collaborative agreements aim to improve the team performance, for instance, by mitigating risk, compensating for one another's

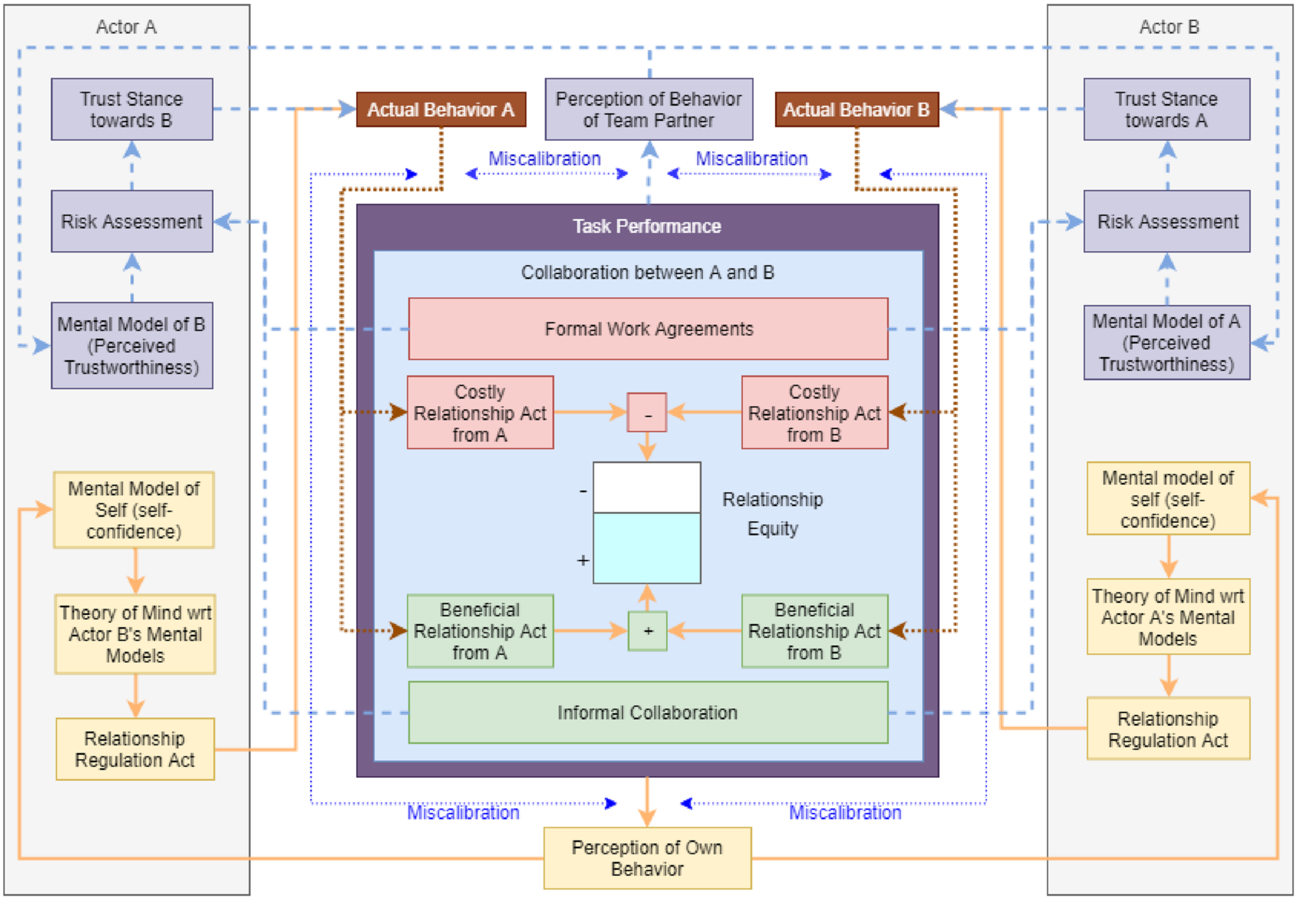
Actor A
Trust Stance towards B
Risk Assessment
Mental Model of B (Perceived Trustworthiness)
Mental Model of Self (self-confidence)
Theory of Mind wrt Actor B's Mental Models
Relationship Regulation Act
Actual Behavior A
Perception of Behavior of Team Partner
Actual Behavior B
Miscalibration
Miscalibration
Task Performance
Collaboration between A and B
Formal Work Agreements
Costly Relationship Act from A
-
Costly Relationship Act from B
-
+
Relationship Equity
Beneficial Relationship Act from A
+
Beneficial Relationship Act from B
Informal Collaboration
Miscalibration
Miscalibration
Perception of Own Behavior
Actor B
Trust Stance towards A
Risk Assessment
Mental Model of A (Perceived Trustworthiness)
Mental model of self (self-confidence)
Theory of Mind wrt Actor A's Mental Models
Relationship Regulation Act

Fig. 2 The Human–robot team (HRT) trust model. The collaboration itself is represented in the middle of the figure, describing how each action from either of the team members adds to or takes from the relationship equity bank, and how the level of this bank influences the preferred way of collaboration, i.e. through informal and implicit agreements, or through formal and explicit agreements. The blue-grey boxes represent the passive trust calibration process, whereas the yellow boxes describe the active trust calibration process

limitations, coordinating parallel activities, or communicating information relevant to the team [20,21,84,92,150].

2.3 Perceptions of Team Partner (Grey)

The blue-grey boxes indicate the *passive trust calibration process*: Based on team members' perceptions of one another, actors predict one another's trustworthiness. Taking into account their current formal work agreements and informal way of collaboration, they then (sub)consciously assess the risk involved in the collaboration as it currently is, and decide upon a trust stance towards one another [16,90]. They then may decide to adjust their collaboration to mitigate the assessed risks, for example by proposing formal work agreements or by relaxing the existing work agreements. During the next collaborative occasion, the actors obtain additional information concerning their team member's trustworthiness. This information may deviate from the original prediction, resulting in a prediction error, or miscalibration.

Adequately calibrated trust stances among the team members lead to more effective collaboration: Overtrust can condition team members into complacent states and misuse, whereas undertrust can cause inefficient monitoring and unbalanced workload. In other words, trust calibration is crucial for optimal team performance. Through the feedback loops described in the model, the HRT trust process leads to continuous incremental updates of the team members' trust stances towards one another and an overall reduction of miscalibrations. We assume that, for team members that are benevolent and sincere, the development of appropriate trust stances will benefit their collaborative efforts; team members can compensate for each others' flaws, while relying on each others' strengths.

2.4 Perceptions of Self (Yellow)

The yellow boxes indicate the *active trust calibration process*: This process is based on an actor's awareness concerning their involvement in team trust calibration. This awareness enables both actors to engage in deliberate attempts to influence, aid, or hamper the trust calibration process. This is achieved first and foremost through the formation of a theory of mind, allowing an actor to reason about the other actors' mental models. If the actor concludes, based on their self-confidence and their theory of mind, that another team member may be mistaken about their performance level, the actor may decide to actively intervene in the trust calibration process, through a relationship regulation act, such as an explanation or an apology.

The next few sections describe the various parts presented by the HRT trust model in more detail.

3 Relationship Equity: Benefits of Building Trust Over Time

Relationship equity represents the interaction history between two actors and is the cumulative positive or negative assessment with respect to the relationship between the actors. Relationship equity affects future perceptions of trustworthiness and the trust stance by functioning as a lens through which future interactions are perceived and interpreted. Relationships that have accumulated a lot of positive equity may be able to absorb trust violations, without stirring the relationship equity all that much. Alternatively, relationships showing negative equity may be rattled even by small trust violations compared to relations that have positive equity. We believe relationship equity is a critical construct that is needed to predict long-term human–robot interaction. The relationship equity between team mates is influenced not only by the collaborative experience itself but can also be actively and deliberately affected through relationship regulation acts, as we will see during the last two steps in the HRT trust model.

The feedback loops (passive and active trust calibration) presented in Fig. 2 occur continuously while interacting with other actors. The feedback obtained from these loops is remembered and stored in what we propose as a new construct known as *relationship equity* (also see Fig. 3).

The first part of this section focuses on healthy human relationships that are maintained by partners who actively engage in relationship regulation acts to contribute to relationship equity [67,68]. Research on human–robot interaction has applied some of the core concepts related to this work to HRTs to explore the similarities and differences [24–27].

3.1 Emotion Regulation as the Key Activity to Build Relationship Equity in Human Relationships

Emotion and emotion regulation play an important role in the formation of trust [33,151]. Expressions of emotion are a crucial mechanism by which people determine how to relate to each other and whether to trust each other. Social functional accounts of emotion highlight this important role of emotional expressions by conceptualizing them as "interpersonal communication systems that help individuals navigate the basic problems that arise in dyadic and group relations" [103]. Interactions thus involve a constant coordination on affect as participants of an interaction jointly determine how behavior should be interpreted and responded to [73]. Some emotional expressions increase interpersonal trust while others reduce it. For example, negative and especially hostile expressions, such as anger and contempt, have been found to impair trust formation [2,103], whereas positive expressions, e.g. expressions of embarrassment, have been found to facilitate trust formation [83].

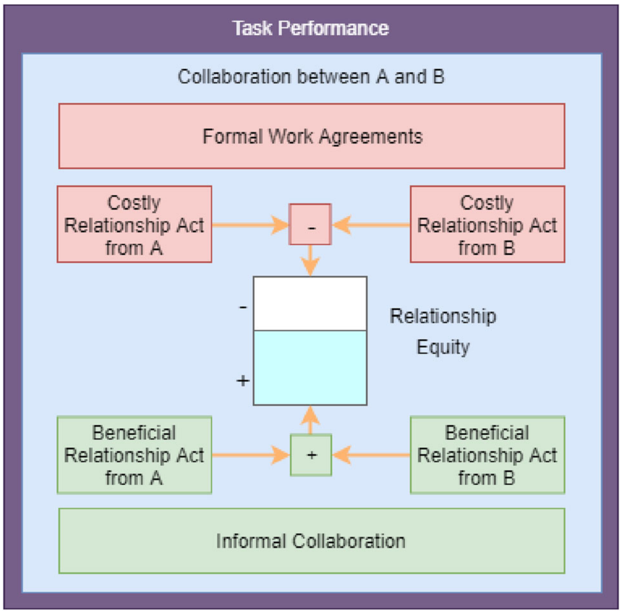
Task Performance
Collaboration between A and B
Formal Work Agreements
Costly Relationship Act from A
-
Costly Relationship Act from B
-
Relationship Equity
+
Beneficial Relationship Act from A
+
Beneficial Relationship Act from B
Informal Collaboration

Fig. 3 The core of the HRT trust model is the relationship equity bank, which accumulates the net result of repeated interactions over time

Researchers on emotion expression in couples suggest that in order to understand trust formation and maintenance, we need to take the temporal dynamics of emotion expression into account [66,68]. This important role of temporal dynamics for our understanding of trust formation has also recently been highlighted in the area of human–automation interaction, e.g. [157]. A point graph method can be used to account for different temporal dynamics of emotion expressions [48]. Point graphs plot the cumulative sum of positive minus negative expressions over time (see Fig. 4), describing emotional interaction dynamics over time. An upwards directed point graph (regulated) indicates an interaction in which participants are able to shape the emotional dynamics in such a way that more positive than negative expressions are consistently produced, whereas a downwards directed point graph (non-regulated) indicates an inability to do so. While interactions in couples exhibit different levels of intensity, research shows that as long as a surplus of positive over negative behavior

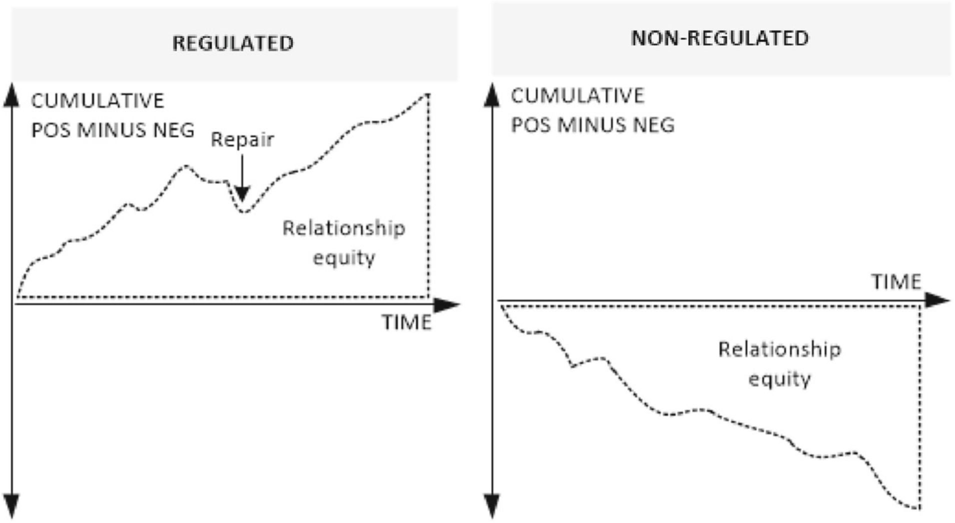

Fig. 4 Regulated and non-regulated point graphs

can be maintained (regulated interaction) it has positive consequences for short and long-term outcomes.

Evidence from several studies suggests that the ratio of positive to negative behaviors assessed through the point graphs generalizes as a predictor of outcomes from couples to teams. For example, one study showed that this ratio predicted satisfaction with group membership and team performance [75]. Further studies [43,72,89] all demonstrate the importance of the ratio of positive to negative expressions for team performance.

The point graphs also highlight that each positive or negative behavior exhibited by participants of an interaction has a cumulative impact on the relationship. An interaction that is characterized by mostly positive behaviors is likely more resilient towards occasional negative or hostile behaviors than an interaction that has been less positive over time. We use the term relationship equity to refer to the idea that trust is built up through moment-to-moment exchanges of positive and negative behaviors that accumulate over time. The term "equity" highlights that the impact one negative behavior has on the relationship depends on the equity accumulated throughout prior exchanges. Operationally, as suggested also by [157], relationship equity is best understood as the area under the curve of a point graph as shown in Fig. 4 [157].

When integrating robots into teams it is crucial to understand how their presence and behavior influences a team's exchange of positive and negative behaviors and, through that, overall trust formation. Currently, only little is understood about how robots influence the dynamics of the teams they are embedded in [74,76]. Recent work, however, has shown that robots can actively shape interaction dynamics in teams through repair of negative behaviors [75], through the expression of vulnerable behavior [135], or through the expression of group based emotion [19]. It is thus important to understand not only how robots influence the formation of a team's relationship equity through their behavior, but also how robots might be used to actively regulate interpersonal exchanges to promote relationship equity buildup and trust formation.

Proposition 1 *When designing a robotic team partner, it should have access to a relationship equity bank that allows the robot to maintain an understanding of the current relationship equity, rising with each positive relationship act, and falling with each negative one.*

3.2 Relationship Equity as a Predictor for Team processes that Manage Risk

3.2.1 Formal and Informal Work Agreements

Based on their relationship equity and current knowledge of one another's capabilities, the team members may jointly define a series of work agreements [102,107]. Work agreements ensure smooth collaboration between team members by explicitly defining collaborative agreements, such as communication, coordination, and task allocation between the team members. Work agreements evolve over time as they are either learned implicitly through training or collaboration, or formed explicitly through formalized rules in the form of obligations and permissions/prohibitions. Scientific research on the formalization of work agreements (in the literature often referred to as "social commitments") comes form the field of normative multi-agent systems, where work agreements are used to support robustness and flexibility [17]. According to this framework, a work agreement is an explicit agreement made between two parties, denoted as a four-place relation: $\langle debtor, creditor, antecedent, consequent \rangle$, where the debtor owes it to the creditor to effectuate the consequent once the antecedent is valid. Work agreements can, for instance, be used to specify the extent to which team members monitor and check one another's work and/or ask for permission to continue with their next activity or task before doing so. In sum, work agreements can be used to mitigate the risks involved in collaboration (assessed based on the actor's current knowledge of its team members' capabilities), by introducing rules that restrict the team members in their autonomy, especially when it comes to task allocation, assessment, and completion. For more information, please refer to related works, such as [22,82,102].

Proposition 2 *Relationship equity will negatively predict the degree to which formal work agreements will be constructed as a method for reducing risk. Lower degrees of relationship equity will predict more formal work agreements.*

3.2.2 Collaboration

The team proceeds to collaborate in compliance with the work agreements. During collaboration, the actors observe one another's capabilities and inspect one another's deliverables and performance. Based on their observations, the agents continuously update their mental models and corresponding trust stances. This potentially leads to revisions of the work agreements.

Collaboration does not merely entail the team's ability to adequately perform the task. Collaboration is also characterized by social interaction between team members, i.e. non-task related peripheral interactions with team members (jokes, humor, being able to honestly call each other out, informal things). Being able to show vulnerability at work, especially during non-work related interactions, has been shown to increase team effectiveness, as it facilitates positive trust development [37,39].

Proposition 3 *Relationship equity will positively predict the degree to which informal collaboration will occur as the primary manner of interaction between team members. In a team with a high relationship equity (through social interaction and successful past team performance) the need for regulative formal agreements and corrections will decrease, leading to a less controlling atmosphere in the team environment.*

3.2.3 Team Trust Dynamics

Unique trust effects can occur within a team as a result of feedback loops. For example, *ripple effects* can occur when one behavior of a team member is copied by another. In a striking example of this type of effect, a robot expressing vulnerability caused other human team members in the group to express vulnerability as well [135]. Another example are *spiraling effects* when negative behaviors and especially hostile behaviors have a tendency to be reciprocated and trigger a spiral of increasing negativity with detrimental consequences for trust [2]. Finally, *maladaptive feedback loops* can occur when team members are simply out of sync with one another. For example, a teammate A may try to compensate for a perceived failure that teammate B is not aware of. The lack of reciprocation by teammate B may cause frustration by teammate A that leads to confusion in teammate B. The lack of empathy by teammate B may inspire even more frustration in teammate A. What started as a minor miscommunication and ensuing adaptation strategy can cause maladaptive feedback that further escalates the situation.

Proposition 4 *In our model, ripple and negative spiral effects can occur as a result of subsequent reactive tit-for-tat behaviors. We predict that these will be positive with higher relationship equity and negative with lower relationship equity. Maladaptive feedback loops may occur when there is sustained miscalibration between the two actors.*

4 Minimizing Social Calibration Errors as a Way to Build Relationship Equity

This section presents a signal detection approach to social trust calibration, proposing that trust violations can either be correctly anticipated (hit), incorrectly anticipated (false alarm), incorrectly unanticipated (miss), or correctly unanticipated (correct rejection).

While many of the terms described in our team trust model have been researched extensively, the proposed concept of social trust calibration is new. The focus of the current section is to describe a signal detection approach for social trust calibration with the goal to create intelligent systems that can recognize when their social behavior may cause a trust violation, e.g. when a machine performs unlike its usual performance standards, or when it fails to meet others' expectations.

4.1 Social Trust Calibration

When teammates collaborate, they engage in a constant process of social trust calibration (see Fig. 1). Calibrated trust between team members is defined such that someone's perceived trustworthiness of a teammember matches that teammember's actual trustworthiness. When looking at the passive trust clibration part of our model, depicted in Fig. 5, perfectly calibrated trust would mean that the prediction error (miscalibration) is 0. Undertrust is defined as the situation in which the trustor has lower trust in the trustee than the trustee deserves. In situations of undertrust, the trustor fails to take full advantage of the trustee's capabilities. In teams, situations of undertrust can result in suboptimal solutions to problems, a lack of communication, and increased

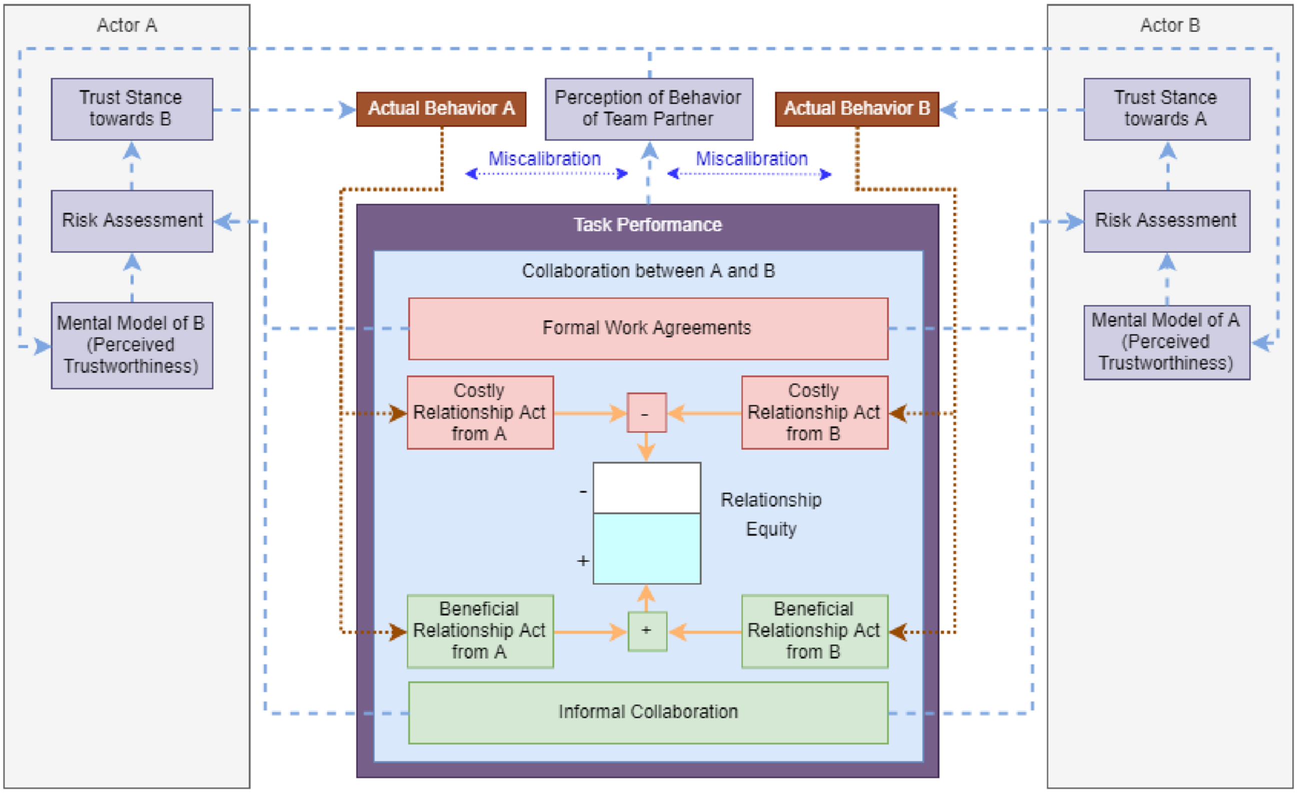
Actor A
Trust Stance towards B
Risk Assessment
Mental Model of B (Perceived Trustworthiness)
Actual Behavior A
Perception of Behavior of Team Partner
Actual Behavior B
Miscalibration
Miscalibration
Actor B
Trust Stance towards A
Risk Assessment
Mental Model of A (Perceived Trustworthiness)
Task Performance
Collaboration between A and B
Formal Work Agreements
Costly Relationship Act from A
-
Costly Relationship Act from B
-
+
Relationship Equity
Beneficial Relationship Act from A
+
Beneficial Relationship Act from B
Informal Collaboration

Fig. 5 Passive trust calibration between team partners

workload for individual members (as opposed to distributed workload). Undertrust can be detrimental to the effectiveness and efficiency of the HRT, as it may lead to disuse or micromanagement. Actors can also trust each other too much. Overtrust is defined as the situation in which the trustor trusts the trustee to a greater extent than deserved given the trustee's true capabilities. In situations of overtrust, the trustor allows the trustee to act autonomously, even in situations where the trustee is not capable of performing the task adequately. Overtrust is a dangerous condition that should be avoided for all critical systems because it can lead to disastrous outcomes. Many examples exist of accidents caused by overtrust [6,45,111,133]. Overtrust can be hazardous as it may lead to a lack of guidance and control for systems not fully capable of performing a given task.

Proposition 5 *Socially calibrated trust will increase relationship equity. Socially miscalibrated trust will decrease relationship equity.*

4.2 A Signal Detection Model for Social Trust Calibration

Trust (mis-)calibration may occur in any team situation with a great potential for frustration. Since trust violations have a detrimental effect on trust, it is vital to minimize their overall impact. Early detection and accurate awareness of potential trust violations may allow team partners to engage in active trust calibration (see Fig. 6), so as to prevent escalation of minor issues into larger problems within the HRT. Accordingly, we created a simple model of anticipated and unanticipated trust violations with the use of trust repair, dampening, and transparency methods. We may imagine four situations depending on whether a trust violation is anticipated and whether a trust violation occurred using a signal detection classification approach (see Fig. 7). A "Hit" situation may be one where a trust violation is anticipated prior to the occurrence of the trust violation and a trust violation also occurs. A system may actively attempt to lower expectations in anticipation of expected failures prior to the attempted action. Such an action would effectively dampen trust and may more appropriately calibrate trust. When the trust violation occurs, the robot can refer back to the lowered initial expectations. A "False Alarm" situation is one where a trust violation is anticipated, but does not occur. It may not be bad overall to dampen expectations that are not borne out in an actual trust violation. Lowering expectations may alert the operator to guide the interaction more. However, dampening without the trust violation occurring may result in the system downgrading itself unnecessarily which may lead to reduced trust over time. A "Miss" situation is one where no trust violation was anticipated, but one occurs. This is the situation

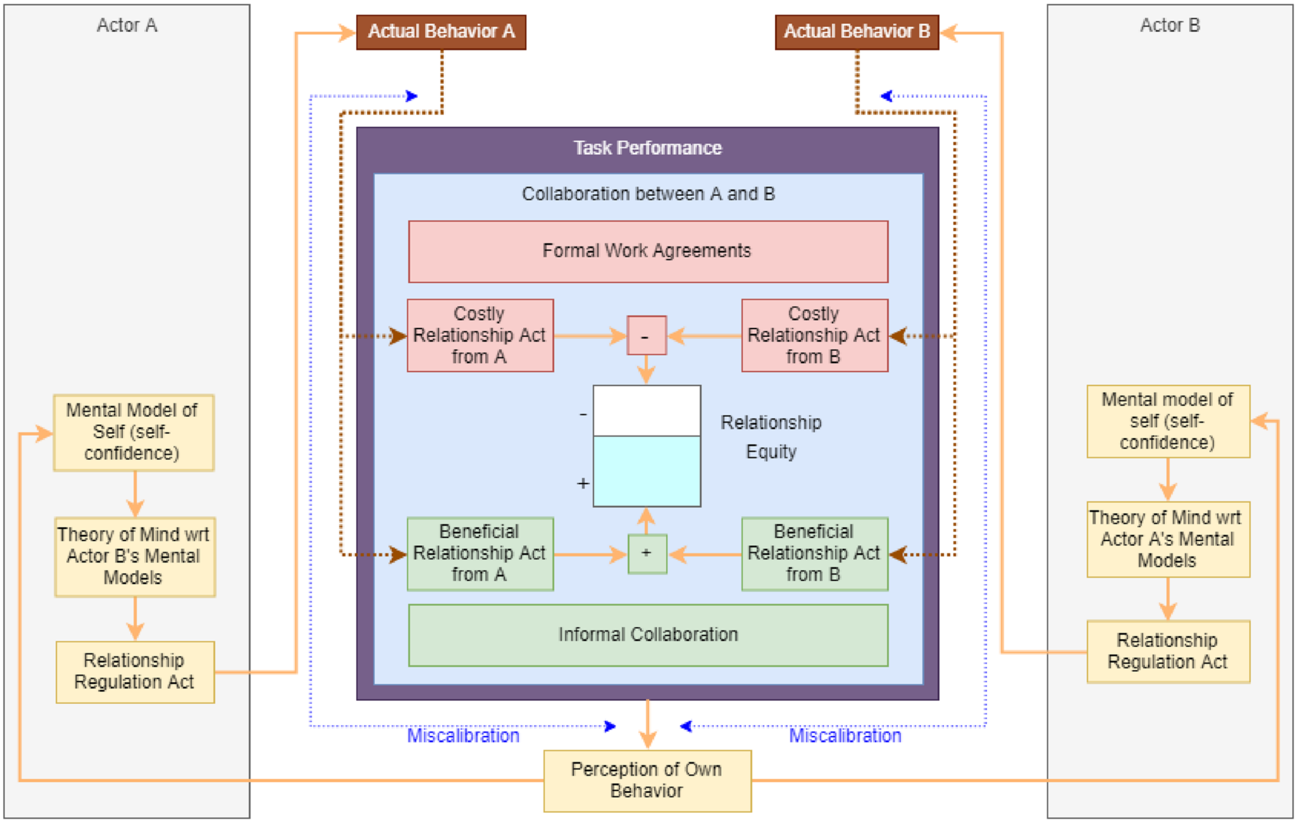
Actor A
Actual Behavior A
Actual Behavior B
Actor B
Task Performance
Collaboration between A and B
Formal Work Agreements
Costly Relationship Act from A
-
Costly Relationship Act from B
-
+
Relationship Equity
Beneficial Relationship Act from A
+
Beneficial Relationship Act from B
Informal Collaboration
Mental Model of Self (self-confidence)
Theory of Mind wrt Actor B's Mental Models
Relationship Regulation Act
Mental model of self (self-confidence)
Theory of Mind wrt Actor A's Mental Models
Relationship Regulation Act
Miscalibration
Miscalibration
Perception of Own Behavior

Fig. 6 Active trust calibration between team partners

where trust repair activities are most needed since a trust violation was not anticipated. This may also be the situation that occurs most frequently in team interactions. The "Correct Rejection" situation is one where no trust violation is anticipated and no trust violation occurs. No specific action is required. General transparency methods can enhance this baseline situation in a passive, but proactive manner.

The signal detection model focuses on single-shot interactions, yet when looking at longitudinal teaming, one should actually be looking at a series of single-shot interactions, that accumulate over time, as outlined in Sect. 3.1. Figure 4 displays how relationship equity may be constructed (or broken down) as a result of repeated interactions. Ultimately, the relationship equity bank, described in Sect. 3.1, serves as a lens through which each subsequent interaction is reviewed, and thereby mediates trust calibration and mental model updating, as described by the HRT trust model presented in Fig. 2.

Proposition 6 *Social hits and correct rejections will decrease the probability of mis-calibration. Social false alarms and misses will increase the probability of mis-calibration.*

5 Methods for Building Relationship Equity through Social Trust Calibration

In the previous sections, we have outlined components of a theory that describes social trust calibration in human–robot teams. The theory assumes a number of calibration methods, such as trust repair, trust dampening, transparency and explanation. We describe each method in detail in Table 2 as well as in the following subsections.

Recognizing that a potential trust violation is going to happen or that an actual trust violation has occurred is important to determine whether trust calibrating acts are appropriate. Please note that a trust violation can go both ways: either an actor is lucky and performs unusually well, or it fails and performs substantially worse than it normally would. In both cases, trust calibrating acts are warranted. This section introduces a range of actions and utterances an actor can perform to calibrate trust either by mitigating a state of undertrust through trust repair or a state of overtrust through trust dampening. In addition we discuss methods of transparency and explanation as viable options for trust calibration.

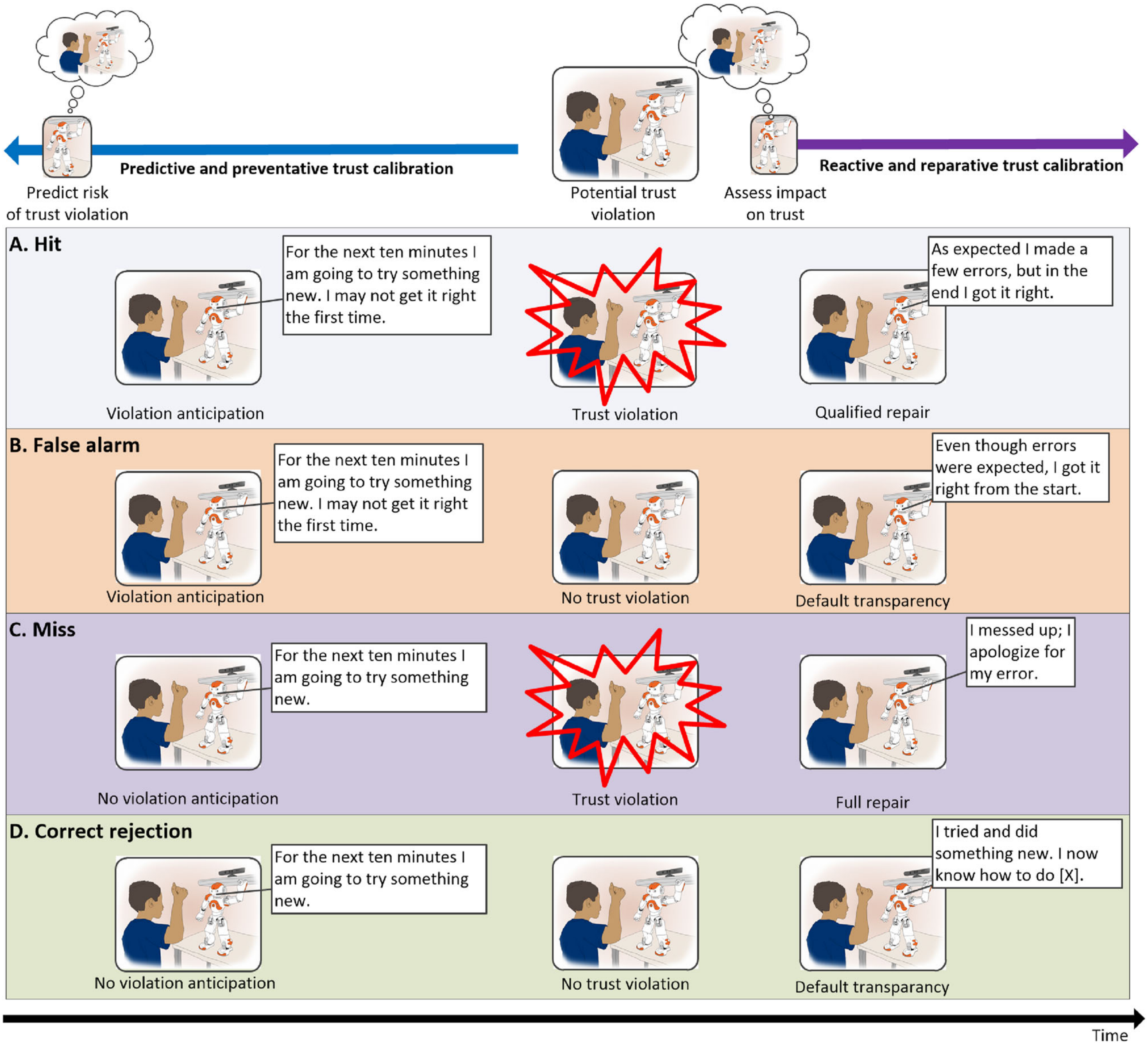
Predict risk of trust violation
Predictive and preventative trust calibration
Potential trust violation
Assess impact on trust
Reactive and reparative trust calibration
A. Hit
For the next ten minutes I am going to try something new. I may not get it right the first time.
Violation anticipation
Trust violation
As expected I made a few errors, but in the end I got it right.
Qualified repair
B. False alarm
For the next ten minutes I am going to try something new. I may not get it right the first time.
Violation anticipation
No trust violation
Even though errors were expected, I got it right from the start.
Default transparency
C. Miss
For the next ten minutes I am going to try something new.
No violation anticipation
Trust violation
I messed up; I apologize for my error.
Full repair
D. Correct rejection
For the next ten minutes I am going to try something new.
No violation anticipation
No trust violation
I tried and did something new. I now know how to do [X].
Default transparency
Time

Fig. 7 Over time, an actor can use its predictions about itself and its impact on the other actor's trust stance to dampen or repair trust, or to prevent breaking trust. Credit for illustration: USC Viterbi School of Engineering, with permission from Prof. M. Matarić [49]

5.1 Methods for Trust Repair

Trust repair is a reactive approach to restore undertrust after the machine has made a mistake, caused trouble, or displayed unexpected or inappropriate behavior [7,29,94]. Trust repair seeks to repair situations where trust is broken or where there is an initial distrust bias, e.g. by explaining the cause and/or situational nature of a mistake, or making promises about future behavior.

Proposition 7 *Trust repair activities will help to increase trust if they are appropriately timed and commensurate with the degree of trust violation.*

5.2 Methods for Dampening Trust

Trust dampening is a reactive approach to quell overtrust after a machine has made a lucky guess, or when a machine makes a mistake that has not been noted by its collaborators or users. Trust dampening approaches seek to lower expectations when too much faith has been put into a machine. Dampening methods may include showing a user what a system failure looks like, showing a history of performance, and providing likelihood alarms [88]. Dampening approaches may need to be applied especially in the beginning of interacting with a new machine. Often, people tend to have high expectations of machines and robots known as automation

Table 2 Methods for trust repair, trust dampening, transparency, and explanation

Methods for trust repair	Description
\Apology	Admission of fault and statement of remorse [85,105,131]
\Apology w/Improvement	Apology and promise to improve in the future
\Apology w/Entity Attribution	Apology and faulting another party
\Apology w/Process Attribution	Apology and faulting a process
\Apology before next potential error	Apology that only occurs when the actor next approaches a situation similar to the one where a previous error occurred
\Denial	Recognition of error and assertion that they are not at fault [85,131]
\Denial w/Entity attribution	Denial and faulting another party
\Denial w/Process Attribution	Denial and faulting a process
\Recognition	Recognition of the error's cause without acknowledging fault
\Explanation	Explanation of the error's cause without acknowledging fault
\Control	Statement of control/awareness of situation
\Proficiency	Statement of expertise or proficiency [65]
\Downgrading	Acknowledging error, but downgrading severity
\Blaming	Blaming of another entity, without denial or apology [71,93]
\Gas-lighting	Stating that no error occurred
Methods for trust dampening	**Description**
\Lowering expectations	Warning that one may not perform well under current circumstances
\Expressing reduced confidence	Express that one is not confident about executing the current task
\Providing caveats to behavior	Conveying that one is unable to perform certain tasks
\Convey system limitations	Provide information about the range of tasks and capabilities one is able to perform, within which domain, and under what circumstances, potentially explicitly pointing to one's own limitations
\Request for assistance	Asking for assistance when one fails to figure out what to do or to perform the task [69]
\Provide timely warnings	Provide timely warnings about one's own potential limitations, enabling teammates to adapt and take over part of the task load [105,119]
Methods for transparency	**Description**
\Convey a history of performance	Displaying errors directly and for specific situations [95]
\Provide performance feedback	Adjusting feedback based on the biases of the operator [38]
\Convey uncertainty directly	Conveying uncertainty directly by providing facial cues [8] or machine confidence indicators [36,57,62,86,97,157]
\Provide verification methods	Offering methods for detailed verification of the system's behavior and performance to reduce the number of commission errors [6]
\Enhance mode awareness	Providing guidance on different system modes, decision-making, and transparency of the algorithm [30]
\Show critical states	Enhance an operator's mental state by showing only the critical states, situations that require immediate action, of an algorithm's policy [52,60,153]
\Explain why things fail	Provide a way to know the nature of errors, why they occur, and what they mean in the context of the designer's intent and goals for the automated aid [38]

Table 2 continued

Methods for transparency	Description
\Mimic the social behavior of a user	Actively mimicking a human driver, for instance by matching the goals of an autonomous driving system with the goals of the driver [146] or making a virtual autonomous driver look and act like the driver [146,147]
\Equip machine with social behaviors	Designing automation that conforms to social norms, such as politeness or etiquette [46,56,64,100,109,112], or adding anthropomorphic features to an interface [1,26,54,110,118]
Methods for explanation	**Description**
\Intuitive confidence measure	Explaining the likelihood of a correct single prediction (output) based on similarity and previous experiences [140,141]
\Contrastive explanations	Explaining the current output (“fact”) in relation to an output of interest (“foil”) [65,142,143]

bias [38]. When actual robotic behavior is observed, people may punish machines more deeply than their human counter parts [47].

Proposition 8 *Trust dampening activities will help to stimulate trust resilience by appropriately adjusting expectations in the face of anticipated errors.*

5.3 Methods for Transparency

Some methods apply to facilitate both processes of repair and dampening, such as making the internal processes and processing steps more transparent or inspectable for other team members. These methods revolve around the central method of increasing the transparency of a system [13], such as its performance, its process, and its intent and purpose [3,13,14,99]. The benefits of increasing transparency have been demonstrated through empirical research. Previous work has emphasized design approaches to increase the transparency of the system to a user to promote trust calibration. Transparency design methods focus on conveying trust cues which convey information about an agent’s uncertainty, dependence, and vulnerability [25,136].

Proposition 9 *Transparency activities will help to calibrate trust by providing accurate meta-information about the robotic partner.*

5.4 Methods for Explanation

Recently, the research on explainable AI (XAI) has expanded rapidly (e.g. see [23,51,104]). XAI refers to (1) the ability to offer a meaningful explanation for a specific human actor when needed, and (2) the ability to ask for an interpretable explanation from a specific human actor when needed. So far, research centered primarily on the first type of ability, often with a focus on a specific (classification) task or machine learning model. However, more integrative methods are evolving, which include both bottom-up data-driven (perceptual) processes and top-down model-based (cognitive) processes ([106]; cf. dual process theories [79], [40]). Such methods could help to assess the trustworthiness of AI output (i.e. robot’s own task performance) and, subsequently, explain the foundation and reasons of this performance to establish an adequate trust stance.

At the perceptual level, the provision of an Intuitive Confidence Measure (ICM) enhances human judgment of the quality of the data processing and corresponding inferences [140,141]. The ICM explains the likelihood of a correct single prediction (output) based on similarity and previous experiences (e.g. “I am reasonably certain that there is a victim at location A”). At the cognitive level, available models of the user’s goals, beliefs, and emotions can be used to provide explanations that provide the reasons of specific output (e.g. advice) and behaviors (e.g. “It is important to drive around this area, because there is an explosion risk”; cf. [81]). Personalization of these explanations is crucial to accommodate a user’s goal and emotional state [80]. At the perceptual-cognitive level, contrastive explanations provide the reasons of a specific output (the “fact”) in relation to an alternative output that is of interest (the “foil”) [142,143]. Humans often use this type of explanation. Contrastive explanations narrow down the amount of features that are included in the explanation, making it more easy to interpret [101].

For the construction of meaningful explanations, challenges are to establish at run-time (1) an adequate level of detail, specificity and extensiveness, (2) an effective dynamic adaptation to the human and context, and (3) an appropriate choice for allocating the initiative of an explanation dialogue. The development of Ontology Design Patterns can help to meet these challenges, particularly for an artificial actor’s reasoning and communication [130]. Interaction Design Patterns are being constructed for shaping mixed-initiative communicative acts of explanation [106]. Developing the ability for a robot to ask for an interpretable explanation from a specific human actor when needed, is yet a rather unexplored area.

Proposition 10 *Explanation activities will help to calibrate trust by providing accurate meta-information about the robotic partner.*

6 Implications of the Framework

The HRT trust model describes the role and purpose of trust calibrating acts, i.e. trust repair and trust dampening, in improving HRT collaboration over longer periods of time. Our model assumes that trust calibration benefits all actors as long as they are sincere and benevolent in their collaboration. If accurately executed, trust calibration results in optimized collaboration through the adoption of implicit and explicit work agreements that appropriately benefit from the strengths of the actors involved, while mitigating risk by compensating for team members' shortcomings and/or limitations.

From our model, there are a number of different research directions that can be explored given the longitudinal interaction between humans and artificial team members (e.g. robots, agents, and other AI-based systems) that actively employ trust calibration methods. Some of these topics have been raised in other publications [27], each of which could merit its own research program with a set of experiments. We discuss several of these research directions as well as the implications of our framework next.

6.1 A Common Framework for Mixed Human–Robot Teams

The first implication of our model is that it presents a flexible and common way of modeling relationship equity as it relates to trust calibration in human–human, human–robot or robot–robot teams. This is useful because currently the social capabilities of robots are still limited, but expected to greatly improve in the next few decades. For example, if a robot can function as a human in this type of relationship, the robot would be expected to have models and an understanding of its own behavior, its teammate, and the process of collaboration itself. Progress indeed has been made in each of these areas, but work remains to build the types of teams that can regulate emotion and manage relationships in mixed human–robot teams. Currently, these relationships are asymmetrical where the human compensates for the lack of a robot's social abilities. For longer-term interaction to be sustained, those deficiencies will have to be resolved.

6.2 Adaptive Trust Calibration Systems for Mixed Human–Robot Teams

The second implication of our work is that it provides a step in the direction of mutually adaptive trust calibration systems (ATCS). These systems are a special form of adaptive automation [10,42,63,78,129] that can measure the trust state of a human and then adapt their behavior accordingly to provide a positive impact on team performance. When implemented well, these systems have the potential to calibrate trust on the fly and provide immediate benefits for human–robot team performance.

In recent pioneering work that exemplifies ATCS, researchers designed and developed a robot that calibrates trust through automatically generated explanations [149]. Especially when the robot had low capabilities, its explanations led to improved transparency, trust, and team performance. This important work shows how the impact of expected trust violations can be mediated with the use of a trust dampening strategy, i.e. explanations. Other pioneering work demonstrating the utility of ATCS has used various modeling techniques to incorporate trust measures and adapt team performance. For instance, Chen et al. (2018) [15] used trust-POMDP modeling to infer a human teammate's trust and only engage in moving a critical object when a human teammate has built up enough trust in a robot's arm's ability to move objects carefully.

With the use of our model, researchers have a means to place this work in a larger context of challenges related to trust development in HRTs. This framework may therefore serve as a guide by providing an overview of what parts the research community currently understands and/or is able to successfully implement in an artificial team mate, and which parts still require additional research.

6.3 Towards Long-Term Interaction for Human–Robot Teams

There is much to be learned about effective team behavior in general, both in human-only teams and teams with a mix of humans and artificial team members. By achieving a better understanding of trust dynamics, researchers may learn a great deal about the impact of particular team behaviors on the trust relationships within a team, e.g. what behaviors contribute to team effectiveness and what behaviors are detrimental.

More importantly, one might be able to recognize negative team behavior patterns, such as the ripple effect or downward spiraling mentioned in Sect. 3.2.3. In this way, artificial actors can be designed with a range of individual behaviors serving as subtle interventions that help steer away from such destructive patterns and simultaneously promote healthy, emotionally regulated, and effective team behavior patterns [144].

In addition, the issues below are primary research challenges that researchers could address in the future.

6.3.1 Ontology and Model Development

There is a great need to develop a method that allows for reasoning about humans, agents, and robots alike. For this purpose, we are developing an *ontology*, that provides the vocabulary and semantics of multi-modal HRT communication and the foundation for automated reasoning [5,77,114–116]. Furthermore, a computational model of trust development and repair -providing quantitative predictions of potential trust violations and their prospective impact- would help to focus and integrate research within the HRI community. Progress towards this goal has already been made by research that has indicated how trust can modeled to extend across tasks [134].

6.3.2 Trust Measurement and Modeling Development

Trust measurement remains a key issue for HRT interaction [34,35,44,53]. It is important to know how trust is initiated, how it develops, how it breaks down, and how it recovers. This requires a convergence of behavioral, self-report, observational, and neuro-physiological measures and correlates [28,53], as well as the development and validation of new measurements specific to the process of HRT trust. Recent work has demonstrated the multi-faceted nature of trust as it relates to delegation decisions [155] as well as the ability to model trust with dynamic Baysian models [156].

6.3.3 Implementation and Experimentation

In addition to the conceptual and computational models, it is important to develop software modules for robots and other artificial team members enabling them to compute and/or reason about trust and engage in trust calibrating acts where needed or appropriate [138,139,145]. This will also allow for hypothesis testing based on predictions and prescriptions provided by our model.

6.3.4 Validation and Verification

Endeavors of implementation and experimentation will facilitate theory development and refinement as we gather empirical data on the actual effectiveness of trust calibrating acts, as well as relationship equity models and predictions of potential trust violations, all situated in field exercises with prospective end users [91].

7 Conclusion

Future societies will rely substantially on human–robot teams (HRTs) in a wide variety of constellations. Enabling such teams to effectively work together towards the achievement of shared goals will be paramount to their success. The theory, models, methods, and research agenda presented in this paper will contribute to this endeavor, and may lead to the design and implementation of artificial team members and corresponding team behaviors that support healthy trust development, in turn contributing to high performing HRTs.

Acknowledgements This material is based upon work supported by the Air Force Office of Scientific Research under award numbers 16RT0881 and FA9550-18-1-0455, as well as the Dutch Ministry of Defence's exploratory research program (project Human-AI Teaming).

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Abubshait A, Wiese E (2017) You look human, but act like a machine: agent appearance and behavior modulate different aspects of human–robot interaction. Front Psychol 8:1393
2. Andersson LM, Pearson CM (1999) Tit for tat? the spiraling effect of incivility in the workplace. Acad Manag Rev 24(3):452–471
3. Alonso V, de la Puente P (2018) System transparency in shared autonomy: a mini review. Front Neurorobot 12:83
4. Atkinson DJ, Clancey WJ, Clark MH (2014) Shared awareness, autonomy and trust in human-robot teamwork. In: 2014 AAAI fall symposium series
5. Bagosi T, Hindriks KV, Neerincx MA (2016) Ontological reasoning for human-robot teaming in search and rescue missions. In: 2016 11th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 595–596
6. Bahner JE, Hüper AD, Manzey D (2008) Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. Int J Hum–Comput Stud 66(9):688–699
7. Baker AL, Phillips EK, Ullman D, Keebler JR (2018) Toward an understanding of trust repair in human–robot interaction: current research and future directions. ACM Trans Interact Intell Syst (TiiS) 8(4):30
8. Beller J, Heesen M, Vollrath M (2013) Improving the driver-automation interaction: an approach using automation uncertainty. Hum Factors 55(6):1130–1141
9. Billings DR, Schaefer KE, Chen JYC, Hancock PA (2012) Human-robot interaction: developing trust in robots. In: 2012 7th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 109–110
10. Byrne EA, Parasuraman R (1996) Psychophysiology and adaptive automation. Biol Psychol 42(3):249–268
11. Castelfranchi C, Falcone R (2010) Trust theory: a socio-cognitive and computational model, vol 18. Wiley, Hoboken
12. Chen JY, Barnes MJ (2014) Human-agent teaming for multirobot control: a review of human factors issues. IEEE Trans Hum–Mach Syst 44(1):13–29

13. Chen JYC, Barnes MJ, Wright JL, Stowers K, Lakhmani SG (2017) Situation awareness-based agent transparency for human-autonomy teaming effectiveness. In: Micro-and nanotechnology sensors, systems, and applications IX, vol 10194. International Society for Optics and Photonics, p 101941V
14. Chen JY, Lakhmani SG, Stowers K, Selkowitz AR, Wright JL, Barnes M (2018) Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theor Issues Ergon Sci 19(3):259–282
15. Chen M, Nikolaidis S, Soh H, Hsu D, Srinivasa S (2018) Planning with trust for human-robot collaboration. In: Proceedings of the 2018 ACM/IEEE international conference on human–robot interaction. ACM, pp 307–315
16. Chien SY, Lewis M, Sycara K, Kumru A, Liu JS (2019) Influence of culture, transparency, trust, and degree of automation on automation use. IEEE Trans Hum Mach Syst (submitted)
17. Chopra AK, Singh MP (2016) From social machines to social protocols: Software engineering foundations for sociotechnical systems. In: Proceedings of the 25th international conference on world wide web. International World Wide Web Conferences Steering Committee, pp 903–914
18. Cohen MS, Parasuraman R, Freeman JT (1998) Trust in decision aids: a model and its training implications. In: Proceedings of the command and control research and technology symposium. Citeseer
19. Correia F, Mascarenhas S, Prada R, Melo FS, Paiva A (2018) Group-based emotions in teams of humans and robots. In: Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction. ACM, pp 261–269
20. Correia F, Mascarenhas SF, Gomes S, Arriaga P, Leite I, Prada R, Melo FS, Paiva A (2019) Exploring prosociality in human-robot teams. In: 2019 14th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 143–151
21. Correia F, Melo FS, Paiva A (2019) Group intelligence on social robots. In: 2019 14th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 703–705
22. Dastani M, van der Torre L, Yorke-Smith N (2017) Commitments and interaction norms in organisations. Auton Agents Multi-Agent Syst 31(2):207–249
23. de Graaf MM, Malle BF (2019) People's explanations of robot behavior subtly reveal mental state Inferences. In: 2019 14th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 239–248
24. de Visser EJ, Parasuraman R (2011) Adaptive aiding of human–robot teaming: effects of imperfect automation on performance, trust, and workload. J Cognit Eng Decis Mak 5(2):209–231
25. de Visser EJ, Cohen M, Freedy A, Parasuraman R (2014) A design methodology for trust cue calibration in cognitive agents. In: International conference on virtual, augmented and mixed reality. Springer, pp 251–262
26. de Visser EJ, Monfort SS, McKendrick R, Smith MAB, McKnight PE, Krueger F, Parasuraman R (2016) Almost human: anthropomorphism increases trust resilience in cognitive agents. J Exp Psychol Appl 22(3):331
27. de Visser EJ, Pak R, Neerincx MA (2017) Trust development and repair in human-robot teams. In: Proceedings of the companion of the 2017 ACM/IEEE international conference on human–robot interaction. ACM, pp 103–104
28. de Visser EJ, Beatty PJ, Estepp JR, Kohn S, Abubshait A, Fedota JR, McDonald CG (2018) Learning from the slips of others: neural correlates of trust in automated agents. Front Hum Neurosci. https://doi.org/10.3389/fnhum.2018.00309
29. de Visser EJ, Pak R, Shaw TH (2018) From 'automation'to 'autonomy': the importance of trust repair in human-machine interaction. Ergonomics 61(10):1409–1427
30. Degani A, Shafto M, Kirlik A (1999) Modes in human-machine systems: constructs, representation, and classification. Int J Aviat Psychol 9(2):125–138
31. Demir M, McNeese NJ, Cooke NJ (2017) Team situation awareness within the context of human-autonomy teaming. Cognit Syst Res 46:3–12
32. Demir M, McNeese NJ, Johnson C, Gorman JC, Grimm D, Cooke NJ (2019) Effective team interaction for adaptive training and situation awareness in human-autonomy teaming. In: 2019 IEEE conference on cognitive and computational aspects of situation management (CogSIMA). IEEE, pp 122–126
33. Deutsch M (1960) The effect of motivational orientation upon trust and suspicion. Hum Relat 13(2):123–139
34. Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013, March). Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction* (pp. 251–258). IEEE Press
35. Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., ... & Yanco, H. (2012, March). Effects of changing reliability on trust of robot systems. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (pp. 73–80). ACM
36. Du N, Huang KY, Yang XJ (2019) Not all information is equal: effects of disclosing different types of likelihood information on trust, compliance and reliance, and task performance in human-automation teaming. Hum Factors. https://doi.org/10.1177/0018720819862916
37. Duhigg C (2016) What google learned from its quest to build the perfect team. The New York Times Magazine. https://www.nytimes.com/2016/02/28/magazine/what-google-learned-from-its-quest-to-build-the-perfect-team.html
38. Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP (2003) The role of trust in automation reliance. Int J Hum–Comput Stud 58(6):697–718
39. Edmondson AC, Kramer RM, Cook KS (2004) Psychological safety, trust, and learning in organizations: a group-level lens. Trust Distrust in Organ Dilemmas Approaches 12:239–272
40. Evans JSB, Frankish KE (2009) In two minds: dual processes and beyond. Oxford University Press, Oxford
41. Falcone R, Castelfranchi C (2001) The socio-cognitive dynamics of trust: Does trust create trust? In: Trust in cyber-societies. Springer, pp 55–72
42. Feigh KM, Dorneich MC, Hayes CC (2012) Toward a characterization of adaptive systems: a framework for researchers and system designers. Hum Factors 54(6):1008–1024
43. Fredrickson BL, Losada MF (2005) Positive affect and the complex dynamics of human flourishing. Am Ppsychol 60(7):678
44. Freedy A, de Visser E, Weltman G, Coeyman N (2007) Measurement of trust in human-robot collaboration. In: International symposium on collaborative technologies and systems, 2007 (CTS 2007). IEEE, pp 106–114
45. Goddard K, Roudsari A, Wyatt JC (2011) Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 19(1):121–127
46. Goodrich MA, Crandall JW, Oudah M, Mathema N (2018) Using narrative to enable longitudinal human-robot interactions. In: Proceedings of the HRI2018 workshop on longitudinal human–robot teaming, Chicago, IL
47. Goodyear K, Parasuraman R, Chernyak S, de Visser EJ, Madhavan P, Deshpande G, Krueger F (2017) An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents. Soc Neurosci 12(5):570–581
48. Gottman JM, Levenson RW (1992) Marital processes predictive of later dissolution: behavior, physiology, and health. J Personal Soc Psychol 63(2):221

49. Greczek J, Kaszubski E, Atrash A, Matarić M (2014) Graded cueing feedback in robot-mediated imitation practice for children with autism spectrum disorders. In: The 23rd IEEE international symposium on robot and human interactive communication (2014 RO–MAN). IEEE, pp 561–566
50. Groom V, Nass C (2007) Can robots be teammates? Benchmarks in human–robot teams. Interact Stud 8(3):483–500
51. Guidotti R, Monreale A, Turini F, Pedreschi D, Giannotti F (2018) A survey of methods for explaining black box models. arXiv preprint arXiv:1802.01933
52. Guznov S, Lyons J, Pfahler M, Heironimus A, Woolley M, Friedman J, Neimeier A (2019) Robot transparency and team orientation effects on human–robot teaming. Int J Hum Comput Interact. https://doi.org/10.1080/10447318.2019.1676519
53. Hancock PA, Billings DR, Schaefer KE, Chen JYC, De Visser EJ, Parasuraman R (2011) A meta-analysis of factors affecting trust in human–robot interaction. Hum Factors 53(5):517–527
54. Haring KS, Watanabe K, Velonaki M, Tossell CC, Finomore V (2018) FFAB-The form function attribution bias in human–robot interaction. IEEE Trans Cognit Dev Syst 10(4):843–851
55. Haring KS, Mosley A, Pruznick S, Fleming J, Satterfield K, de Visser EJ, Tossell CC, Funke G, (2019) Robot authority in human-machine teams: effects of human-like appearance on compliance. In: Chen J, Fragomeni G (eds) Virtual, augmented and mixed reality. Applications and case studies. HCII, (2019) Lecture notes in computer science, vol 11575. Springer, Cham, pp 63–78
56. Hayes CC, Miller CA (2010) Human-computer etiquette: Cultural expectations and the design implications they place on computers and technology. CRC Press, Boca Raton
57. Helldin T, Falkman G, Riveiro M, Davidsson S (2013) Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In: International conference on automotive user interfaces and interactive vehicular applications. ACM, pp 210–217
58. Hertz N, Shaw T, de Visser EJ, Wiese E (2019) Mixing it up: how mixed groups of humans and machines modulate conformity. J Cogn Eng Decis Mak. https://doi.org/10.1177/1555343419869465
59. Hoff KA, Bashir M (2015) Trust in automation: Integrating empirical evidence on factors that influence trust. Hum Factors 57(3):407–434
60. Huang SH, Bhatia K, Abbeel P, Dragan AD (2018) Establishing appropriate trust via critical states. In: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 3929–3936
61. Huseman RC, Hatfield JD, Miles EW (1987) A new perspective on equity theory: the equity sensitivity construct. Acad Mmanag Rev 12(2):222–234
62. Hutchins AR, Cummings ML, Draper M, Hughes T (2015) Representing autonomous systems' self-confidence through competency boundaries. In: The Human factors and ergonomics society annual meeting, vol 59. SAGE Publications Sage CA, Los Angeles, CA, pp 279–283
63. Inagaki T et al (2003) Adaptive automation: sharing and trading of control. Handb Cognit Task Des 8:147–169
64. Inbar O, Meyer J (2019) Politeness counts: perceptions of peacekeeping robots. IEEE Trans Hum Mach Syst 49(3):232–240
65. Israelsen BW, Ahmed NR (2019) "Dave... I can assure you... that it's going to be all right..." A definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. ACM Comput Surv 51(6):113
66. Gottman JM (1994) What predicts divorce?. L. Erlbaum, USA
67. Gottman JM (2005) The mathematics of marriage: dynamic nonlinear models. MIT Press, Boston
68. Gottman JM (2011) The science of trust: emotional attunement for couples. WW Norton & Company, New York
69. Jackson RB, Williams T (2019) Language-capable robots may inadvertently weaken human moral norms. In: 2019 14th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 401–410
70. Jarrold W, Yeh PZ (2016) The social-emotional turing challenge. AI magazine 37(1):31–39
71. Jensen T, Albayram Y, Khan MMH, Fahim MAA, Buck R, Coman E (2019) The apple does fall far from the tree: user separation of a system from its developers in human-automation trust repair. In: Proceedings of the 2019 on designing interactive systems conference. ACM, pp 1071–1082
72. Jung MF (2016) Coupling interactions and performance: predicting team performance from thin slices of conflict. ACM Trans Comput–Hum Interact (TOCHI) 23(3):18
73. Jung MF (2017) Affective grounding in human-robot interaction. In: Proceedings of the 2017 ACM/IEEE international conference on human–robot interaction. ACM, pp 263–273
74. Jung MF, Beane M, Forlizzi J, Murphy R, Vertesi J (2017) Robots in group context: rethinking design, development and deployment. In: Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems. ACM, pp 1283–1288
75. Jung MF, Martelaro N, Hinds PJ (2015) Using robots to moderate team conflict: the case of repairing violations. In: Proceedings of the tenth annual ACM/IEEE international conference on human–robot interaction. ACM, pp 229–236
76. Jung MF, Šabanović S, Eyssel F, Fraune M (2017) Robots in groups and teams. In: Companion of the 2017 ACM conference on computer supported cooperative work and social computing. ACM, pp 401–407
77. Juvina I, Collins MG, Larue O, Kennedy WG, Visser ED, Melo CD (2019) Toward a unified theory of learned trust in interpersonal and human–machine interactions. ACM Trans Interact Intell Syst 9(4):24
78. Kaber DB, Endsley MR (2004) The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. Theor Issues Ergon Sci 5(2):113–153
79. Kahneman D (2011) Thinking, fast and slow. Macmillan, London
80. Kaptein F, Broekens J, Hindriks K, Neerincx M (2017) Personalised self-explanation by robots: the role of goals versus beliefs in robot-action explanation for children and adults. In: 2017 26th IEEE international symposium on robot and human interactive communication (RO–MAN). IEEE, pp 676–682
81. Kaptein F, Broekens J, Hindriks K, Neerincx M (2017) The role of emotion in self-explanations by cognitive agents. In: 2017 seventh international conference on affective computing and intelligent interaction workshops and demos (ACIIW). IEEE, pp 88–93
82. Kayal A (2017) Normative social applications: user-centered models for sharing location in the family life domain. Ph.D. thesis, Delft University of Technology
83. Keltner D, Young RC, Buswell BN (1997) Appeasement in human emotion, social practice, and personality. Aggress Behav 23(5):359–374
84. Kiesler S (2005) Fostering common ground in human–robot interaction. In: ROMAN 2005. IEEE international workshop on robot and human interactive communication. IEEE, pp 729–734
85. Kohn SC, Quinn D, Pak R, de Visser EJ, Shaw TH (2018) Trust repair strategies with self-driving vehicles: an exploratory study. In: Proceedings of the human factors and ergonomics society annual meeting, vol 62. Sage, Los Angeles, pp 1108–1112
86. Kunze A, Summerskill SJ, Marshall R, Filtness AJ (2019) Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. Ergonomics 62(3):345–360
87. Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. Hum Factors 46(1):50–80

88. Lewicki RJ, Tomlinson EC, Gillespie N (2006) Models of interpersonal trust development: theoretical approaches, empirical evidence, and future directions. J Mnag 32(6):991–1022
89. Losada M, Heaphy E (2004) The role of positivity and connectivity in the performance of business teams: a nonlinear dynamics model. Am Behav Sci 47(6):740–765
90. Lyons JB, Guznov SY (2019) Individual differences in human-machine trust: a multi-study look at the perfect automation schema. Theor Issues Ergon Sci 20(4):440–458
91. Lyons JB, Clark MA, Wagner AR, Schuelke MJ (2017) Certifiable trust in autonomous systems: making the intractable tangible. AI Mag 38(3):37–49
92. Malle BF, Scheutz M (2018) Learning how to behave: moral competence for social robots. In: Handbuch Maschinenethik, pp 1–24
93. Malle BF, Scheutz M, Forlizzi J, Voiklis J (2016) Which robot am i thinking about? The impact of action and appearance on people's evaluations of a moral robot. In: The eleventh ACM/IEEE international conference on human robot interaction. IEEE Press, pp 125–132
94. Marinaccio K, Kohn S, Parasuraman R, De Visser EJ (2015) A framework for rebuilding trust in social automation across health-care domains. In: Proceedings of the international symposium on human factors and ergonomics in health care, vol 4. Sage, New Delhi, pp 201–205
95. Masalonis AJ, Parasuraman R (2003) Effects of situation-specific reliability on trust and usage of automated air traffic control decision aids. In: The human factors and ergonomics society annual meeting, vol 47. SAGE Publications Sage, Los Angeles, CA, pp 533–537
96. Mayer RC, Davis JH, Schoorman FD (1995) An integrative model of organizational trust. Acad Manag Rev 20(3):709–734
97. McGuirl JM, Sarter NB (2006) Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. Hum Factors 48(4):656–665
98. McNeese N, Demir M, Chiou E, Cooke N, Yanikian G (2019) Understanding the role of trust in human-autonomy teaming. In: Proceedings of the 52nd Hawaii international conference on system sciences
99. Mercado JE, Rupp MA, Chen JY, Barnes MJ, Barber D, Procci K (2016) Intelligent agent transparency in human-agent teaming for Multi-UxV management. Hum Factors 58(3):401–415
100. Meyer J, Miller C, Hancock P, de Visser EJ, Dorneich M (2016). Politeness in machine–human and human–human interaction. In: Proceedings of the human factors and ergonomics society annual meeting, vol 60. Sage, Los Angeles, pp 279–283
101. Miller T (2017) Explanation in artificial intelligence: Insights from the social sciences. arXiv preprint arXiv:1706.07269
102. Mioch T, Peeters MMM, Neerincx MA (2018) Improving adaptive human-robot cooperation through work agreements. In: 27th IEEE international symposium on robot and human interactive communication (RO–MAN 2018), Nanjing, China, August 27–31, 2018, pp 1105–1110. https://doi.org/10.1109/ROMAN.2018.8525776
103. Morris MW, Keltner D (2000) How emotions work: the social functions of emotional expression in negotiations. Res Organ Behav 22:1–50
104. Mueller ST, Hoffman RR, Clancey W, Emrey A, Klein G (2019) Explanation in human–AI systems: a literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv preprint arXiv:1902.01876
105. Nayyar M, Wagner AR (2018) When should a robot apologize? Understanding how timing affects human–robot trust repair. International conference on social robotics. Springer, Cham, pp 265–274
106. Neerincx M, Van der Waa J, Kaptein F, Van Diggelen J (2018) Using perceptual and cognitive explanations for enhanced human-agent team performance. In: Engineering psychology and cognitive ergonomics. Springer
107. Neerincx MA, van Diggelen J, van Breda L (2016) Interaction design patterns for adaptive human–agent–robot teamwork in high-risk domains. In: International conference on engineering psychology and cognitive ergonomics, pp 211–220. Springer
108. Ososky S, Schuster D, Phillips E, Jentsch FG (2013) Building appropriate trust in human-robot teams. In: AAAI spring symposium: trust and autonomous systems
109. Oudah M, Rahwan T, Crandall T, Crandall JW (2018) How AI wins friends and influences people in repeated games with cheap talk. In: Thirty-second AAAI conference on artificial intelligence
110. Pak R, Fink N, Price M, Bass B, Sturre L (2012) Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. Ergonomics 55(9):1059–1072
111. Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: an attentional integration. Hum Factors 52(3):381–410
112. Parasuraman R, Miller CA (2004) Trust and etiquette in high-criticality automated systems. Commun ACM 47(4):51–55
113. Parasuraman R, Riley V (1997) Humans and automation: use, misuse, disuse, abuse. Hum Factors 39(2):230–253
114. Peeters MMM (2016) ReMindMe: agent-based support for self-disclosure of personal memories in people with alzheimer's disease. In: Proceedings of the ICT4AWE. ScitePress, Rome, pp 61–66
115. Peeters MMM, Neerincx MA (2016) Human-agent experience sharing: creating social agents for elderly people with dementia. In: UMAP (extended proceedings)
116. Peeters MMM, van den Bosch K, Neerincx MA, Meyer JJC (2014) An ontology for automated scenario-based training. Int J Technol Enhanc Learn 6(3):195–211
117. Phillips E, Ososky S, Grove J, Jentsch F (2011) From tools to teammates: toward the development of appropriate mental models for intelligent robots. In: Proceedings of the human factors and ergonomics society annual meeting, vol 55. Sage, Los Angeles, pp 1491–1495
118. Phillips E, Zhao X, Ullman D, Malle BF (2018) What is human-like?: Decomposing robots' human-like appearance using the anthropomorphic robot (abot) database. In: Proceedings of the 2018 ACM/IEEE international conference on human–robot interaction. ACM, pp 105–113
119. Robinette P, Howard AM, Wagner AR (2015) Timing is key for robot trust repair. International conference on social robotics. Springer, Cham, pp 574–583
120. Robinette P, Howard AM, Wagner AR (2017) Effect of robot performance on human–robot trust in time-critical situations. IEEE Trans Hum Mach Syst 47(4):425–436
121. Robinette P, Li W, Allen R, Howard AM, Wagner AR (2016) Overtrust of robots in emergency evacuation scenarios. In: The eleventh ACM/IEEE international conference on human robot interaction. IEEE Press, pp 101–108
122. Rossi A, Dautenhahn K, Koay KL, Saunders J (2017) Investigating human perceptions of trust in robots for safe HRI in home environments. In: Proceedings of the companion of the 2017 ACM/IEEE international conference on human–robot interaction. ACM, pp 375–376
123. Salas E, Dickinson TL, Converse SA, Tannenbaum SI (1992) Toward an understanding of team performance and training. In: Teams: their training and performance. Ablex Publishing
124. Salas E, Sims DE, Burke CS (2005) Is there a "big five" in teamwork? Small group research 36(5):555–599
125. Salas E, Bisbey TM, Traylor AM, Rosen MA (2019) Can teamwork promote safety in organizations? Annu Rev Organ Psy-

chol Organ Behav. https://doi.org/10.1146/annurev-orgpsych-012119-045411

126. Salem M, Lakatos G, Amirabdollahian F, Dautenhahn K (2015) Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In: Proceedings of the tenth annual ACM/IEEE international conference on human–robot interaction. ACM, pp 141–148
127. Salomons N, van der Linden M, Strohkorb Sebo S, Scassellati B (2018) Humans conform to robots: disambiguating trust, truth, and conformity. In: Proceedings of the 2018 ACM/IEEE international conference on human–robot interaction. ACM, pp 187–195
128. Sauer J, Schmutz S, Sonderegger A, Messerli N (2019) Social stress and performance in human-machine interaction: a neglected research field. Ergonomics 62(11):1377–1391
129. Scerbo MW (1996) Theoretical perspectives on adaptive automation. Theory and applications, automation and human performance, pp 37–63
130. Schulte A, Donath D, Lange DS (2016) Design patterns for human-cognitive agent teaming. In: International conference on engineering psychology and cognitive ergonomics. Springer, pp 231–243
131. Sebo SS, Krishnamurthi P, Scassellati B (2019) "I don't believe you": investigating the effects of robot trust violation and repair. In: 2019 14th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 57–65
132. Shively RJ, Lachter J, Brandt SL, Matessa M, Battiste V, Johnson WW (2017) Why human-autonomy teaming? In: International conference on applied human factors and ergonomics. Springer, pp 3–11
133. Singhvi A, Russel K (2016) Inside the self-driving tesla fatal accident. The New York Times Magazine. https://www.nytimes.com/interactive/2016/07/01/business/inside-tesla-accident.html
134. Soh H, Shu P, Chen M, Hsu D (2018) The transfer of human trust in robot capabilities across tasks. arXiv preprint arXiv:1807.01866
135. Strohkorb Sebo S, Traeger M, Jung M, Scassellati B (2018) The ripple effects of vulnerability: the effects of a robot's vulnerable behavior on trust in human–robot teams. In: Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction. ACM, pp 178–186
136. Tenhundfeld NL, de Visser EJ, Haring KS, Ries AJ, Finomore VS, Tossell CC (2019) Calibrating trust in automation through familiarity with the autoparking feature of a Tesla Model X. J Cognit Eng Decis Mak. https://doi.org/10.1177/0018720819865412
137. Tenhundfeld NL, de Visser EJ, Ries AJ, Finomore VS, Tossell CC (2019) Trust and distrust of automated parking in a Tesla Model X. Hum Factors. https://doi.org/10.1177/0018720819865412
138. van der Vecht B, van Diggelen J, Peeters MMM, van Staal W, van der Waa J (2018) The SAIL framework for implementing human-machine teaming concepts. International conference on practical applications of agents and multi-agent systems. Springer, Cham, pp 361–365
139. van der Vecht B, van Diggelen J, Peeters MMM, Barnhoorn J, van der Waa J (2018) SAIL: a social artificial intelligence layer for human–machine teaming. International conference on practical applications of agents and multi-agent systems. Springer, Cham, pp 262–274
140. Van der Waa J, van Diggelen J, Neerincx M (2018) The design and validation of an intuitive certainty measure. IUI 2018 workshop on explainable smart systems. In: IUI 2018 workshop on explainable smart systems. ACM
141. Van der Waa J, van Diggelen J, Neerincx M, Raaijmakers S (2018) ICM: an intuitive, model independent and accurate certainty measure for machine learning. In: 10th internantional conference on agents and AI. ICAART
142. Van der Waa J, Robeer M, van Diggelen J, Brinkhuis M, Neerincx M (2018) Contrastive explanations with local foil trees. In: IJCAI 2018 workshop on human interpretability in machine learning. WHI
143. van der Waa J, van Diggelen J, van den Bosch K, Neerincx M (2018) Contrastive explanations for reinforcement learning in terms of expected consequences. Retrieved from arXiv:1807.08706
144. Van Diggelen J, Neerincx M, Peeters M, Schraagen JM (2018) Developing effective and resilient human-agent teamwork using team design patterns. IEEE Intell Syst 34(2):15–24
145. van Diggelen J, Barnhoorn JS, Peeters MMM, van Staal W, van Stolk M, van der Vecht B, van der Waa J, Schraagen JM (2019) Pluggable social artificial intelligence for enabling human-agent teaming. arXiv preprint arXiv:1909.04492
146. Verberne FMF, Ham J, Midden CJH (2012) Trust in smart systems: sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. Hum Factors 54(5):799–810
147. Verberne FMF, Ham J, Ponnada A, Midden CJH (2013) Trusting digital chameleons: the effect of mimicry by a virtual social agent on user trust. In: International conference on persuasive technology. Springer, pp 234–245
148. Walliser JC, de Visser EJ, Wiese E, Shaw TH (2019) Team structure and team building improve human–machine teaming with autonomous agents. J Cognit Eng Decis Mak. https://doi.org/10.1177/1555343419867563
149. Wang N, Pynadath DV, Hill SG (2016) Trust calibration within a human-robot team: comparing automatically generated explanations. In: The eleventh ACM/IEEE international conference on human robot interaction. IEEE Press, pp 109–116
150. Wen J, Stewart A, Billinghurst M, Dey A, Tossell C, Finomore V (2018) He who hesitates is lost (... in thoughts over a robot). In: Proceedings of the technology, mind, and society. ACM, p 43
151. Williams M (2007) Building genuine trust through interpersonal emotion management: a threat regulation model of trust and collaboration across boundaries. Acad Manag Rev 32(2):595–621
152. Wiltshire TJ, Barber D, Fiore SM (2013) Towards modeling social-cognitive mechanisms in robots to facilitate human-robot teaming. In: Proceedings of the human factors and ergonomics society annual meeting, vol 57. SAGE Publications Sage, Los Angeles, CA, pp 1278–1282
153. Wright JL, Chen JY, Lakhmani SG (2019) Agent transparency and reliability in human–robot interaction: the influence on user confidence and perceived reliability. IEEE Trans Hum Mach Syst. https://doi.org/10.1109/THMS.2019.2925717
154. Wynne KT, Lyons JB (2018) An integrative model of autonomous agent teammate-likeness. Theor Issues Ergon Sci 19(3):353–374
155. Xie Y, Bodala IP, Ong DC, Hsu D, Soh H (2019) Robot capability and intention in trust-based decisions across tasks. In: 2019 14th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 39–47
156. Xu A, Dudek G (2015) Optimo: Online probabilistic trust inference model for asymmetric human–robot collaborations. In: Proceedings of the tenth annual ACM/IEEE international conference on human–robot interaction. ACM, pp 221–228
157. Yang XJ, Unhelkar VV, Li K, Shah JA (2017) Evaluating effects of user experience and system transparency on trust in automation. In: 2017 12th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 408–416

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ewart J. de Visser currently serves as the technical director for the Warfighter Effectiveness Research Center at the United States Air Force Academy. He further holds affiliations with Drexel University, George Mason University, Clemson University, and the University of Alabama in Huntsville. Dr. de Visser's research focusses on how to create effective human-robot teams by modeling, measuring and designing relationships between people and autonomous machines. He holds a PhD in Human Factors Psychology from George Mason University and a BA in Film Studies from the University of North Carolina Wilmington.

Dr. Marieke M. M. Peeters is a senior research scientist in Artificial Intelligence at TNO, the Netherlands Organization for Applied Scientific Research. She works in the department of Perceptual and Cognitive Systems, as part of the research team on human-agent-robot teaming. Furthermore, she is a co-founder of the human-agent teaming lab at TNO, and is responsible for the research agenda of various research projects focusing on human-agent teaming, including topics such as work agreements, team mental models, dynamic task allocation, proactive communication, emergent teams, and trust calibration.

Malte F. Jung is an Assistant Professor in Information Science at Cornell University and the Nancy H. '62 and Philip M. '62 Young Sesquicentennial Faculty Fellow. He leads the Robots in Groups Lab, which explores not only how robots influence the dynamics of groups and teams, but also how robots can be designed to shape group dynamics in such a way that outcomes improve. Malte Jung holds a PhD in Mechanical engineering and a PhD Minor in Psychology from Stanford University.

Spencer Kohn received a M.A. in Human Factors Psychology from George Mason University in 2016. He is currently pursuing a Ph.D. in Human Factors Psychology at George Mason, focusing on trust repairs administered by automation.

Tyler H. Shaw is an Associate Professor in the Human Factors and Applied Cognition Program at George Mason University. He received his PhD in Experimental Psychology from the University of Cincinnati in 2008.

Richard Pak is a professor of psychology at Clemson University. He received his PhD in psychology from the Georgia Institute of Technology.

Mark A. Neerincx is full professor Human-Centered Computing at the Delft University of Technology and principal scientist Perceptual & Cognitive Systems at TNO (Netherlands organisation of applied research). Recent projects focus on the socio-cognitive engineering of artificial, virtual or physical, agents (ePartners) that show social, cognitive and affective behaviours to enhance performance, resilience, health and/or wellbeing. ePartner prototypes are being developed (1) for sharing situation awareness, harmonizing workload distributions and supporting stress-coping in high risk domains (e.g., robot-assisted disaster response teams), and (2) for continual learning and behavior support (e.g., robotic partners for diabetes self-management or elderly care).