Behavioral Measures of Trust in Human-autonomy Teams

DANIEL ALEJANDRO GONZÁLEZ RUEDA, Rensselaer Polytechnic Institute, USA
DAVID PIORKOWSKI, IBM Research, USA
DAVID MENDONÇA, MITRE Corporation, USA

Trust has long been acknowledged as a crucial aspect of teamwork, whether in all-human or in mixed human/autonomy teams. However, typical approaches to the measurement of trust rely chiefly on psychometric approaches that are not well suited to capturing data on trust among non-human members of a team and can constitute interference in the workflow of all-human teams. This paper explores prospects for conceptualizing and measuring trust at the team level through the measurement of observable behaviors associated with trust. Here, three aspects of trust are considered—competence, predictability, and integrity–and existing behavioral measures of trust are examined in relation to them, using as criteria the reliability, validity, and extensibility of each measure. This paper concludes with a summative assessment of the current state of behavioral measures on trust in teams, as well as recommendations for future work. Further research along these lines will be critical for understanding the role of trust in human/autonomy teams in general, but particularly when the proportion of non-human members on a team is large, or when "autonomies" participate in vital activities in the team's workflow.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**; *HCI design and evaluation methods.*

Additional Key Words and Phrases: behavioral measures, trust, trust aspects, human-agent teams (HATs), human-agent interaction

ACM Reference Format:
Daniel Alejandro González Rueda, David Piorkowski, and David Mendonça. 2022. Behavioral Measures of Trust in Human-autonomy Teams. In *New Orleans '22: Workshop on Trust and Reliance in AI-Human Teams, April 30, 2022, New Orleans, LA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Trust— which may be defined as the belief or expectation that a team will plan and execute with a certain consistency and quality during the development of task while maintaining contracts agreed by team members [20] —is at once a critical factor in teamwork [15, 24] and one that is inherently unobservable—at least via direct methods [11]. Not surprisingly, the measurement of trust in human teams has tended to rely on psychometric instruments administered before and/or after teamwork has been undertaken [29, 30].

In human/machine teams, the measurement of trust has generally been restricted to trust as experienced by human members. There are exceptions, however, including behavioral measures that rely either on automated methods for data collection (e.g., via sensors) or on observation by human observers. An issue with these measures is that they are seldom tied into rigorously defined theoretical conceptualizations of trust, thereby creating a gap with the rich base of theory that informs work on human teams.

The psychometric, questionnaire-driven approach is known to be limiting in situations where the ebb and flow of trust during teamwork is the object of inquiry [6]. In these situations, pre/post measures of trust offer minimal insights.

Manuscript submitted to ACM

On the other hand, automated methods hold out the prospect of capturing moment-to-moment actions and interactions, yet typically lack a strong and validated tie back to basic theories of trust. Furthermore, even though non-human team members lack specific cognitive and social functions for trust, behaviors that arise from the interactions between both human and non-human team members provide useful quantitative measures that can be used to further the analysis of trust for a team level measure.

This paper unites these two strains of research—theoretically grounded perspectives on trust and behavioral measures of trust—through a review and synthesis of a curated selection of empirical studies in the area of human-autonomy teams (HATs). The main contributions are (i) the connection of theoretical conceptualizations of trust with behavioral measures of trust; (ii) a critique of those measures in terms of reliability, validity, and extensibility; and (iii) conclusions and recommendations regarding the further use and development of behavioral measures of trust in HATs.

Section 2 defines three main aspects of trust as well as approaches to evaluating behavioral measures of those aspects. Section 3 describes the methods used in selecting experimental tasks in which behavioral measures of trust were used. Section 4 presents an analysis, classification, and evaluation of these measures with respect to the criteria of reliability, validity and extensibility. Section 5 provides recommendations on how these measures can be improved and deployed in future empirical studies.

2 BACKGROUND

Trust in human-autonomy teams (HATs) is here conceptualized as extending to all members of the team, whether human or not [12]. Trust has many aspects (e.g., based in affect or cognition) and is dependent on a wide range of precursors [13, 15, 28]. Affect-based trust is defined as trust that arises from social and affective characteristics related to the moral sense and intentions of another person, group or institution [22, 24]. Different aspects compose this type of trust, including honesty and integrity, benevolence and faith [24]. Cognition-based trust is based on the competence and predictability of group or team members [22].

These three aspects of trust are indicative of others in emphasizing the role of underlying psychological phenomena in shaping trust, meaning that these aspects are inherently human-centered. Yet for trust among all HAT members to be understood, it must be conceptualized and measured in a way that captures trust across all agents in the HAT, not merely among its human members, for which there is a diversity of operationalizations of trust [14, 18, 31]. The categorization of these aspects is based in both previous research [15, 20] and the analysis of characteristics that have been explored through quantitative measurements [11, 33] based on actions of both human and non-human team members.

The remainder of this section briefly presents the specific aspects of trust considered here, and introduces the criteria used to assess behavioral measures associated with those aspects.

2.1 Aspects of Trust

Three aspects of trust—integrity (from the affective perspective) and competence and predictability (both from the cognitive perspective) are considered here in order to provide a degree of breadth.

(1) *Competence.* Competence is defined as the ability of a team member or agent to perform a series of required functions or actions under given constraints to further the objective(s) of a given task [13, 16, 20, 25]. Competence therefore relates directly to an ability to perform: that is, some function of actual performance is used as a proxy for competence. Competence is associated with the cognitive perspective on trust, and has been shown

to influence a human operator´s trust in automation [10, 13]. This aspect of trust in the HAT and psychology literatures is sometimes framed as reliability [8], ability [3, 21], or capability [1]. This aspect is often expressed as the number or rate of failures during task performance [9, 11]. Recent work has introduced finer-grained measures based on physical movements [2].

(2) *Predictability* is defined as the consistency of a team member's behaviors and the extent to which these behaviors conform to the expectations of other team members [7, 19, 20, 25]. Related conceptualizations are variability [16] and reliability [1]. Predictability in HATs has been modeled in a probabilistic framework as changes in behavior based on particular stimuli [5, 17, 32].

(3) *Integrity* is defined as the capacity of an agent to adhere to agreements among team members [20]. For a non-human team member, an agreement is represented by constraints on the actions and performance of team members [23]. Integrity has been estimated from behaviors presented in economic game setups where a person can make an agreement on the amount of money that s/he will give to another and how frequently the commitment to give is fulfilled [35].

2.2 Evaluation of Behavioral Measurements of Trust

Consistent with prior research on the development of survey instruments, we evaluate measures based on the criteria of reliability and validity, adding the criterion of extensibility as a way of assessing prospects for further development of measures [4].

(1) *Validity* is an encompassing term that refers to the ability of a measurement instrument to capture or otherwise express the value of a theoretical (and typically unobservable) concept [4]. One of the more widely addressed types of validity is construct validity, which expresses the extent to which an operationalization measures the true value of some construct/concept [4]. In HATs, one approach to the assessment of construct validity is to compare data obtained through simulations of a given task with behavioral data obtained in a real-life task [2].

(2) *Reliability* addresses the degree to which results obtained from a measure are consistent through repeated measurements and/or trials [4]. In behavioral terms, reliability may be understood as the extent to which behavioral responses to fixed stimuli are consistent.

(3) *Extensibility* refers to the ability of a system or measurement to be extended through new functionality or in new areas of application [26]. In behavioral terms, extensibility is analyzed as the range of behavioral responses that can be introduced and adapted to a measure.

3 METHODOLOGY

This section describes the criteria used to identify, classify and assess behavioral measures of the competence, predictability and integrity aspects of trust, using the three criteria (validity, reliability, and extensibility) described previously.

3.1 Identification of Candidate Behavioral Measures

The area of focus for the identification of papers was human-autonomy teams/teaming, an expansive literature that contains a wide variety of quantitative measures on teams and team performance. The approach taken was to identify papers that had trust as an important focus, and where behavioral measures were used to estimate or infer trust. The search covered journal and conference papers published from 2010 to 2022 and listed in Google Scholar, Scopus,

Academic Search Premier, and ACM. The queries used were as follows, with 'ABS:' indicating a search limited to the contents of the abstract:

(1) "human robot interaction AND trust" (2) "[[ABS: human robot team] OR [ABS: human agent team]] AND [ABS: trust]" (3) "human robot collaboration AND trust" (4) "human agent collaboration AND trust" (5) "trust inference model AND human robot collaboration" (6) "trust inference model AND human-robot collaboration".

These queries yielded approximately 100 papers that were clearly experiment-based. The next step was to determine whether behavioral measures of trust were used. Only studies that included at least one non-human team member were retained. This step was accomplished by reading the Abstract and Methodology sections, resulting in 15 papers. These were all classified using the criteria below. A subset were then selected in order to provide a reasonable and manageable degree of coverage across all combinations of classification criteria.

It should be emphasized that this selection of studies is not meant to be comprehensive nor to be necessarily representative of the literature as a whole. Instead, the selected studies are illustrative, in that they may be used to explore the prospects for developing and assessing behavioral measures of trust. A comprehensive examination of the literature would complement the present study by providing an assessment of the state of the art of the field—an item for future work.

3.2 Classification of Measures on Aspects of Trust

For the selected papers, we searched for the presence or absence of measures associated with the three target aspects of trust. As suggested by the foregoing discussions, the terminology across different studies differed, though the measures may have been operationally similar (or *vice versa*). We began with the definitions given previously as the filter through which we classified the measures we found.

(1) *Competence* was reflected in a measure to the extent that one or more of the following properties were expressed by it: (a) a demonstrated ability to undertake a task, (b) the presence of an objective which team members pursued, or (c) the presence of task-related constraints.. If the measurement had these three elements then there was the element of competence as a aspect of trust in the study.
(2) *Predictability* was reflected in a measure to the extent that it expressed a notion of behavioral consistency given fixed inputs. One approach is to consider the history of behaviors performed by a team member in response to fixed experimental stimuli. These histories of interaction are used as inputs for a probability function that expresses an expected value that is used to define the consistency [1, 33].
(3) *Integrity* was reflected in a measure if it captured some notion of adherence to agreements between team members. Of course, agreements may vary considerably in form and function across different tasks. Two examples are in maintaining agreements (or information) during task performance (e.g., how often information shared by an agent was true and how often did the agent change the information previously presented [32].

3.3 Evaluation of Behavioral Measurements for Trust

Measures of each aspect of trust were assessed using the criteria of reliability, validity, and extensibility, using the approaches described below.

(1) *Reliability* was assessed by examining the Results and Discussion sections of each paper. For these sections we considered the relationship between stimuli and behavioral outcomes, the number of simulation runs or experimental trials, and any replications of the experiments.

(2) *Validity* was assessed by examining any evidence of typical validation activities as described in the Methodology and Results sections of the selected papers. This included searching for proof of comparisons performed between the data obtained with behavioral data taken from behaviors of people or the results from other measures of trust, among other approaches.
(3) *Extensibility* was assessed by examining the Methodology, Results and Discussion sections for any evidence that the study represented extensions from work in others domains, or that there were claims or proof that the measurement approach could be (or had been) applied in different settings.

4 RESULTS

The results of the classification and assessment of the measures are given in Table 1. Task domains included semi-automated vehicles [1], cyber-physical systems [33], robot patrol [27], shared-control systems [2], and search [11]. The remainder of this section presents an analysis of the measures for the different aspects of trust with respect to the evaluation criteria of reliability, validity, and extensibility.

4.1 Measures of Trust

Each aspect of trust was found in at least one experiment, with competence found in all experiments. Predictability was found in all but one study, while Integrity was present in only one study. (While formal classification of the full set of 15 experiments originally identified is not presented here, this distribution of results approximately mirrors that of the full set.) The following analysis of the measures makes reference to the experiments as labeled in Table 1 (e.g., M2 refers to the experiment in [33], as indicated in the *Citation* column).

(1) *Competence* is reflected in the performance of the team or team members, and was present in measures of trust in all experiments. M1 analyzed the performance as the capacity or ability of an agent to perform or not a task depending on the requirements of the task for driving a car. M2 measured performance of the agent as the length of the routes developed with the objective of minimizing them while avoiding obstacles. M3 measured performance by analyzing the capacity of the agent to transmit correct information to other nodes with a limitation in the range of each agent. M4 measured performance as the amount of time a robot takes to patrol a certain area. M5 analyzed performance as the capacity of the human controller to follow a reference trajectory obtained from an optimization algorithm without hitting obstacles. Finally, M6 measured performance as the amount of destinations a robot has to reach under a safety and time parameter defined by the human operator.
(2) *Predictability* in the experiments was presented as the expected value given by a probability function of the possible behaviors depending on the task. All experiments but one (M2) included at least one measure that reflected predictability. M1 and M4 present predictability as the belief of one agent in the ability to succeed in a task depending on the history of tasks completed and the capacities of the agent. M3 proposes predictability as a part of the integrity measurement and is considered as the variance on the quality of the information transmitted to other agents. M5 analyzes predictability as the probability that a human operator will deviate from the reference trajectory based on an optimization algorithm. M6 presents predictability as the amount of trustworthy and untrustworthy behaviors classified using a Case-base Reasoning algorithm.
(3) *Integrity* was identified in M3 as the adherence to contracts by agents. This was done in the experiment with the history of information transmitted to other notes and the amount of time a given agent changed the type

of information sent with respect to previous information. As mentioned previously, no other tasks included a measure that reflected this aspect of trust.

4.2 Assessment Criteria

Assessment criteria were treated for the measures of each experiment in general, as opposed to being applied to each measure. In other words, this section discusses the use of techniques for assessing reliability, validity and extensibility in each experiment, regardless of the specific aspect of trust being measured.

Each assessment criterion was found at least once in each of the selected experiments. In particular, extensibility was found in all the experiments. On the other hand, reliability was assessed in four of the experiments and validity in two of the experiments. Table 1 contains the results of these assessments for each study.

(1) *Reliability* of measures was assessed in M3 through M6. Specifically, M3 and M6 used deterministic parameters with the initial values defined by the experimenters and simulations showed the consistency of the results. M4 and M5 used the inputs obtained from the movements of the team agents and through replications these were analyzed for consistency.
(2) *Validity* of measures was assessed in M4 and M5. In M4, data on performance of robots in a synthetic (simulation) environment was compared with data on the behavior of robots in a physical space, specifically the paths taken in a patrolling task. Furthermore, the concepts were obtained from previously defined aspects that affect trust (such as competence) using performance as a proxy [34]. In M5 the model was validated by comparing the results of simulations done to analytic results (via control theory) concerning the trajectories directed by an operator, with different levels of automation controlled by the agent teammate for a crane.
(3) *Extensibility* of measures was assessed to some extent in all the experiments. Behavioral measures of trust included concepts (such as performance) that can be redefined and modified to a variety of tasks. For example, in M1 the measure of trust is defined using capabilities of the team members and requirements of the task. These two aspects are used in different situations and can be defined by a considerable amount of problems. M1 required defining the capabilities and requirements necessary for driving a vehicle. However, this approach can also be used in a search task by analyzing the capabilities and necessary requirements and using those as inputs for the trust measure.

The following section presents an analysis of insights obtained through the analysis of these behavioral measurements of trust.

5 DISCUSSION AND CONCLUSIONS

In this section we draw some general conclusions from the foregoing results, focusing on prospects for further development and assessment of behavioral measures of trust in HATs. Both conceptually and practically, the further development of behavioral measures of trust faces two main challenges. First, typical (i.e., questionnaire-based) approaches to measuring trust cannot be applied to machines. Second, behavioral measures (whether on human or machine behavior) must always be taken as proxies for trust, and indeed ones that (on the human side) are more distant than those obtained via questionnaires. Further work on methods for assessing measure validity, reliability and extensibility (perhaps along with other criteria) will therefore be essential in ongoing measure development.

Table 1. Classification of behavioral measures of trust

ID	Task	Trust Measure	Aspects of Trust			Assessment of Measures			Citation
			C	P	I	R	V	E	
M1	Car Driving	Function the probability of an agent to perform a task given the task requirements an the agent's own capabilities.	Yes	Yes	No	No	No	Yes	[1]
M2	Path planning	Function of human and robot performance and faults of cooperation.	Yes	No	No	No	No	Yes	[33]
M3	Cyber-physical systems	Independent measurements of ability, benevolence and integrity	Yes	Yes	Yes	Yes	No	Yes	[32]
M4	Robot Patrolling	Belief about the capacities (competence) of an agent given a history of interactions (predictability).	Yes	Yes	No	Yes	Yes	Yes	[27]
M5	Shared-control System	Combination of predictability (consistency of trajectories) and competence (deviation of movements)	Yes	Yes	No	Yes	Yes	Yes	[2]
M6	Search Task	Combination of competence (times a robot succeeds, fails or is interrupted) and predictability (repetition or not of trustworthy behaviors).	Yes	Yes	No	Yes	No	Yes	[11]

The following abbreviations are used in the table: (1) C: Competence (2) P: Predictability (3) I: Integrity (4) R: Reliability (5) V: Validity (6) E: Extensibility.

5.1 Performance as a Proxy for Competence in Behavior Measures

The use of (outcome) performance as a proxy for competence means that the inputs for a measure of trust are obtained once task performance has concluded. The implication is that to analyze the competence of an agent there has to be a clear way to evaluate task performance. When the objective is clear, this can be relatively straightforward. However, it is precisely in situations where the task is less well-defined that human/autonomy collaboration will be needed.

5.2 Predictability Based on History of Behaviors

The assessment of predictability is likely to require access to a prior history of performance on the same task or on similar tasks. With human agents the history of behaviors may be obtained by comparing the relationship between task inputs and outputs over time. However, in certain tasks there are difficulties in the analysis of the reasoning behind actions. For example, the reasoning of human agents might vary considerably even across similar tasks. On the other hand, machines use algorithms previously defined when performing actions during a task.

5.3 Integrity for Humans and Agents in a Team

The measures of the integrity aspect of trust considered here follow from an explicit model of the beliefs and intentions of (human or synthetic) team members. Nevertheless, there are difficulties in the methods used to analyze the intentions of human agents. These intentions are executed through thought processes that vary for each human agent and are not easily obtainable unless a behavior is generated. Furthermore, the expectations for and adherence to contracts for humans might be neither objective nor optimal. For example, one human might feel discouraged to work with a robot that does not adhere to all the agreements, while another human agent might find the adherence to all contracts irrelevant as long as the task as a whole is done successfully. In either case the difficulty lies in which method to use for the analysis of the beliefs and intentions generated by other team members. Ín contrast, synthetic agents present an advantage as adherence to contracts can be modeled as a function of task requirements.

5.4 Assessment of Behavioral Measures of Trust

The assessment of behavioral measures of trust presents a number of challenges and, therefore, opportunities for methodological innovation. At a conceptual level, further work may be needed in distinguishing competence from predictability in that both aspects are connected via behaviors associated with task performance, whether in the moment (in the case of competence) or over repeated instances of task performance (in the case of predictability). Predictability and integrity can also be linked, given that the consistency of behaviors presented in team members' history of interactions may be associated with the degree to which they have adhered to agreements over time. Finally, extensibility can also be informed by aspects of team performance, including errors.

A second broad area of methodological challenge is in the development of structured, commonly accepted techniques for measure assessment, particularly for the critical criteria of validity and reliability. For example, over many decades of research, tools such as confirmatory factor analysis have been developed for validating survey instruments. One general approach is to assess the degree to which items (i.e., survey questions) associated with the same concept are highly correlated with each other, and orthogonal to items associated with concepts that are quite different from them. There may be opportunities for pursuing similar approaches with behavioral measures (e.g., through the development and analysis of multiple measures of a given aspect of trust).

Finally, a third area is in the further development of measures of trust that are not merely simple aggregations of values obtained from individual measures members of the team. When subjective measures of trust (such as questionnaires) are used, computation of the team-level measure is typically achieve via linear combination of the individual responses (e.g., by computing the mean). Clearly this approach breaks down conceptually and empirically when the number of human members of the team is considerably smaller than the number of non-human members. Measures are needed that engage team-level phenomena (such as shared situation awareness, workflow, and transactive memory) and that will be valid regardless of the number of human vs. non-human members of the team.

5.5 Summary

This paper has explored prospects for measuring three aspects of trust (competence, predictability, and integrity) in behavioral terms, using as evaluation criteria the reliability, validity and extensibility of these measures. This review has identified numerous opportunities for further development of measures, as well as some of the conceptual and methodological challenges associated with doing so. Although measures of trust associated with psychometric instruments are, in general, far more advanced than those associated with human behavior, they are clearly impossible to implement with synthetic members of a team. However, non-human team members will still have to take decisions that affect the team which requires operalizations of trust and/or aspects of trust that reflect the application of satisfactory and ethical decisions by the non-human members of the team. Put simply, if we are to understand the evolution of trust in HATs as a whole, then we must continue to explore behavioral measures for doing so. This need becomes even more urgent if the goal is to understand how trust evolves during task execution, when stopping the action to ask participants questions can easily introduce interruptions and confounds into task performance. Indeed, as the sophistication, portability, and precision of on-board sensors (whether affixed to humans or robots) continue to grow, we envision a dramatically expanded set of tools for achieving high degrees of precision in estimating trust.

ACKNOWLEDGMENTS

This work was supported by the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons).

REFERENCES

[1] Hebert Azevedo-Sa, X Jessie Yang, Lionel Robert, and Dawn Tilbury. 2021. A Unified bi-directional model for natural and artificial trust in human-robot collaboration. IEEE Robotics and Automation Letters (2021).

[2] Alexander Broad, Jarvis Schultz, Matthew Derry, Todd Murphey, and Brenna Argall. 2016. Trust adaptation leads to lower control effort in shared control of crane automation. IEEE Robotics and Automation Letters 2, 1 (2016), 239–246.

[3] Christopher S Calhoun, Philip Bobko, Jennie J Gallimore, and Joseph B Lyons. 2019. Linking precursors of interpersonal trust to human-automation trust: An expanded typology and exploratory experiment. Journal of Trust Research 9, 1 (2019), 28–46.

[4] Edward G Carmines and Richard A Zeller. 1979. Reliability and Validity Assessment. Sage publications.

[5] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S Melo, and Ana Paiva. 2018. Exploring the impact of fault justification in human-robot trust. In Proceedings of the 17th international conference on autonomous agents and multiagent systems. 507–513.

[6] Sylvain Daronnat, Leif Azzopardi, Martin Halvey, and Mateusz Dubiel. 2019. Human-agent collaborations: Trust in negotiating control. CHI 2019 (2019).

[7] Sylvain Daronnat, Leif Azzopardi, Martin Halvey, and Mateusz Dubiel. 2021. Inferring trust From users' behaviours: Agents' predictability positively affects trust, task performance and cognitive load in human-agent real-time collaboration. Frontiers in Robotics and AI 8 (2021), 194.

[8] Munjal Desai, Mikhail Medvedev, Marynel Vázquez, Sean McSheehy, Sofia Gadea-Omelchenko, Christian Bruggeman, Aaron Steinfeld, and Holly Yanco. 2012. Effects of changing reliability on trust of robot systems. In 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 73–80.

[9] Balbir S Dhillon and ARM Fashandi. 1997. Safety and reliability assessment techniques in robotics. Robotica 15, 6 (1997), 701–708.

[10] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. International journal of human-computer studies 58, 6 (2003), 697–718.

[11] Michael W Floyd, Michael Drinkwater, and David W Aha. 2016. Learning trustworthy behaviors using an inverse trust metric. In Robust Intelligence and Trust in Autonomous Systems. Springer, 33–53.

[12] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. 2007. Measurement of trust in human-robot collaboration. In 2007 international symposium on collaborative technologies and systems. IEEE, 106–114.

[13] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals 14, 2 (2020), 627–660.

[14] Yaohui Guo and X Jessie Yang. 2020. Modeling and predicting trust dynamics in human-robot teaming: A Bayesian inference approach. International Journal of Social Robotics (2020), 1–11.

[15] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. Human factors 53, 5 (2011), 517–527.

[16] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors 57, 3 (2015), 407–434.

[17] Trung Dong Huynh, Nicholas R Jennings, and Nigel R Shadbolt. 2006. An integrated trust and reputation model for open multi-agent systems. Autonomous Agents and Multi-Agent Systems 13, 2 (2006), 119–154.

[18] Retno Larasati, Anna De Liddo, and Enrico Motta. 2020. The Effect of Explanation Styles on User's Trust. In ExSS-ATEC@ IUI.

[19] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. Ergonomics 35, 10 (1992), 1243–1270.

[20] Peter R Lewis and Stephen Marsh. 2022. What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. Cognitive Systems Research 72 (2022), 33–49.

[21] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. Academy of management review 20, 3 (1995), 709–734.

[22] Daniel J McAllister. 1995. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. Academy of management journal 38, 1 (1995), 24–59.

[23] Nathan J McNeese, Mustafa Demir, Erin K Chiou, and Nancy J Cooke. 2021. Trust and team performance in human–autonomy teaming. International Journal of Electronic Commerce 25, 1 (2021), 51–72.

[24] Bonnie M Muir. 1987. Trust between humans and machines, and the design of decision aids. International journal of man-machine studies 27, 5-6 (1987), 527–539.

[25] Sherry Ogreten, Stephanie Lackey, and Denise Nicholson. 2010. Recommended roles for uninhabited team members within mixed-initiative combat teams. In 2010 International symposium on collaborative technologies and systems. IEEE, 531–536.

[26] David Lorge Parnas. 1979. Designing software for ease of extension and contraction. IEEE transactions on software engineering 2 (1979), 128–138.

[27] Charles Pippin and Henrik Christensen. 2014. Trust modeling in multi-robot patrolling. In 2014 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 59–66.

[28] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. Human factors 58, 3 (2016), 377–400.

[29] Stephen J Selcon, Ronald M Taylor, and Eva Koritsas. 1991. Workload or situational awareness?: TLX vs. SART for aerospace systems design evaluation. In Proceedings of the Human Factors Society Annual Meeting, Vol. 35. SAGE Publications Sage CA: Los Angeles, CA, 62–66.

[30] Indramani L Singh, Robert Molloy, and Raja Parasuraman. 1993. Automation-induced" complacency: Development of the complacency-potential rating scale. The International Journal of Aviation Psychology 3, 2 (1993), 111–122.

[31] Jennifer Wang and Angela Moulden. 2021. AI trust score: A user-centered approach to building, designing, and measuring the success of intelligent workplace features. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. 1–7.

[32] Yan Wang. 2018. Trust quantification for networked cyber-physical systems. IEEE Internet of Things Journal 5, 3 (2018), 2055–2070.

[33] Yue Wang, Laura R Humphrey, Zhanrui Liao, and Huanfei Zheng. 2018. Trust-based multi-robot symbolic motion planning with a human-in-the-loop. ACM Transactions on Interactive Intelligent Systems (TiiS) 8, 4 (2018), 1–33.

[34] Anqi Xu and Gregory Dudek. 2015. Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 221–228.

[35] Sebastian Zörner, Emy Arts, Brenda Vasiljevic, Ankit Srivastava, Florian Schmalzl, Glareh Mir, Kavish Bhatia, Erik Strahl, Annika Peters, Tayfun Alpay, et al. 2021. An immersive investment game to study human-robot trust. Frontiers in Robotics and AI 8 (2021), 139.