The Role of Trust in Human-Robot Interaction

Michael Lewis[1], Katia Sycara[2], and Phillip Walker[1]

[1] Department of Information Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. (ml@sis.pitt.edu,pmw19@pitt.edu)
[2] Robotics Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. (katia@cs.cmu.edu)

1 Introduction

Robots and other complex autonomous systems offer potential benefits through assisting humans in accomplishing their tasks. These beneficial effects, however, may not be realized due to maladaptive forms of interaction. While robots are only now being fielded in appreciable numbers, a substantial body of experience and research already exists characterizing human interactions with more conventional forms of automation in aviation and process industries.

In human interaction with automation, it has been observed that the human may fail to use the system when it would be advantageous to do so. This has been called *disuse (underutilization or under-reliance)* of the automation [1]. People also have been observed to fail to monitor automation properly (e.g. turning off alarms) when automation is in use, or they accept the automation's recommendations and actions when inappropriate [2, 1]. This has been called *misuse, complacency, or over-reliance*. Disuse can decrease automation benefits and lead to accidents if, for instance, safety systems and alarms are not consulted when needed. Another maladaptive attitude is automation bias [3–7], a user tendency to ascribe greater power and authority to automated decision aids than to other sources of advice (e.g. humans). When the decision aid's recommendations are incorrect, automation bias may have dire consequences [8–11] (e.g. errors of omission, where the user does not respond to a critical situation, or errors of commission, where the user does not analyze all available information but follows the advice of the automation).

Both naïve and expert users show these tendencies. In [12], it was found that skilled subject matter experts had misplaced trust in the accuracy of diagnostic expert systems. (see also [13]). Additionally the Aviation Safety Reporting System contains many reports from pilots that link their failure to monitor to excessive trust in automated systems such as autopilots or FMS [14, 15]. On the other hand, when corporate policy or federal regulations mandate the use of automation that is not trusted, operators may "creatively disable" the device [16]. In other words: disuse the automation.

Studies have shown [17, 18] that trust towards automation affects reliance (i.e. people tend to rely on automation they trust and not use automation they do not trust). For example, trust has frequently been cited [19, 20] as a contributor to human decisions about monitoring and using automation. Indeed,

within the literature on trust in automation, complacency is conceptualized interchangeably as the overuse of automation, the failure to monitor automation, and lack of vigilance [21–23]. For optimal performance of a human-automation system, *human trust in automation should be well-calibrated.* Both disuse and misuse of the automation has resulted from improper calibration of trust, which has also led to accidents [1, 24].

In [25], trust is conceived to be an "attitude that an agent (automation or another person) will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." A majority of research in trust in automation has focused on the relation between automation reliability and operator usage often without measuring the intervening variable, trust. The utility of introducing an intervening variable between automation performance and operator usage, however, lies in the ability to make more precise or accurate predictions with the intervening variable than without it. This requires that trust in automation be influenced by factors in addition to automation reliability/performance. The three dimensional (Purpose, Process, & Performance) model proposed by Lee & See [25], for example, presumes that trust (and indirectly, propensity to use) is influenced by a person's knowledge of what the automation is supposed to do (purpose), how it functions (process), and its actual performance. While such models seem plausible, support for the contribution of factors other than performance has typically been limited to correlation between questionnaire responses and automation use. Despite multiple studies in trust in automation, the conceptualization of trust and how it can be reliably modeled and measured is still a challenging problem.

In contrast to automation where system behavior has been pre-programmed and the system performance is limited to the specific actions it has been designed to perform, autonomous systems/robots have been defined as having intelligence-based capabilities that would allow them to have a degree of self governance, which enables them to respond to situations that were not pre-programmed or anticipated in the design. Therefore, the role of trust in interactions between humans and robots is more complex and difficult to understand.

In this chapter, we present the conceptual underpinnings of trust in Section 2, and then discuss models of, and the factors that affect, trust in automation in Sections 3 and 4, respectively. Next, we will discuss instruments of measuring trust in Section 5, before moving on to trust in the context of human-robot interaction (HRI) in Section 6 both in how humans influence robots, and vice versa. We conclude in Section 7 with open questions and areas of future work.

2 Conceptualization of Trust

Trust has been studied in a variety of disciplines (including social psychology, human factors, and industrial organization) for understanding relationships between humans or between human and machine. The wide variety of contexts within which trust has been studied leads to various definitions and theories of trust. The different context within which trust has been studied has led to

definitions of trust as an attitude, an intention, or a behavior [26–28]. Both within the inter-personal literature and human-automation trust literature, a widely accepted definition of trust is lacking [29]. However, it is generally agreed that trust is best conceptualized as a *multidimensional psychological attitude* involving beliefs and expectations about the trustee's trustworthiness derived from experience and interactions with the trustee in situations involving uncertainty and risk [30]. Trust has also been said to have both *cognitive and affective features*. In the interpersonal literature, trust is also seen involving affective processes, since trust development requires seeing others as personally motivated by care and concern to protect the trustor's interests [31]. In the automation literature, cognitive (rather than affective) processes may play a dominant role in the determination of trustworthiness, i.e., the extent to which automation is expected to do the task that it was designed to do [32]. In the trust in automation literature, it has been argued that trust is best conceptualized as an attitude [25] and a relatively well accepted definition of trust is: "...an attitude which includes the belief that the collaborator will perform as expected, and can, within the limits of the designer's intentions, be relied on to achieve the design goals" [33].

3 Modeling Trust

The basis of trust can be considered as a set of attributional abstractions (trust dimensions) that range from the trustee's competence to its intentions. [32] combined the dimensions of trust from two works ([34] and [35]). Barber's model [34] is in terms of human expectations that form the basis of trust between human and machine. These expectations are persistence, technical competency, and fiduciary responsibility. Although in the subsequent literature, the number and concepts in the trust dimensions vary [25], there seems to be a convergence on the three dimensions—*Purpose, Process, and Performance* [25]—mentioned earlier, along with correspondences of those to earlier concepts, such as the dimensions in [34], and those of *Ability, Integrity, and Benevolence* [26]. *Ability* is the trustee competence in performing expected actions, *benevolence* is the trustee intrinsic and positive intentions towards the trustor, and *integrity* is trustee's adherence to a set of principles that are acceptable to the trustor. [26].

Both trust in automation [17] and interpersonal relations literature [36–39] agree that trust relations are *dynamic* and varying over time. There are three phases that characterize trust over time: *trust formation*, where trustors choose to trust trustees and potentially increase their trust over time, *trust dissolution*, where trustors decide to lower their trust in trustees after a trust violation has occurred, and *trust restoration* where trust stops decreasing after a trust violation and gets restored (although potentially not to the same level as before the trust violation). Early in the relationship, the trust in the system is based on the predictability of the system's behavior. Work in the literature has shown shifts in trust in response to changes in properties and performance of the automation [32, 19]. When the automation performed reliably, operator trust increased over time and vice versa. Varying levels of trust were also positively correlated with

the varying levels of automation use. As trust decreased, for instance, manual control became more frequent. As the operator interacts with the system, he/she attributes dependability to the automation. Prolonged interaction with the automation leads the operator to make generalizations about the automation and broader attributions about his belief in the future behavior of the system (faith). There is some difference in the literature as to when exactly faith develops in the dynamic process of trust development. Whereas [35] argue that interpersonal trust progresses from predictability to dependability to faith, [17] suggest that for trust in automation, faith is a better predictor of trust early rather than late in the relationship.

Some previous work has explored trust with respect to *automation vs. human trustee* [18]. Their results indicate (a) the dynamics of trust are similar, in that faults diminish trust both towards automation or another human, (b) the sole predictor of reliance on automation was the difference between trust and self-confidence, and (c) participants, in human-human experiments, were more likely to delegate a task to a human when the human was thought to have a low opinion of their own trustworthiness. In other words, when participants thought their own trustworthiness in the eyes of others was high, they were more likely to retain control over a task. However, trustworthiness played no role when the collaborative partner was an automated controller, i.e. only participants' own confidence in their performance determined their decision to retain/obtain control. Other work on trust in humans vs. trust in automation [40] explored the extent to which participants trusted identical advice given by an expert system under the belief that it was given by a human or a computer. The results of these studies were somewhat contradictory however. In one study, participants were more confident in the advice of the human (though their agreement with the human advice did not vary vs. their agreement on the expert system's advice), while in the second study, *participants agreed more with the advice of the expert system, but had less confidence in the expert system.* Similar contradictory results have been shown in HRI studies, where work indicated that errors by a robot did not affect participants' decisions of whether or not to follow the advice of a robot [41], yet did affect their subjective reports of the robot's reliability and trustworthiness [42]. Study results by [2], however, indicated that reliance on a *human* aid was reduced in situations of higher risk.

4 Factors Affecting Trust

The factors that are likely to affect trust in automation have generally been categorized as those pertaining to *automation*, the *operator*, and the *environment*. Most work on factors that have been empirically researched pertains to characteristics of the automation. Here we briefly present relevant work on the most important of these factors.

4.1 System Properties

The most important correlates of use of automation have been system reliability and effects of system faults. Reliability typically refers to automation that has some error rate—for example, misclassifying targets. Typically this rate is constant and data is analyzed using session means. Faults are typically more drastic, such as controller that fails making the whole system behave erratically. Faults are typically single events and studied as time series.

System reliability: Prior literature has provided empirical evidence that there is a relationship between trust in automation and the automation's reliability [33, 43, 23, 1, 44]. Research shows [27] that declining system reliability can lead to systematic decline in trust and trust expectations, and most crucially, these changes can be measured over time. There is also some evidence that only the most recent experiences with the automation affect trust judgments [19, 24].

System faults: System faults are a form of system reliability, but are treated separately because they concern discrete system events and involve different experimental designs. Different aspects of faults influence the relation between trust and automation. Lee and Moray [19] showed that in the presence of continual system faults, trust in the automation reached its lowest point only after six trials, but trust did recover gradually even as faults continued. The magnitude of system faults has differential effects on trust (smaller faults had minimal effect on trust while large faults negatively affected trust and were slower to recover the trust). Another finding [17] showed that faults of varying magnitude diminished trust more than large constant faults. Additionally, it was found that when faults occurred in a particular subsystem, the corresponding distrust did spread to other functions controlled by the same subsystem. The distrust did not, however, spread to independent or similar subsystems.

System predictability: Although system faults affect the trust in the automation, this happens when the human has little *a priori* knowledge about the faults. Research has shown that when people have prior knowledge of faults, these faults do not necessarily diminish trust in the system [43, 18]. A plausible explanation is that knowing that the automation may fail reduces the uncertainty and consequent risk associated with use of the automation. In other words, predictability may be as (or more) important as reliability.

System intelligibility and transparency: Systems that can explain their reasoning will be more likely to be trusted, since they would be more easily understood by their users [45–48]. Such explanatory facility may also allow the operator to query the system in periods of low system operation in order to incrementally acquire and increase trust.

Level of Automation: Another factor that may affect trust in the system is its level of automation (i.e. the level of functional allocation between the human and the system). It has been suggested [20, 32] that system understandability is an important factor for trust development. In their seminal work on the subject [49], Sheridan and Verplank propose a scale for assessing the level of automation in a system from 0-10, with 0 being no autonomy and 10 being fully autonomous. Since higher levels of automation are more complex, thus potentially

more opaque to the operator, higher levels of automation may engender less trust. Some limited empirical work suggests that different levels of automation may have different implications for trust [27]. Their work based on Level 3 [49] automation did not show same results when conducted with Level 7 (higher) automation.

4.2 Properties of the Operator

Propensity to trust: In the sociology literature [50] it has been suggested that people have different propensity to trust others and it has been hypothesized that this is a stable personality trait. In the trust in automation literature, there is very limited empirical work on the propensity to trust. Some evidence is provided in [1] suggests that operator's overall propensity to trust is distinct from trust towards a specific automated system. In other words, it may be the case that an operator has high propensity to trust in automation in general, but faced with a specific automated system, their trust may be very low.

Self Confidence: Self-confidence is a factor of individual difference and one of the few operator characteristics that has been studied in the trust in automation literature. Work in [51] suggested that when trust was higher than self-confidence, automation, rather than manual control would be used and vice versa when trust was lower than self-confidence. However, later work [27], which was conducted with a higher level of automation than [51], did not obtain similar results. It was instead found that trust was influenced by properties of the system (e.g., real or apparent false diagnoses) while self-confidence was influenced by operator traits and experiences (e.g. whether they had been responsible for accidents). Furthermore, it was also found that self-confidence was not affected by system reliability. This last finding was also suggested in the work of [18] which found that self-confidence was not lowered by shifts in automation reliability.

Individual Differences and Culture: It has been hypothesized, and supported by various studies, that individual differences [15, 51–53] and culture [54] affect the trust behavior of people. The interpersonal relations literature has identified many different personal characteristics of a trustor, such as self-esteem [50, 55], secure attachment [56], and motivational factors [57] that contribute to the different stages in the dynamics of trust. Besides individual characteristics, socio-cultural factors that contribute to differences in trust decisions in these different trust phases have also been identified [58–60, 39]. For example, combinations of socio-cultural factors that may result in quick trust formation (also called "swift trust" formation in temporary teams [61]) are time pressure [62] and high power distance with authority [63]. People in high power distance (PD) societies expect authority figures to be benign, competent and of high integrity. Thus people in high power distance societies will engage in less vigilance and monitoring for possible violations by authority figures. To the extent then that people of high PD cultures perceive the automation as authoritative, they should be quick to form trust. On the other hand, when violations occur, people in high PD cultures should be slow to restore trust once violations have occurred [64]. Additionally,

it has been shown [65] via replication of Hofstede's [66] cultural dimensions for a very large-scale sample of pilots, that even in such a highly specialized and regulated profession, national culture still exerts a meaningful influence on attitude and behavior over and above the occupational context.

To date, only a handful of studies consider cultural factors and potential differences in the context of trust in automation, with [67] [68] and [69] being exceptions. As the use of automation gets increasingly globalized, it is imperative that we gain an understanding on how trust in automation is conceptualized across cultures and how it influences operator reliance and use of automation, and overall human-system performance.

4.3 Environmental Factors

In terms of environmental factors that influence trust in automation, risk seems most important. Research in trust in automation suggests that reliance on automation is modulated by the risk present in the decision to use the automation [70]. People are more averse to using the automation if negative consequences are more probable and, once trust has been lowered, it takes people longer to re-engage the automation in high-risk vs. low risk situations [43]. However, knowing the failure behavior of the automation in advance may modify the perception of risk, in that people's trust in the system does not decrease [70].

5 Instruments for Measuring Trust

While a large body of work on trust in automation and robots has developed over the past two decades, standardized measures have remained elusive with many researchers continuing to rely on short idiosyncratically worded questionnaires. Trust (in automation) refers to a cognitive state or attitude, yet it has most often been studied *indirectly through its purported influence on behavior often without any direct cognitive measure*. The nature and complexity of the tasks and failures studied has varied greatly ranging from simple automatic target recognition (ATR) classification [4], to erratic responses of a controller embedded within a complex automated system [51] to robots misreading QR codes [71]. The variety of reported effects (automation bias, complacency, reliance, compliance, etc.) mirror these differences in tasks and scenarios. [72] and [73] have criticized the very construct of trust in automation on the basis of this diversity as an unfalsifiable "folk model" without clear empirical grounding. Although the work cited in the reply to these criticism in [44] as well as the large body of work cited in the review by [23] have begun to examine the interrelations and commonalities of concepts involving trust in automation, empirical research is needed to integrate divergent manifestations of trust within a single task/test population so that common and comparable measures can be developed.

Most "measures" of trust in automation since the original study [17] have been created for individual studies based on face validity and have not in general

benefited from the same rigor in development and validation that has characterized measures of interpersonal trust. "Trust in automation" has been primarily understood through its analogy to interpersonal trust and more sophisticated measures of trust in automation have largely depended on rationales and dimensions developed for interpersonal relations, such as ability, benevolence, and integrity.

Three measures of trust in automation, Empirically Derived (ED), Human-Computer Trust (HTC), and SHAPE Automation Trust Index (SATI) have benefited from systematic development and validation. The Empirically Derived 12 item scale developed by [74] was systematically developed, subjected to a validation study [75] and used in other studies [76]. In [74], they developed their scale in three phases beginning with a word elicitation task. They extracted a 12-factor structure used to develop a 12-item scale based on examination of clusters of words. The twelve items roughly correspond to the classic three dimensions: benevolence (purpose), integrity (process), and ability (performance).

The Human-Computer Trust (HTC) instrument developed in [28] demonstrated construct validity and high reliability within their validation sample and has subsequently been used to assess automation in air traffic control (ATC) simulations, most recently in [77]. Subjects initially identified constructs that they believed would affect their level of trust in a decision aid. Following refinement and modification of the constructs and potential items, the instrument was reduced to five constructs (reliability, technical competence, understandability, faith, and personal attachment). A subsequent principal components analysis limited to five factors found most scale items related to these factors.

The SHAPE Automation Trust Index, SATI, [78] developed by the European Organization for the Safety of Air Navigation is the most pragmatically oriented of the three measures. Preliminary measures of trust in ATC systems were constructed based on literature review and a model of the task. This resulted in a seven dimensional scale (reliability, accuracy, understanding, faith, liking, familiarity, and robustness). The measure was then refined in focus groups with air traffic controllers from different cultures rating two ATC simulations. Scale usability evaluations, and construct validity judgments were also collected. The instrument/items have reported reliabilities in the high 80's but its constructs have not been empirically validated.

All three scales have benefited from empirical study and systematic development yet each has its flaws. The ED instrument in [74], for instance, addresses trust in automation in the abstract without reference to an actual system and as a consequence appears to be more a measure of propensity to trust than trust in a specific system. A recent study [79] found scores on the ED instrument to be unaffected by reliability manipulations that produced significant changes in ratings of trust on other instruments. The HTC was developed from a model of trust and demonstrated agreement between items and target dimensions but stopped short of confirmatory factor analysis. Development of the SATI involved the most extensive pragmatic effort to adapt items so they made sense to users

and captured aspects of what users believed contributed to trust. However, SATI development neglected psychometric tests of construct and content validity.

A recent effort [80, 81] has led to a general measure of trust in automation validated across large populations in three diverse cultures, US, Taiwan and Turkey, as representative of Dignity, Face, and Honor cultures [82]. The Cross-cultural measure of trust is consistent with the three (performance, purpose, process) dimensions of [83, 25] and contains two 9 item scales, one measuring the propensity to trust as in [74] and the other measuring trust in a specific system. The second scale is designed to be administered repeatedly to measure the effects of manipulations expected to affect trust while the propensity scale is administered once at the start of an experiment. The scales have been developed and validated for US, Taiwanese, and Turkish samples and are based on 773 responses (propensity scale) and 1673 responses (specific scale).

The Trust Perception Scale-HRI [84, 79] is a psychometrically-developed 40 item instrument intended to measure human trust in robots. Items are based on data collected identifying robot features from pictures and their perceived functional characteristics. While development was guided by the triadic (human, robot, environment) model of trust inspired by the meta-analysis in [85], a factor analysis of the resulting scale found four components corresponding roughly to capability, behavior, task, and appearance. Capability and behavior correspond to two of the dimensions commonly found in interpersonal trust [83] and trust in automation [25], while appearance may have a special significance for trust in robots. The instrument was validated in same-trait and multi-trait analyses producing changes in rated trust associated with manipulation of robot reliability. The scale was developed based on 580 responses and 21 validation participants.

The HRI Trust Scale [86] was developed from items based on five dimensions (team configuration, team process, context, task, and system) identified by 11 subject matter experts (SMEs) as likely to affect trust. A 100 participant Mechanical Turk sample was used to select 37 items representing these dimensions. The HRI Trust Scale is incomplete as a sole measure of trust and is intended to be paired with Rotter's [50] interpersonal trust inventory when administered. While Lee & See's dimensions [25] other than "process" are missing from the HRI scale, they are represented in Rotter's instrument.

Because trust in automation or robots is an attitude, self-report through psychometric instruments such as these provides the most direct measurement. Questionnaires, however, suffer from a number of weaknesses. Because they are intrusive, measurements cannot be conveniently taken during the course of a task but only after the task is completed. This may suffice for automation such as ATR where targets are missed at a fixed rate and the experimenter is investigating the effect of that rate on trust [4], but it does not work in measuring moment to moment trust in a robot reading QR codes to get its directions [71].

6 Trust in Human Robot Interaction

Robots are envisioned to be able to process many complex inputs from the environment and be active participants in many aspects of life, including work environments, home assistance, battlefield and crisis response, and others. Therefore, robots are envisioned to transition from tool to teammate as humans transition from operator to teammate in an interaction more akin to human-human teamwork. These envisioned transitions raise a number of general questions: How would human interaction with the robot be affected? How would performance of the human-robot team be affected? How would human performance or behavior be affected? Although there are numerous tasks, environments, and situations of human-robot collaboration, in order to best clarify the role of trust we distinguish two general types of interactions of humans and robots: *performance-based interactions*, where the focus is on the *human influencing/controlling the robot* so it can perform useful tasks for the human, and *social-based interactions*, where the focus is on how the *robot's behavior influences human's beliefs and behavior*. In both these cases, the human is the trustor and the robot the trustee. In particular, in performance based interactions there is a particular task with a clear performance goal. An example of performance-based interactions is where human and robot collaborate in manufacturing assembly, or a UAV performing surveillance and recognition of victims in a search and rescue mission. Here measures of performance could be accuracy and timing to complete the task. On the other hand, in social interactions, the performance goal is not as crisply defined. An example of such a task is the ability of a robot to influence a human to reveal private knowledge, or how a robot can influence a human to take medicine or do useful exercises.

6.1 Performance-Based Interaction: Humans Influencing Robots

A large body of HRI research investigating factors thought to affect behavior via trust, such as reliability, rely strictly on behavioral measures without reference to trust. Meyer's [87] expected value (EV) theory of alarms provides one alternative by describing the human's choice as one between compliance (responding to an alarm) and reliance (not responding in the absence of an alarm). The expected values of these decisions are determined by the utilities associated with an uncorrected fault, the cost of intervention and the probabilities of misses (affecting reliance) and false alarms (affecting compliance). Research in [88], for example, investigated the effects of unmanned aerial vehicle (UAV) false alarms and misses on operator reliance inferred from longer reaction times for misses and compliance inferred from shorter reaction times to alarms. While reliance/compliance effects were not found, higher false alarm rates correlated with poorer performance on a monitoring task, while misses correlated with poorer performance on a parallel inspection task. A similar study by [89] of unmanned ground vehicle (UGV) control found participants with higher perceived attentional control were more adversely affected by false alarms (under-compliance) while those with low perceived attentional control were more strongly affected

by misses (over-reliance). Reliance and compliance can be measured in much the same way for homogeneous teams of robots as illustrated by a follow up study of teams of UGVs [90] of similar design and results. A similar study [91] involved multiple UAVs manipulating ATR reliability and administering a trust questionnaire, again finding that ratings of trust increased with reliability.

Transparency, common ground, or shared mental models involve a second construct ("process" [25] or "integrity" [26]) believed to affect trust. According to these models, the extent to which a human can understand the way in which an autonomous system works and predict its behavior will influence trust in the system. There is far less research on effects of transparency, with most involving level of automation manipulations. An early study [92] in which all conditions received full information found best performance for an intermediate level of automation that facilitated checks of accuracy (was transparent). Participants, however, made substantially greater use of a higher level of automation that provided an opaque recommendation. In this study, ratings of trust were affected by reliability but not transparency. More recent studies have equated transparency with additional information providing insight into robot behavior. Researchers in [93] compared conditions in which participants observed a simulated robot represented on a map by a status icon (level of transparency 1), overlaid with environmental information such as terrain (level 2), or with additional uncertainty and projection information (level 3). Note that these levels are distinct from Sheridan's Levels of Automation mentioned previously. What might appear as erratic behavior in level 1, for example, might be "explained"' by the terrain being navigated in level 2. Participant's ratings of trust were higher for levels 2 and 3. A second study manipulated transparency by comparing minimal (such as static image) contextual (such as video clip) and constant (such as video) information for a simulated robot team mate with which participants had intermittent interactions but found no significant differences in trust. In [94], researchers took a different approach to transparency by having a simulated robot provide "explanations" of its actions. The robot guided by a POMDP model can make different aspects of its decision making such as beliefs (probability of dangerous chemicals in building) or capabilities (ATR has 70% reliability) available to its human partner. Robot reliability affected both performance and trust. Explanations did not improve performance but did increase trust among those in the high reliability condition. As these studies suggest, reliability appears to have a large effect on trust, reliance/compliance, and performance, while transparency about function has a relatively minor one, primarily influencing trust. The third component of trust in robot's "purpose" [25] or "benevolence" [26] has been attributed [95–97] to "transparency" as conveyed by appearance discussed in 6.2. By this interpretation, matching human expectations aroused by a robot's appearance to its purpose and capabilities can make interactions more transparent by providing a more accurate model to the human.

Studies discussed to this point have treated trust as a dependent variable to be measured at the end of a trial and have investigated whether or not it had been affected by characteristics of the robot or situation. If trust of a robot

is modified through a process of interaction, however, it must be continuously varying as evidence accumulates of its trustworthiness or untrustworthiness. This was precisely the conception of trust investigated by Lee and Moray [19] in their seminal study but has been infrequently employed since. An recent example of such a study is reported in [98] where a series of experiments addressing temporal aspects of trust involving levels of automation and robot reliability have been conducted using a robot navigation and barriers task. In that task, a robot navigates through a course of boxes with labels that the operator can read through the robot's camera and QR codes presumed readable by the robot. The labels contain directions such as "turn right" or "U turn". In automation modes, robots follow a predetermined course with "failures" appearing to be misread QR codes. Operators can choose either the automation mode or a manual mode in which they determine the direction the robot takes. An initial experiment [98] investigated the effects of reliability drops at different intervals across a trial, finding that decline in trust as measured by post trial survey was greatest if the reliability decline occurred in the middle or final segments. In subsequent experiments, trust ratings were collected continuously by periodic button presses indicating increase or decrease in trust. These studies [99, 71] confirmed the primacy-recency bias in episodes of unreliability and the contribution of transparency in the form of confidence feedback from the robot.

Work in [100] collected similar periodic measures of trust using brief periodically presented questionnaires to participants performing a multi-UAV supervision task to test effects of priming on trust. These same data were used to fit a model similar to that formalized by [101] using decision field theory to address the decision to rely on the automation/robot's capabilities or to manually intervene based on the balance between the operator's self-confidence and her trust in the automation/robot. The model contains parameters characterizing information conveyed to operator, inertia in changing beliefs, noise, uncertainty, growth-decay rates for trust and self-confidence, and an inhibitory threshold for shifting between responses. By fitting these parameters to human subject data, the time course of trust (as defined by the model) can be inferred. An additional study of UAV control [102] has also demonstrated good fits for dynamic trust models with matches within 2.3% for control over teams of UGVs. By predicting effects of reliability and initial trust on system performance, such models might be used to select appropriate levels of automation or provide feedback to human operators. In another study involving assisted driving [103], the researchers use both objective (car position, velocity, acceleration, and lane marking scanners) and subjective (gaze detection and foot location) to train a mathematical model to recognize and diagnose over-reliance on the automation. The authors show that their models can be applied to other domains outside automation-assisted driving as well.

Willingness to rely on the automation has been found in the automation literature to correlate with user's self-confidence in their ability to perform the task [51]. It has been found that if a user is more confident in their own ability to perform the task, they will take control of the automation more frequently

if they perceive that the automation does not perform well. However, as robots are envisioned to be deployed in increasingly risky situations, it may be the case that a user (e.g. a soldier) may elect to use a robot for bomb disposal irrespective of his confidence in performing the task. Another factor that has considerably influenced use of automation is user workload. It has been found in the literature that users exhibit over-reliance [104, 105] on the automation in high workload conditions.

Experiments in [42] show that people over-trusted a robot in fire emergency evacuation scenarios conducted with a real robot in a campus building, although the robot was shown to be defective in various ways (e.g. taking a circuitous route rather then the efficient route in guiding the participant in a waiting room before the emergency started). It was hypothesized by the experimenters that the participants, having experienced an interaction with a defective robot, would decrease their trust (as opposed to a non-defective robot), and also that participants' self-reported trust would correlate with their behavior (i.e their decision to follow the robot or not). The results showed that, in general, participants did not rate the non-efficient robot as a bad guide, and even the ones that rated it poorly still followed it during the emergency. In other words, trust rating and trust behavior were not correlated. Interestingly enough, participants in a previous study with similar scenarios of emergency evacuation *in simulation* by the same researchers [106] behaved differently, namely participants rated less reliant simulated robots as less trustworthy and were less prone to follow them in the evacuation. The results from the simulation studies of emergency evacuation, namely positive correlation between participants' trust assessment and behavior, are similar to results in low risk studies [71]. These contradictory results point strongly that more research needs to be done to refine robot, operator and task-context variables and relations that would lead to correct trust calibration, and better understanding of the relationship between trust and performance in human robot interaction.

One important issue is how an agent forms trust on agents it has not encountered before. One approach from the literature in multiagent systems (MAS) investigates how trust forms in ad hoc groups, where agents that had not interacted before come together for short periods of time to interact and achieve a goal, after which they disband. In such scenarios, a decision tree model based on both trust and other factors (such as incentives and reputation) can be used [107]. A significant problem in such systems, known as the *cold start problem*, is that when such groups form there is little to no prior information on which to base trust assessments. In other words, how does an agent choose who to trust and interact with when they have no information on any agent? Recent work has focused on bootstrapping such trust assessments by using stereotypes [108]. Similar to stereotypes used in interpersonal interactions among humans, stereotypes in MAS are quick judgements based on easily observable features of the other agent. However, whereby human judgements are often clouded by cultural or societal biases, stereotypes in MAS can be constructed in a way that maximizes the accuracy. Further work by the researchers in [109] shows how stereotypes in

MAS can be spread throughout the group to improve others' trust assessments, and can be used by agents to detect unwanted biases received from others in the group. In [110], the authors show how this work can be used by organizations to create decision models based on trust assessments from stereotypes and other historical information about the other agents.

Towards Co-adaptive Trust In other studies [111, 112], Xu and Dudek create an online trust model to allow a robot or other automation to assess the operator's trust in the system while a mission is ongoing, using the results of the model to adjust the automation behavior on the fly to adapt to the estimated trust level. Their end goal is *trust-seeking adaptive robots*, which seek to actively monitor and adapt to the estimated trust of the user to allow for greater efficiency in human-robot interactions. Importantly, the authors combined common objective, yet indirect, measures of trust (such as quantity and type of user interaction), with a subjective measure in the form of periodical queries to the operator about their current degree of trust.

In an attempt to develop an objective and direct measure of trust the human has in the system, the authors of [113] use a mathematical decision model to estimate trust by determining the expected value of decisions a trusting operator would make, and then evaluate the user's decisions in relation to this model. In other words, if the operator deviates largely from the expected value of their decisions, they are said to be less trusting, and vice versa. In another study [114], the authors use two-way trust to adjust the relative contribution of the human input to that of the autonomous controller, as well as the haptic feedback provided to the human operator. They model both robot-to-human and human-to-robot trust, with lower values of the former triggering higher levels of force feedback, and lower values of the latter triggering a higher degree of human control over that of the autonomous robot controller. The authors demonstrate their model can significantly improve performance and lower the workload of operators when compared to previous models and manual control only.

These studies help introduce the idea of "inverse trust". The inverse trust problem is defined in [115] as determining how "an autonomous agent can modify it's behavior in an attempt to increase the trust a human operator will have in it". In this paper, the authors base this measure largely on the number of times the automation is interrupted by a human operator, and uses this to evaluate the autonomous agent's assessment of change in the operator's trust level. Instead of determining an absolute numerical value of trust, the authors choose to have the automation estimate *changes* in the human's trust level. This is followed in [116] by studies in simulation validating their inverse trust model.

6.2 Social-Based Interactions: Robots Influencing Humans

Social robotics deals with humans and robots interacting in ways humans typically interact with each other. In most of these studies, the robot—either by its appearance or its behavior—influences the human's beliefs about trustworthiness, feelings of companionship, comfort, feelings of connectedness with the

robot, or behavior (such as whether the human discloses secrets to the robot or follows the robot's recommendations). This is distinct from the prior work discussed, such as ATR, where a robot's actions are not typically meant to influence the feelings or behaviors of its operator. These social human-robot interactions contain affective elements that are closer to human-human interactions. There is a body of literature that looked at how robot characteristics affected ratings of animacy and other human-like characteristics, as well as trust in the robot, without explicitly naming a performance or social goal that the robot would perform. It has been consistently found in the social robotics literature that people tend to judge robot characteristics, such as reliability and intelligence, based on robot appearance. For example, people ascribe human qualities to robots that look more anthropomorphic. Another result of people's tendency to anthropomorphize robots is that they tend to ascribe animacy and intent to robots. This finding has not been reported just for robots [117] but even for simple moving shapes [118, 119]. Kiesler and Goetz [120] found that people rated more anthropomorphic looking robots as more reliable. Castro-Gonzalez et al. [121] investigated how the combination of movement characteristics with body appearance can influence people's attributions of animacy, liekeability, trustworthiness, and unpleasantness. They found that naturalistic motion was judged to be more animate, but only if the robot had a human appearance. Moreover, naturalistic motion improved ratings of likeability irrespective of the robot's appearance. More interestingly, a robot with human-like appearance was rated as more disturbing when its movements were more naturalistic. Participants also ascribe personality traits to robots based on appearance. For instance, in [122], robots with spider legs were rated as more aggressive whereas robots with arms rated as more intelligent than those without arms. Physical appearance is not the only attribute that influences human judgment about robot intelligence and knowledge. For example, [123] found that robots that spoke a particular language (e.g. Chinese) were rated higher in their purported knowledge of Chinese landmarks than robots that spoke English.

Robot appearance, physical presence [124], and matched speech [125] are likely to engender trust in the robot. [126] found that empathetic language and physical expression elicits higher trust. [127] found that highly expressive pedagogical interfaces engender more trust. A recent meta-analysis by Hancock et al. [85] found that robot characteristics such as reliability, behaviors and transparency influenced people's rating of trust in a robot. Besides these characteristics, the researchers in [85] also found that anthropomorphic qualities also had a strong influence on ratings of trust, and that trust in robots is influenced by experience with the robot.

Martelato et al. [128] found that if the robot is more expressive, this encourages participants to disclose information about themselves. However, counter to their hypotheses, disclosure of private information by the robot, a behavior that the authors labelled as making the robot more vulnerable, did not engender increased willingness to disclose on the part of the participants. In a study on willingness of children to disclose secrets, Bethel et al. [129] found in a qualita-

tive study that preschool children were found to be as likely to share a secret with an adult as with a humanoid robot.

An interesting study is reported in [41], where the authors studied how errors performed by the robot affect human trustworthiness and willingness of the human to subsequently comply with the robot's (somewhat unusual) requests. Participants interacted with a home companion robot, in the experimental room that was the pretend home of the robot's human owner in two conditions, (a) where the robot did not make mistakes and (b) where the robot made mistakes. The study found that the participants' assessment of robot reliability and trustworthiness was decreased significantly in the faulty robot condition; nevertheless, the participants were not substantially influence in their decisions to comply with the robot's unusual requests. It was further found that the nature of the request (revocable vs. irrevocable) influenced the participants' decisions on compliance. Interestingly, the results in this study also show that participants attributed less anthropomorphism when the robot made errors, which contradict those found by an earlier study the same authors had performed [130].

7 Conclusions and Recommendations

In this chapter we briefly reviewed the role of trust in human-robot interaction. We draw several conclusions, the first of which is that there is no accepted definition of what "trust" is in the context of trust in automation. Furthermore, when participants are asked to answer questions as to their level of trust in a robot or software automation, they are almost never given a definition of trust, leaving open the possibility that different participants are viewing the question of trust differently. From a review of the literature, it is apparent that robots still have not achieved full autonomy, and still lack the attributes that would allow them to be considered true teammates by their human counterparts. This is especially true because the literature is largely limited to simulation, or to specific, scripted interactions in the real world. Indeed, in [131], the authors argue that without human-like mental models and a sense of agency, robots will never be considered equal teammates within a mixed human-robot team. They argue that the reason researchers include robots in common HRI tasks is due to their ability to complement the skills of humans. Yet, because of the tendency of humans to anthropomorphize things they interact with, the controlled interactions researchers develop for HRI studies are more characteristic of human-human interactions. While this tendency to anthropomorphize can be helpful in some cases, it poses a serious risk if this naturally gives humans a higher degree of trust in robots than is warranted. The question of how a robot's performance influences anthropomorphization is also unclear—with recent studies finding conflicting results ([41] and [130]).

There is a general agreement that the notion of trust involves vulnerability of the trustor to the trustee in circumstances of risk and uncertainty. In the performance-based literature, where the human is relying on the robot to do the whole task or part of the task, it is clear that the participant is vulnerable

to the robot with respect to the participant's performance in the experimental task. In most of the studies in social robotics, however, where the robot is trying to get the participant to do something (e.g. comply with instructions to throw away someone else's mail, or disclose a secret) it is not clear that the participant is truly vulnerable to the robot (unless we regard breaking a social convention as making oneself vulnerable), merely enjoying the novelty of robots, or feeling pressure to follow experimental procedure. Therefore, the notion that was measured in those studies may not have been trust in the sense that the term is defined in the trust literature. For example in [42], where participants showed compliance with a robot guide even when reliability was ranked lower after an error, the researchers admit several confounding factors (e.g., participants did not have enough time to deliberate). The findings on human tendencies to ascribe reliability, trustworthiness, intelligence and other positive characteristics to robots may prohibit correct estimation of robot's abilities and prevent correct trust calibration. This is dangerous especially since the use of robots is envisioned to increase, especially in high risk situations such as emergency response and the military.

This overview enables us to provide several recommendations for how future work investigating trust in human-autonomy and human-robot interaction would proceed. First, it would be useful for the community to have a clear definition in each study as to what autonomy and what teammate characteristics the robot in the study possesses. Second, it would be useful for each study to define the notion of trust the author's espouse, as well as which dimensions of the notion of trust they believe are relevant to the task being investigated. The experimenters should also try to understand, via surveys or other means, what definition of trust the participants have in their heads. A possible idea is that experimenters could even give their definition of trust to the participants and see how this may affect the participants' answers.

Another recommendation is that, given the novelty of robots for the majority of the population, along with the well-known fact from in-group/out-group studies that people seem to be influenced very easily and for trivial reasons, it would be useful to perform longer duration studies to investigate the transient nature of trust assessments. In other words, how does trust in automation change as a function of how familiar users are with the automation and how much they interact with it over time? One could imaging someone unfamiliar with automation or robots placing a high degree of trust in them due to prior beliefs (which may be incorrect). Over time, this implicit trust may fade as they work more with automation and realize that it is not perfect.

Furthermore, we believe in a need to increase research in the multi-robot systems area, as well as the area of robots helping human teams. As the number of robots increase and hardware and operation costs decrease, it is inevitable that humans will be interacting with larger numbers of robots to perform increasingly complex tasks. Furthermore, trust in larger groups and collectives of robots is no doubt influenced by different factors—specifically those regarding the robots' behaviors—in addition to single robot control. Similarly, there is little work

investigating how multiple humans working together with robots affect each others' trust levels, which needs to be addressed.

Finally, it would be helpful for the community to define a set of task categories of human-robot interaction with characteristics that involve specific differing dimensions of trust. Such characteristics could be the degree of risk to the trustor, the degree of uncertainty, the degree of potential gain, whether the trustor's vulnerability is to the reliability of the robot, or the robot's integrity or benevolence. Other studies should expand on the notion of co-adaptive trust to improve how robots assess their own behavior and how it affects the trust in them by their operator. As communication is key to any collaborative interaction, research should not focus merely on how the human sees the robot, but also how the robot sees the human.

8 Acknowledgments

This work is supported by awards FA9550-13-1-0129 and FA9550-15-1-0442.

References

1. R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 39, no. 2, pp. 230–253, 1997.
2. J. B. Lyons and C. K. Stokes, "Human–human reliance in the context of automation," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, p. 0018720811427034, 2011.
3. K. L. Mosier and L. J. Skitka, "Human decision makers and automated decision aids: Made for each other?" *Automation and human performance: Theory and applications*, pp. 201–220, 1996.
4. M. T. Dzindolet, L. G. Pierce, H. P. Beck, and L. A. Dawe, "The perceived utility of human and automated aids in a visual detection task," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 44, no. 1, pp. 79–94, 2002.
5. C. Layton, P. J. Smith, and C. E. McCoy, "Design of a cooperative problem-solving system for en-route flight planning: An empirical evaluation," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 36, no. 1, pp. 94–119, 1994.
6. J. M. McGuirl and N. B. Sarter, "Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 48, no. 4, pp. 656–665, 2006.
7. N. B. Sarter and B. Schroeder, "Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 43, no. 4, pp. 573–583, 2001.
8. K. L. Mosier, E. A. Palmer, and A. Degani, "Electronic checklists: Implications for decision making," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 36, no. 1. SAGE Publications, 1992, pp. 7–11.

9. K. L. Mosier, L. J. Skitka, S. Heers, and M. Burdick, "Automation bias: Decision making and performance in high-tech cockpits," *The International journal of aviation psychology*, vol. 8, no. 1, pp. 47–63, 1998.
10. E. Alberdi, A. Povyakalo, L. Strigini, and P. Ayton, "Effects of incorrect computer-aided detection (cad) output on human decision-making in mammography," *Academic radiology*, vol. 11, no. 8, pp. 909–918, 2004.
11. K. A. McKibbon and D. B. Fridsma, "Effectiveness of clinician-selected electronic information resources for answering primary care physicians' information needs," *Journal of the American Medical Informatics Association*, vol. 13, no. 6, pp. 653–659, 2006.
12. R. P. Will, "True and false dependence on technology: Evaluation with an expert system," *Computers in human behavior*, vol. 7, no. 3, pp. 171–183, 1991.
13. K. E. Weick, "Enacted sensemaking in crisis situations," *Journal of management studies*, vol. 25, no. 4, pp. 305–317, 1988.
14. K. Mosier, L. Skitka, and K. Korte, "Cognitive and social psychological issues in flight crew/automation interaction," *Human performance in automated systems: Current research and trends*, pp. 191–197, 1994.
15. I. L. Singh, R. Molloy, and R. Parasuraman, "Individual differences in monitoring failures of automation," *The Journal of General Psychology*, vol. 120, no. 3, pp. 357–373, 1993.
16. P. M. Satchell, *Cockpit Monitoring and Alerting Systems*. Routledge, 1993.
17. B. M. Muir and N. Moray, "Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.
18. S. Lewandowsky, M. Mundy, and G. Tan, "The dynamics of trust: comparing humans to automation." *Journal of Experimental Psychology: Applied*, vol. 6, no. 2, p. 104, 2000.
19. J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, 1992.
20. B. M. Muir, "Operators' trust in and use of automatic controllers in a supervisory process control task," Ph.D. dissertation, University of Toronto, 1990.
21. C. Billings, J. Lauber, H. Funkhouser, E. Lyman, and E. Huff, "Nasa aviation safety reporting system," NASA Ames Research Center, Tech. Rep. Technical Report TM-X-3445, 1976.
22. J. Llinas, A. Bisantz, C. Drury, Y. Seong, and J.-Y. Jian, "Studies and analyses of aided adversarial decision making. phase 2: Research on human trust in automation," DTIC Document, Tech. Rep., 1998.
23. R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 52, no. 3, pp. 381–410, 2010.
24. C. Kelly, M. Boardman, P. Goillau, and E. Jeannot, "Guidelines for trust in future atm systems: A literature review," *European Organization for the Safety of Air Navigation*, 2003.
25. J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, 2004.
26. R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Academy of management review*, vol. 20, no. 3, pp. 709–734, 1995.

27. N. Moray, T. Inagaki, and M. Itoh, "Adaptive automation, trust, and self-confidence in fault management of time-critical tasks." *Journal of Experimental Psychology: Applied*, vol. 6, no. 1, p. 44, 2000.
28. M. Madsen and S. Gregor, "Measuring human-computer trust," in *11th australasian conference on information systems*, vol. 53. Citeseer, 2000, pp. 6–8.
29. B. D. Adams, D. J. Bryant, and R. Webb, "Trust in teams: Literature review," Report to Defense and Civil Insitute of Environmental Medicine. Humansystems Inc., Tech. Rep. Technical Report CR-2001-042, 2001.
30. G. R. Jones and J. M. George, "The experience and evolution of trust: Implications for cooperation and teamwork," *Academy of management review*, vol. 23, no. 3, pp. 531–546, 1998.
31. J. D. Lewis and A. Weigert, "Trust as a social reality," *Social forces*, vol. 63, no. 4, pp. 967–985, 1985.
32. B. M. Muir, "Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994.
33. N. Moray and T. Inagaki, "Laboratory studies of trust between humans and machines in automated systems," *Transactions of the Institute of Measurement and Control*, vol. 21, no. 4-5, pp. 203–211, 1999.
34. B. Barber, *The logic and limits of trust*. Rutgers University Press, 1983.
35. J. K. Rempel, J. G. Holmes, and M. P. Zanna, "Trust in close relationships." *Journal of personality and social psychology*, vol. 49, no. 1, p. 95, 1985.
36. R. E. Miles and W. D. Creed, "Organizational forms and managerial philosophies-a descriptive and analytical review," *Research in Organizational Behavior: An Annual Series of Anaylytical Essays and Critical Reviews*, vol. 17, pp. 333–372, 1995.
37. D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not so different after all: A cross-discipline view of trust," *Academy of management review*, vol. 23, no. 3, pp. 393–404, 1998.
38. P. H. Kim, K. T. Dirks, and C. D. Cooper, "The repair of trust: A dynamic bilateral perspective and multilevel conceptualization," *Academy of Management Review*, vol. 34, no. 3, pp. 401–422, 2009.
39. A. Fulmer and G. M, *Models for intercultural collaboration and negotiation*. Springer, 2012, ch. Dynamic trust processes: trust dissolution, recovery and stabilization.
40. F. J. Lerch, M. J. Prietula, and C. T. Kulik, "The turing effect: The nature of trust in expert systems advice," in *Expertise in context*. MIT Press, 1997, pp. 417–448.
41. M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "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, 2015, pp. 141–148.
42. P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 101–108.
43. V. A. Riley, "Human use of automation," Ph.D. dissertation, University of Minneapolis, 1994.
44. R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive

engineering constructs," *Journal of Cognitive Engineering and Decision Making*, vol. 2, no. 2, pp. 140–160, 2008.
45. K. Sycara and M. Lewis, "Forming shared mental models," in *Proc. of the 13th Annual Meeting of the Cognitive Science Society*, 1991, pp. 400–405.
46. A. Simpson, G. Brander, and D. Portsdown, "Seaworthy trust: Confidence in automated data fusion," *The Human-Electronic Crew: Can we Trust the Team*, pp. 77–81, 1995.
47. M. Lewis, "Designing for human-agent interaction," *AI Magazine*, vol. 19, no. 2, p. 67, 1998.
48. K. P. Sycara, M. Lewis, T. Lenox, and L. Roberts, "Calibrating trust to integrate intelligent agents into human teams," in *System Sciences, 1998., Proceedings of the Thirty-First Hawaii International Conference on*, vol. 1. IEEE, 1998, pp. 263–268.
49. T. Sheridan and W. Verplank, "Human and computer control of undersea teleoperators. cambridge, ma: Man-machine systems laboratory, department of mechanical engineering," 1978.
50. J. B. Rotter, "A new scale for the measurement of interpersonal trust1," *Journal of personality*, vol. 35, no. 4, pp. 651–665, 1967.
51. J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *International journal of human-computer studies*, vol. 40, no. 1, pp. 153–184, 1994.
52. A. J. Masalonis and R. Parasuraman, "Effects of situation-specific reliability on trust and usage of automated air traffic control decision aids," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 47, no. 3. SAGE Publications, 2003, pp. 533–537.
53. S. M. Merritt and D. R. Ilgen, "Not all trust is created equal: Dispositional and history-based trust in human-automation interactions," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 50, no. 2, pp. 194–210, 2008.
54. K. Karvonen, L. Cardholm, and S. Karlsson, "Designing trust for a universal audience: a multicultural study on the formation of trust in the internet in the nordic countries." in *HCI*, 2001, pp. 1078–1082.
55. J. B. Rotter, "Generalized expectancies for interpersonal trust." *American psychologist*, vol. 26, no. 5, p. 443, 1971.
56. J. Cassidy, "Child-mother attachment and the self in six-year-olds," *Child development*, vol. 59, no. 1, pp. 121–134, 1988.
57. A. W. Kruglanski, E. P. Thompson, E. T. Higgins, M. Atash, A. Pierro, J. Y. Shah, and S. Spiegel, "To" do the right thing" or to" just do it": locomotion and assessment as distinct self-regulatory imperatives." *Journal of personality and social psychology*, vol. 79, no. 5, p. 793, 2000.
58. M. B. Brewer and R. M. Kramer, "The psychology of intergroup attitudes and behavior," *Annual review of psychology*, vol. 36, no. 1, pp. 219–243, 1985.
59. I. Bohnet, B. Hermann, and R. Zeckhauser, "The requirements for trust in gulf and western countries." *Quarterly Journal of Economics*, vol. 125, pp. 811–828, 2010.
60. P. W. Dorfman, P. J. Hanges, and F. C. Brodbeck, "Leadership and cultural variation: The identification of culturally endorsed leadership profiles," *Culture, leadership, and organizations: The GLOBE study of*, vol. 62, pp. 669–719, 2004.
61. D. Meyerson, K. E. Weick, and R. M. Kramer, "Swift trust and temporary groups," *Trust in organizations: Frontiers of theory and research*, vol. 166, p. 195, 1996.

62. C. K. De Dreu and P. J. Carnevale, "Motivational bases of information processing and strategy in conflict and negotiation," *Advances in experimental social psychology*, vol. 35, pp. 235–291, 2003.
63. D. Carl, V. Gupta, and M. Javidan, "Culture, leadership, and organizations: The globe study of 62 societies," 2004.
64. J. Brockner, T. R. Tyler, and R. Cooper-Schneider, "The influence of prior commitment to an institution on reactions to perceived unfairness: The higher they are, the harder they fall," *Administrative Science Quarterly*, pp. 241–261, 1992.
65. A. Merritt, "Culture in the cockpit do hofstede's dimensions replicate?" *Journal of cross-cultural psychology*, vol. 31, no. 3, pp. 283–301, 2000.
66. G. Hofstede, G. J. Hofstede, and M. Minkov, *Cultures and organizations: Software of the mind*. Citeseer, 1991, vol. 2.
67. P. P. Rau, Y. Li, and D. Li, "Effects of communication style and culture on ability to accept recommendations from robots," *Computers in Human Behavior*, vol. 25, no. 2, pp. 587–595, 2009.
68. L. Wang, P.-L. P. Rau, V. Evers, B. K. Robinson, and P. Hinds, "When in rome: the role of culture & context in adherence to robot recommendations," in *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2010, pp. 359–366.
69. S.-Y. Chien, M. Lewis, K. Sycara, J.-S. Liu, and A. Kumru, "Influence of cultural factors in dynamic trust in automation," in *Proceedings of the Systems, Man, and Cybernetics Society*, 2016.
70. V. A. Riley, *Automation theory and applications*. Mahwah, NJ. Erlbaum, 1996, ch. Operator reliance on automation: theory and data, pp. 19–35.
71. M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2013, pp. 251–258.
72. S. Dekker and E. Hollnagel, "Human factors and folk models," *Cognition, Technology & Work*, vol. 6, no. 2, pp. 79–86, 2004.
73. S. W. Dekker and D. D. Woods, "Maba-maba or abracadabra? progress on human–automation co-ordination," *Cognition, Technology & Work*, vol. 4, no. 4, pp. 240–244, 2002.
74. J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.
75. R. D. Spain, E. A. Bustamante, and J. P. Bliss, "Towards an empirically developed scale for system trust: Take two," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 52, no. 19. SAGE Publications, 2008, pp. 1335–1339.
76. R. Master, X. Jiang, M. T. Khasawneh, S. R. Bowling, L. Grimes, A. K. Gramopadhye, and B. J. Melloy, "Measurement of trust over time in hybrid inspection systems," *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol. 15, no. 2, pp. 177–196, 2005.
77. M. Luz, "Validation of a trust survey on example of mtcd in real time simulation with irish controllers," Ph.D. dissertation, thesis final report. The European Organisation for the Safety of Air Navigation, 2009.
78. P. Goillau, C. Kelly, M. Boardman, and E. Jeannot, "Guidelines for trust in future atm systems-measures," *EUROCONTROL, the European Organization for the Safety of Air Navigation*, 2003.

79. K. E. Schaefer, J. Y. Chen, J. L. Szalma, and P. Hancock, "A meta-analysis of factors influencing the development of trust in automation implications for understanding autonomy in future systems," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, p. 0018720816634228, 2016.
80. S.-Y. Chien, Z. Semnani-Azad, M. Lewis, and K. Sycara, "Towards the development of an inter-cultural scale to measure trust in automation," in *International Conference on Cross-Cultural Design*. Springer, 2014, pp. 35–46.
81. S.-Y. Chien, M. Lewis, S. Hergeth, Z. Semnani-Azad, and K. Sycara, "Cross-country validation of a cultural scale in measuring trust in automation," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 59, no. 1. SAGE Publications, 2015, pp. 686–690.
82. A. K.-Y. Leung and D. Cohen, "Within-and between-culture variation: individual differences and the cultural logics of honor, face, and dignity cultures." *Journal of personality and social psychology*, vol. 100, no. 3, p. 507, 2011.
83. J. Meyer, "Effects of warning validity and proximity on responses to warnings," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 43, no. 4, pp. 563–572, 2001.
84. K. E. Schaefer, "The perception and measurement of human-robot trust," Ph.D. dissertation, University of Central Florida Orlando, Florida, 2013.
85. P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, 2011.
86. R. E. Yagoda and D. J. Gillan, "You want me to trust a robot? the development of a human–robot interaction trust scale," *International Journal of Social Robotics*, vol. 4, no. 3, pp. 235–248, 2012.
87. J. Meyer, "Conceptual issues in the study of dynamic hazard warnings," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 2, pp. 196–204, 2004.
88. S. R. Dixon and C. D. Wickens, "Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 48, no. 3, pp. 474–486, 2006.
89. J. Chen and P. Terrence, "Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment," *Ergonomics*, vol. 52, no. 8, pp. 907–920, 2009.
90. J. Y. Chen, M. J. Barnes, and M. Harper-Sciarini, "Supervisory control of multiple robots: Human-performance issues and user-interface design," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 4, pp. 435–454, 2011.
91. E. de Visser and R. Parasuraman, "Adaptive aiding of human-robot teaming effects of imperfect automation on performance, trust, and workload," *Journal of Cognitive Engineering and Decision Making*, vol. 5, no. 2, pp. 209–231, 2011.
92. T. Lenox, M. Lewis, E. Roth, R. Shern, L. Roberts, T. Rafalski, and J. Jacobson, "Support of teamwork in human-agent teams," in *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on*, vol. 2. IEEE, 1998, pp. 1341–1346.
93. M. W. Boyce, J. Y. Chen, A. R. Selkowitz, and S. G. Lakhmani, "Effects of agent transparency on operator trust," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, ser.

HRI'15 Extended Abstracts. New York, NY, USA: ACM, 2015, pp. 179–180. [Online]. Available: http://doi.acm.org/10.1145/2701973.2702059
94. N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, ser. HRI '16. Piscataway, NJ, USA: IEEE Press, 2016, pp. 109–116. [Online]. Available: http://dl.acm.org/citation.cfm?id=2906831.2906852
95. J. B. Lyons and P. R. Havig, "Transparency in a human-machine context: Approaches for fostering shared awareness/intent," in *International Conference on Virtual, Augmented and Mixed Reality*. Springer, 2014, pp. 181–190.
96. S. Ososky, D. Schuster, E. Phillips, and F. Jentsch, "Building appropriate trust in human-robot teams," in *AAAI Spring Symposium Series*, 2013.
97. J. B. Lyons, "Being transparent about transparency," in *AAAI Spring Symposium*, 2013.
98. M. Desai, "Modeling trust to improve human-robot interaction," Ph.D. dissertation, University of Massachusetts Lowell, 2012.
99. P. Kaniarasu, A. Steinfeld, M. Desai, and H. Yanco, "Robot confidence and trust alignment," in *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2013, pp. 155–156.
100. A. S. Clare, M. L. Cummings, and N. P. Repenning, "Influencing trust for human–automation collaborative scheduling of multiple unmanned vehicles," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 57, no. 7, pp. 1208–1218, 2015.
101. J. Gao and J. D. Lee, "Extending the decision field theory to model operators' reliance on automation in supervisory control situations," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 36, no. 5, pp. 943–959, 2006.
102. F. Gao, A. S. Clare, J. C. Macbeth, and M. Cummings, "Modeling the impact of operator trust on performance in multiple robot control," in *Spring Symposium AAAI*, 2013.
103. K. Takeda, "Modeling and detecting excessive trust from behavior signals: Overview of research project and results," in *Human-Harmonized Information Technology, Volume 1*. Springer, 2016, pp. 57–75.
104. D. P. Biros, M. Daly, and G. Gunsch, "The influence of task load and automation trust on deception detection," *Group Decision and Negotiation*, vol. 13, no. 2, pp. 173–189, 2004.
105. K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: a systematic review of frequency, effect mediators, and mitigators," *Journal of the American Medical Informatics Association*, vol. 19, no. 1, pp. 121–127, 2012.
106. P. Robinette, A. M. Howard, and A. R. Wagner, "Timing is key for robot trust repair," in *International Conference on Social Robotics*. Springer, 2015, pp. 574–583.
107. C. Burnett, T. J. Norman, and K. Sycara, "Decision-making with trust and control in multi-agent systems," *Twenty Second International Joint Conference on Artificial Intelligence*, vol. 10, pp. 241–248, 2011.
108. ——, "Bootstrapping trust evaluations through stereotypes," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 2010, pp. 241–248.
109. ——, "Stereotypical trust and bias in dynamic multiagent systems," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 4, no. 2, p. 26, 2013.

110. C. Burnett, T. J. Norman, K. Sycara, and N. Oren, "Supporting trust assessment and decision making in coalitions," *IEEE Intelligent Systems*, vol. 29, no. 4, pp. 18–24, 2014.
111. A. Xu and G. Dudek, "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, 2015, pp. 221–228.
112. ——, "Maintaining efficient collaboration with trust-seeking robots," in *Intelligent Robots and Systems, 2016.(IROS 2016). Proceedings. 2016 IEEE/RSJ International Conference on*, vol. 16. IEEE, 2016.
113. A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman, "Measurement of trust in human-robot collaboration," in *Collaborative Technologies and Systems, 2007. CTS 2007. International Symposium on*. IEEE, 2007, pp. 106–114.
114. H. Saeidi, F. McLane, B. Sadrfaidpour, E. Sand, S. Fu, J. Rodriguez, J. Wagner, and Y. Wang, "Trust-based mixed-initiative teleoperation of mobile robots," in *2016 American Control Conference (ACC)*. IEEE, 2016, pp. 6177–6182.
115. M. W. Floyd, M. Drinkwater, and D. W. Aha, "Adapting autonomous behavior using an inverse trust estimation," in *International Conference on Computational Science and Its Applications*. Springer, 2014, pp. 728–742.
116. ——, "Learning trustworthy behaviors using an inverse trust metric," in *Robust Intelligence and Trust in Autonomous Systems*. Springer, 2016, pp. 33–53.
117. M. Saerbeck and C. Bartneck, "Perception of affect elicited by robot motion," in *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2010, pp. 53–60.
118. F. Heider and M. Simmel, "An experimental study of apparent behavior," *The American Journal of Psychology*, vol. 57, no. 2, pp. 243–259, 1944.
119. W. Ju and L. Takayama, "Approachability: How people interpret automatic door movement as gesture," *International Journal of Design*, vol. 3, no. 2, 2009.
120. S. Kiesler and J. Goetz, "Mental models of robotic assistants," in *CHI'02 extended abstracts on Human Factors in Computing Systems*. ACM, 2002, pp. 576–577.
121. Á. Castro-González, H. Admoni, and B. Scassellati, "Effects of form and motion on judgments of social robots' animacy, likability, trustworthiness and unpleasantness," *International Journal of Human-Computer Studies*, vol. 90, pp. 27–38, 2016.
122. V. K. Sims, M. G. Chin, D. J. Sushil, D. J. Barber, T. Ballion, B. R. Clark, K. A. Garfield, M. J. Dolezal, R. Shumaker, and N. Finkelstein, "Anthropomorphism of robotic forms: a response to affordances?" in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 49, no. 3. SAGE Publications, 2005, pp. 602–605.
123. S.-l. Lee, I. Y.-m. Lau, S. Kiesler, and C.-Y. Chiu, "Human mental models of humanoid robots," in *Proceedings of the 2005 IEEE international conference on robotics and automation*. IEEE, 2005, pp. 2767–2772.
124. W. A. Bainbridge, J. Hart, E. S. Kim, and B. Scassellati, "The effect of presence on human-robot interaction," in *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2008, pp. 701–706.
125. C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction." *Journal of Experimental Psychology: Applied*, vol. 7, no. 3, p. 171, 2001.

126. A. Tapus, M. J. Mataric, and B. Scassellati, "Socially assistive robotics [grand challenges of robotics]," *IEEE Robotics & Automation Magazine*, vol. 14, no. 1, pp. 35–42, 2007.
127. J. C. Lester, S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone, and R. S. Bhogal, "The persona effect: affective impact of animated pedagogical agents," in *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. ACM, 1997, pp. 359–366.
128. N. Martelaro, V. C. Nneji, W. Ju, and P. Hinds, "Tell me more: Designing hri to encourage more trust, disclosure, and companionship," in *The Eleventh ACM/IEEE International Conference on Human Robot Interation*. IEEE Press, 2016, pp. 181–188.
129. C. L. Bethel, M. R. Stevenson, and B. Scassellati, "Secret-sharing: Interactions between a child, robot, and adult," in *Systems, man, and cybernetics (SMC), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2489–2494.
130. M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 313–323, 2013.
131. V. Groom and C. Nass, "Can robots be teammates?: Benchmarks in human–robot teams," *Interaction Studies*, vol. 8, no. 3, pp. 483–500, 2007.