Trust: Recent Concepts and Evaluations in Human-Robot Interaction

Theresa Law and Matthias Scheutz

December 12, 2019

Abstract

We present a survey of investigations of human trust in robots in the recent human-robot interaction literature. The included papers are all experimental HRI studies and were published in the years 2018 or 2019. We explore how trust is defined in these papers, as well as what types of questions about trust are investigated and how trust is being objectively measured. We divide trust into two categories, performance-based trust and relation-based trust, and examine how the definitions, questions, and measures fall into those two categories. We also examine how these two categories of trust line up across definition, question, and measure for a given paper. We found a conflation between performance-based trust and relation-based trust, and that while there is an interest in asking relation-based trust questions, there is a lack of objective means of measuring that trust. We offer recommendations for the formalization of trust, objective experimental paradigms, and trust-related questions to investigate.

1 Introduction

As robots become more sophisticated and interact with people more frequently, especially people who have little to no robotics experience, it is increasingly important for us to understand what it takes for people to trust or distrust a robot. This understanding begins by examining what we currently know and understand trust to be, and how we can reasonably measure trust, both in the situations where we believe we are can measure it objectively and also those cases where objective measures might not be possible and we have to rely on people's subjective feedback.

While trust as a subject has long been studied in psychology (Dunning & Fetchenhauer, 2011) and human factors (Lee & See, 2004), the human-robot interaction (HRI) community has only recently begun to explore trust in robots in earnest. The growing interest in trust in HRI is evident in the increasing number of publications on the topic, even though there is currently no agreed-upon definition of trust. As trust is a multi-faceted concept (Yamagishi, 1998), and can be affected by a large number of factors (Hancock, Billings, Schaefer, Chen, De Visser, & Parasuraman, 2011), studying trust often means to study one particular aspect of trust, rather than an all-encompassing "trust" concept. Unsurprisingly, there is no accepted paradigm for evaluating trust in HRI either.

The majority of trust studies in HRI rely on different types of questionnaires in order to determine a person's trust in the robot with whom they are interacting. Some of the questionnaires are validated (Schaefer, 2013; Merritt, LaChapell, & Lee, 2012; Mayer, Davis, & Schoorman, 1995; Jian, Bisantz, & Drury, 200; Ullman & Malle, 2018, Larzelere & Huston, 1980), while others are not. These latter studies are more ad hoc. And for the few studies that do have objective trust measures, the lack of a unifying experimental paradigm makes it difficult to compare results across studies.

The purpose of this survey is to investigate the different trust dimensions that have been used in recent HRI studies, to classify those dimensions, and to show what instruments can be used to measure them. The hope is that the resulting framework and analysis will be useful for future investigations of trust in HRI.

The chapter is outlined as follows. First, we describe our methodology for finding papers to include in our survey. Then, we present an overview of how trust has been defined in HRI in recent years. These definitions led us to dividing trust into two categories- performance-based trust and relation-based trust. We then begin to examine the papers relevant to this survey, starting with the questions the papers asked, then how trust was measured. Within these sections, we separated the papers that looked at performance-based trust from relation-based trust, as well as from the studies that looked at a combination of these types of trust. By making this separation, we highlight what these different types of trust encompass and the discrepancy in how they are measured. We then compare what types of questions are being asked and how trust is measured and defined within each study, showing the conflation and inconsistency of studying performance- and relation-based trust. We end with a discussion and recommendations for how the study of trust in HRI should

advance, and a conclusion of our findings.

2 Methodology

The papers included in this study were experimental papers studying trust in HRI. Because we wanted to look at the current state of the field, only papers published in 2018 and up through the summer of 2019 were included. Search engines used were Scopus and Google Scholar, and papers were found using the search terms "trust" and "human-robot interaction," as well as from references from other papers. This resulted in a total of 33 papers being included in the survey. Table 1 presents an overview of all of the papers.

3 Trust Definitions in HRI

The most widely cited trust definitions in recent HRI studies come from Lee & See (2004) and Hancock et al. (2011a; 2011b). Lee and See define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability" (p. 51), while Hancock defines it as "the reliance by an agent that actions prejudicial to their well-being will not be undertaken by influential others" (2011a, p. 24) and as a willingness to accept suggestions from another agent (Hancock et al., 2011b). The concepts covered in these definitions are highly relevant to HRI. For example, if persons who have never worked with or programmed a robot before coming in contact with one, they will likely experience a high level of uncertainty about how the interaction will unfold. Even those familiar with robots in general or with specific robots may be uncertain about the functionality or behavior of a new robot. This uncertainty can leave people vulnerable; for example, they may need to rely on the robot for a task and are uncertain about its performance capabilities in achieving their goal. Alternatively, they may be asked to provide the robot with personal information, and are uncertain about what is going to be done with that information or who will find out about it. Both of these scenarios leave people vulnerable to the behavior of robots. Additionally, interacting with a large and heavy robot may cause a person to be physically vulnerable. Therefore, people choosing to work with robots despite these uncertainties display a certain level of trust in the robot. If trust is present, people may be willing to alter their own behavior based on advice or information provided by the robot. For robots who work directly

and closely with people, this can be an important aspect of a trusting relationship.

For robots that do have roles in which they work with people, especially in a social manner, being emotionally supportive can be another important aspect of trust(cp. to Rotter 1967). Other trust definitions reference the trustee mitigating risk for the truster (Robinette, Howard, & Wagner, 2017; Wagner & Arkin, 2011), an agent's performance being reliable (Lewis, Sycara, & Walker, 2018; Castelfranci & Falcone, 2010; Hodges & Geyer, 2006), and the potential lack of the ability of the truster to monitor the trustee's actions (Mayer et al, 1995). In returning to the above examples, we can see how these are all relevant to HRI. If a human feels vulnerable in their interaction with a robot, perceiving the robot as mitigating the human's risk should help the human to trust it more. Or if the person needs the robot to perform a task properly, the more reliable the robot is at that task, the more likely the person will trust it to continue to perform that task well. Ideally, the robot will be trusted to behave without direct or constant human supervision, allowing humans to focus on their own tasks and trust that the robot can do its own satisfactorily.

We attempted to structure the discussion of trust in HRI based on several important aspects of trust that are experimentally investigated. From the above definitions, we divide trust into two categories for the remainder of this paper: *performance-based trust* and *relation-based trust*.

Performance-based trust centers around the robot being trusted to be reliable, capable, and competent at its task or tasks, without needing to be monitored by a human supervisor. Performance-based trust may also depend on the robot's transparency, responsiveness, and predictability. Relation-based trust, on the other hand, implies that a robot is trusted as a social agent. A person with whom it interacts can be vulnerable emotionally, and may trust that the robot will be sincere and ethical. Relation-based trust means that a person trusts the robot to be part of society in some way, not just off in a factory doing a job without any expectation of knowledge of social norms. This type of trust is becoming increasingly relevant, and therefore it is increasingly pressing to understand what factors may influence it.

These two categories reflect two very important, but very different, aspects of trust. While some situations may require people to trust a robot in both of these manners, many other situations only necessitate one or the other, as will be highlighted in the remainder of the chapter. The presence or absence of these two types of trust will affect robot use in different ways. A robot that garnered a human's trust in a performance-based manner may be tasked with a critical factory job, but

may be sequestered to a room where it will never interact with people. A robot that is trusted in a relation-based manner, conversely, may be brought into a nursing home to be a companion robot, but would not be given a task like sorting a resident's pills. However, as will be discussed below, certain situations such as choosing a robotic teammate for a competitive game, require both types of trust. Because these types of trust affect different types of interactions, it is important for researchers to be aware of which type of trust they are studying based on their definition of trust and their relevant questions, and if their measure of trust aligns with what they are studying. The following sections examine these latter factors- questions that researchers are currently asking about trust, and how they are measuring it.

4 Trust Questions

There is a large number of factors which may have an impact on one agent's trust in another, and those factors could vary depending upon the context and the specific individual or interaction. To understand how these factors play out in establishing and maintaining trust in HRI, we must systematically isolate specific types of interactions, behaviors, and performances to understand their effects on trust. In the following sections we examine the questions researchers were interested in by looking at what they manipulated or used as an independent variable, as these questions are indicative of the aspect of trust the researchers found most important. We start with questions that focus on performance-based or relation-based trust, followed by questions that address both. By dividing the chapter as such, we can categorize papers by the type of trust that the researchers were interested in studying; we can later use this to compare questions to the category of trust that their dependent measure reflected.

In this section, we step through each paper that fit our experimental criteria and categorize it as asking performance-based questions, relation-based questions, or a mix of the two. We then describe the actual questions that each paper addressed, and a summary of their results based on how they measured trust.

4.1 Experimental Manipulations

The following studies are ones in which the researchers manipulated experimental factors to see how different conditions affected trust.

4.1.1 Performance-Based Questions

Performance-based trust questions vary how well the robot performs in different conditions, or how aware the participants are of the robot's ability to perform different tasks. These types of questions do not rely on how the robot interacts socially, but how capable and reliable it is at its task. In interacting with a robot, a person may not be aware of the full extent, or limitations, of a robot's capabilities. Shu et al. (2018) and Soh et al. (2018) informed participants of the robot's capabilities by showing it performing certain tasks, and asked whether their trust in it, indicated by Muir's (1994) questionnaire generalized to other tasks, given what they knew about its capabilities. Both studies found that trust was more directly transferred to unseen tasks that were similar to the observed task, that participants transferred trust more readily to simple tasks rather than difficult tasks, and that the robot's performance affected trust both in the observed and unobserved task.

Rather than manipulate participants' perceptions of robots' capabilities, the following studies manipulated the robot's actual task performance. Xu and Howard (2018) and Geiskkovitch et al. (2019) looked at how the objective correctness of a robot's answer affected trust. The former asked how first impressions, when a robot provides an obviously correct or incorrect answer to a question during a first interaction, impacted trust; the latter asked about how robot errors were perceived by children. Xu and Howard found that a robot who gave a first impression of being faulty by providing an incorrect answer to a question was trusted less than a robot that gave a first impression of working properly by providing the correct answer; participants both accepted a robot's suggestions more often and rated their trust of it higher than that of the faulty robot on a survey question that asked if they trusted its advice. Geiskkovitch et al. found that children age 3-5 trusted a robot who had been previously correct more than one that had been incorrect, as indicated by siding with the correcting one in a binary decision.

Rossi et al. (2018a) asked how severe different robot errors were considered to be, and based on those reports, how much trust was affected by errors of varying severity. When presented with

a final emergency scenario and asked how they wanted to deal with it, participants who had seen the robot cause more severe errors were more likely to indicate that they did not trust the robot to handle the emergency, but were more willing to work as a team if the robot had made small errors. Byrne and Marín (2018) manipulated performance by asking how different levels of task completion in service robots, from failure to success through teamwork, affected trust. They asked participants before and after the experiment about their views on propensity to trust, sociability, competency, team working ability, and responsiveness to measure trust. They found that failure made trust decrease between the surveys, and success through teamwork was the greatest cause of increased trust between the pre and post surveys. Chen et al. (2018) had participants work with a robot to clear off a table, and the robot either did or did not take the person's trust level into account as it performed the task. They found that trust, as indicated by how infrequently a participant intervened in the robot clearing objects off the table, was higher with the robot who took the participant's evolving trust level into account.

Jayaraman et al. (2018) looked at the interaction between autonomous vehicle (AV) behavior and pedestrians by manipulating the performance of an AV in a virtual reality environment. A questionnaire based on Muir's (1987) work showed that participants trusted less aggressive AV drivers more than aggressive ones, and trusted the AV more when it operated in an environment with signalized crosswalks. This trust was coordinated with trusting behaviors of reduced distance between the participant and the AV and increased jaywalking time. Pedersen et al. (2018) also looked at questions concerning trust in AVs. They studied whether the robot's performance causing real-world consequences affected trust. Participants rode in an AV simulator, and received a small electrical shock if the vehicle crashed. Participants had the option to take over driving the car. Participants who believed there would be a consequence to the AV's behavior took over control from the car more often, and when asked, indicated that they did so because they did not trust it. In this way, even though the robot's performance did not vary, the researchers were still asking about whether the humans trusted the robot's performance enough to let the car drive itself.

In summary, these researchers were interested in the following questions: how did knowledge of a robot's capability in one task transfer to trust in the robot's capabilities in another task (Shu et al., 2018; Soh et al., 2018); how did a robot's obvious correctness or incorrectness in a first interaction affect trust (Xu & Howard, 2018); how did robots' previous correctness or incorrectness

affect trust when the robot then made a mistake for children interaction with robots (Geiskkovitch et al., 2019); how did varying levels of robot error severity affect trust (Rossi et al., 2018a); how did varying levels of robot task completion affect trust (Byrne & Marín, 2018); how did a robot that accounted for a human's evolving trust levels affect overall trust (Chen et al., 2018); how did aggressive versus non-aggressive driving of an AV affect pedestrian's trust in it (Jayaraman et al., 2018); and finally did having real-world consequences to an AV's mistake affect trust in the AV (Pedersen et al., 2018). These question all reflect interest in a performance-based trust of the robot, because the factors that were manipulated were all about how the robot performed in specific tasks.

4.1.2 Relation-Based Questions

Relation-based trust questions do not address the robot performance; if there is a task, the robot's performance is the same in all conditions, and the researchers ask about some aspect about social factors, conventions, or norms. There is a wide variety of these questions that have been used in the HRI literature.

Behrens et al. (2018), Kraus et al. (2018), and Ghazali et al. (2018) all looked at how robot gender affects trust. Behrens et al. (2018) manipulated the robot's voice to indicate its gender, and found that participants saw the male robot as more trustworthy and were willing to share more personal information with it. Kraus et al. (2018) also manipulated the robot's voice, as well as its name to give it an explicit gender, and gender stereotypical personalities and tasks to give it an implied gender. They then studied how gender stereotypes affected trust, based on either explicit or implicit robot gender. Participants indicated through a validated trust scale that stereotypically male behaviors and tasks were trusted more than stereotypically female behaviors and tasks.

Ghazali et al. (2018) looked at how gender similarity or dissimilarity between the participant and robot, as well as the robot's facial expression, affected trust. The facial expressions were made to be either trustworthy or untrustworthy, based on social neuroscience research. Participants trusted the trustworthy face more than the untrustworthy face, as indicated via a survey. Unlike the previous two studies, the authors found no gender effects. These results were also seen in the participants' decisions about whether or not to follow the robot's advice. You and Robert (2018) also asked about the similarity and dissimilarity between robot and participant by varying the robot's gender and attitude about work to match or not match the participant, as well as the

danger risk of the task. Based on their answers from Jian et al.'s (2000) Trust in Automation Scale, robots that were similar to the person on a deep level, attitude about work, were trusted more than those that were dissimilar on a deep level. Surface level similarity, the robot's gender, was only significant when the danger risk was low, and then similar robots were trusted more than dissimilar ones.

Sanders et al. (2019) studied how the type of agent (human versus robot) influenced trust. They found that people chose a robot over a human to complete a dangerous task, a human over a robot to complete a menial task, and that participants rated their trust in whichever agent they chose higher than the one they did not choose, based on Jian et al.'s (2000) Trust in Automation Scale. Correia et al. (2018b) also looked at the effect of group dynamics, and varied whether robots in a group that contained both robots and humans expressed individual emotions or group emotions. They found that when a robot teammate expressed group emotions rather than individual, the participants' expressed a higher level of group trust via a post-experiment survey. Finally, Herse et al. (2018) varied the agent's embodiment and whether or not it considered a person's preferences before offering a restaurant suggestion. They found that participants trusted the robot by following its suggestion more if the robot considered a person's preferences.

In summary, these researchers were interested in the following questions: how did the gendered voice of a robot affect trust (Behrens et al., 2018); how did explicit robot gender and implicit gender stereotypes affect trust (Kraus et al., 2018); how did gender similarity or dissimilarity between the participant and the robot, as well as the robot's facial expression, affect trust (Ghazali et al., 2018); how similarity or dissimilarity between the robot and the person's gender and attitude about work affect trust (You & Robert, 2018); how did the type of agent, robot or human, with whom the participant was interacting, affect trust (Sanders et al., 2019); how did group dynamics and expression of group emotions affect (Correia et al., 2018b); how did embodiment and consideration of a person's preferences affect trust (Herse et al., 2018). For all of the above studies, the questions that are being asked are fundamentally social in nature. These are factors that are inherent to social interactions and that will inevitably become increasingly relevant as robots become increasingly prevalent social characters.

4.1.3 Mixed Questions

Experiments that mixed relation-based and performance-based factors either did so explicitly by varying both the performance-based and a relation-based factor, or the robot's performance was tied up implicitly in a relational aspect as well. The main type of question that is being asked which implicitly entails both relation-based and performance-based types of trust is that of a robot's communication style. Volante et al. (2018) investigated robot communication and social conformity by manipulating whether or not the robot communicated, and how other people (confederates) were viewing the robot. They found that other people's views on the robot's behavior affected trust where more positive views resulted in greater trust, as reported by the HRI Trust Perception Scale (Schaefer, 2013). They found no effect of robot communication. Salomons et al. (2018) also looked at trust in a social conformity setting, like the famous Asch (1956) conformity study, by seeing if ambiguity in correct answers resulted in participants trusting the robots' group decision more if they have access to the group's answers before submitting their own final answer. They found that when participants could see the robots' answers before submitting their own, they trusted the robots, as indicated by conforming to their answers, more than when they could not see the robots' answers. Haspiel et al. (2018) and Wang et al. (2018) both looked at how behavioral explanations provided by the robot affected trust. Haspiel et al. (2018) focused on the timing of explanations in regards to expectation violation, and found that providing explanations before behavior resulted in higher trust ratings on Muir's (1987) scale. Wang et al. (2018) focused on the interaction between explanations, embodiment, and error communication strategy, and found that participants rated trust higher on Mayer's (1999) scale and followed the robot's suggestions more when the robot offered an explanation of its decision. One type of error communication strategy, that of vocal warnings of uncertainty, was studied by Christensen et al. (2019), and was found to not have an effect on for how long participants followed a robot's advice.

Fischer et al. (2018) examined how the robot's transparency about its own behavior, as well as how much it adapted to a patient's needs, affected trust. They found that participants rated their trust of the robot higher when the robot was transparent, but adaptability did not have an effect. Other researchers have looked at robot communication when the robot fails at a task, and whether it is able to repair trust. Communication in some form is a necessary part of social

interaction. Correia et al. (2018a) examined how the manner in which a robot justified its mistakes could repair trust, and found that when consequence of the failure was not severe, participants rated their trust as higher when the robot justified the mistake than when it did not, as rated on the Schaefer (2013) scale. Sebo et al. (2019) looked at how the robot framed its actions that violated trust, as well as the manner in which it attempted to repair trust. They found that robots that explained their behavior as a mistake and who then apologized for it were most likely to have repaired trust, as seen by participants performing retaliating behavior less often after the violation, and a higher rating on the Dyadic Trust Scale measurement (Larzelere & Huston, 1980). In these studies, the performance-based trust is being tested because the way that the robot explains its actions may influence how the person interprets its performance. At the same time, communication is an inherently social and relation-based aspect of interacting.

Another type of implicit factor affecting trust is varying a robot's competitiveness in a task (Novitzky, Robinette, Benjamin, Gleason, Fitzgerald, & Schmidt, 2018; Robinette, Novitzky, Fitzgerald, Benjamin, & Schmidt, 2019). Both studies found that participants would trust the more aggressive, competitive robot to be on their team more than the non-competitive robot. This clearly has to do with performance because it is competing to win and therefore needs to do well to do so, but there is also a relation-based aspect to being competitive. These robots may not be implementing the "best" strategy, but they are being aggressive in order to make sure that they are at least better than their competitor. Competitiveness implies a recognition and consideration of another social agent in a task.

Other studies explicitly looked at how multiple factors affected trust, some of which were performance-based trust questions and some of which were relation-based questions. Gombolay et al. (2018) asked about the interaction between embodiment and quality of a suggestion in willingness to accept the suggestion. They found that participants were more likely to comply with the agent's suggestion at an inappropriate time (i.e., low-quality suggestions) when the agent was a non-embodied computer. When the agent was an embodied robot, participants exhibited an appropriate amount of dependence on the agent (i.e., did accept high-quality suggestions but did not accept low-quality suggestions). Xie etal. (2019) varied the type of agent, human versus robot, and the agent's capability and intent. They found that a robot whose intent and capability is similar between an observed task and an unobserved task will be trusted more to perform a task than one

whose capabilities and intent are different from what is needed in the unobserved task.

In summary, these researchers were interested in questions about how the following factors affected trust: robot communication and group perception of the robot (Volante et al., 2018); knowledge about a group of robot's answers to an ambiguous question (Salomons et al., 2018); timing of expectation violation explanations (Haspiel et al., 2018); explanations, embodiment, and error communication strategy (Wang et al., 2018); vocal warnings of uncertainty (Christensen et al., 2019); behavior transparency and adaptation (Fischer et al., 2018); different manners of mistake justification (Correia et al., 2018a); framing of actions that violated trust and manner in which it attempted to repair trust (Sebo et al., 2019); robot competitiveness (Novitzky et al., 2018; Robinette et al., 2019); the interaction of embodiment and robot suggestion quality (Gombolay et al., 2018); type of agent and agent capability and intent (Xie et al., 2019). These questions ask about factors that are a mix of performance- and relation-based trust, either implicitly or explicitly.

4.2 Surveys

The following studies are ones in which the authors did not manipulate anything about the robot or the environment; rather, either all participants experienced the same stimuli, or the researchers examined factors that they themselves did not control. Rossi et al. (2018b) got at a performance-based question by examining whether increasing participants' levels of understanding of how a Pepper robot worked increased their trust in it. They first showed the participants a video of Pepper, then a live interaction demo, then allowed them to program it. Participants trusted Pepper's capability to help them with homework or wake them up from school the most after the live demo. They trusted it to help them in a dangerous situation the most after programming it themselves. Weigelin et al.'s (2018) study is also about performance-based trust; in it, participants' dyad teams pretended to be a healthcare patient and physical therapist, and used a robot to help the "patient" perform physical therapy tasks. They found that the robot's performance affected the participants' distress levels, which in turn affected usability and trust.

Lyons and Guznov (2018) examined whether individual differences and biases in believing automation to be perfect, relation-based factors, influenced trust in a robotic system. They found that there was a positive correlation between the two based on survey responses to Merritt et al.'s (2012) and Mayers et al.'s (1995) questionnaires. Newaz and Saplacan (2018) and van Straten et al. (2018)

both examined mixed performance-based and relation-based factors. Newaz and Saplacan (2018) asked about people's subjective experiences with robotic vacuums based on the feedback it gave, and found that participants reported a lack of trust if there was a lack of feedback. Van Straten et al. (2018) specifically asked whether children differentiate between interpersonal (relation-based) trust and technological (performance-based) trust. In interviews, children differentiated between the two types of trust.

In summary, these researchers were interested in questions about how the following performance-based factors affected trust: knowledge of the robot's capabilities (Rossi et al., 2018b); use of a physical therapy robot (Weigelin et al., 2018).

These researchers were interested in questions about how the following relation-based factors affected trust: individual differences and biases towards believing automation to be perfect (Lyons & Guznov, 2018).

Finally, these researchers were interested in questions about how the following mixed factors affected trust: subjective experience with a home vacuum robot (Newaz & Saplacan, 2018); differentiation between interpersonal and technological trust in children (van Straten et al., 2018).

In the following section, we examine which of the above studies that utilized an objective measure of trust, what that measure was, and what type of trust it was measuring.

5 Objective Measures of Trust

Subjective questionnaires remain the primary means of determining a participant's trust in a robot. While a large number of studies rely on validated questionnaires to measure trust subjectively, others create their own non-validated ad hoc questionnaires. As a general trend, objective measures of trust are used far less frequently than either type of subjective questionnaires. Objective measures, however, allow us to analyze how a person actually interacts with a robot, rather than relying on their own speculation about themselves (e.g., their motivations, reasons, etc.). There are a few common ways in which researchers indirectly and objectively measure trust. We describe these in the following section.

5.1 Types of Objective Measures

Recent objective trust measures can be divided into four categories: task intervention, task delegation, behavioral change, and following advice. In task intervention scenarios, the participant interacts with a robot that is performing a task that is ordinarily done by people. Trust is measured by the number of times the participant intervenes by taking over doing the task from the robot. This type of objective measure is good for when researchers are testing a robot that could be used to take over common tasks that people perform, such as autonomous driving. Task delegation is similar in that the robot may be performing a task, but the participant decides in the end if the robot or a person should be in charge of that task. Or, in some cases, which robot out of multiple should be chosen. The agent that the participant chooses is considered the one they trust the most for that task. Similar to task intervention, this paradigm is useful for tasks that could be done by a human or a robot to investigate which people trust more for that task. These two types of paradigms are primarily appropriate in measuring performance-based trust.

Observing behavioral changes in participants is a useful objective measure when the participant cannot control the robot's behavior the way they might be able to during a task intervention or delegation task. This paradigm is about observing and measuring how participants naturally interact with and behave around different robots. Trust is measured differently in each study based on the specific nature of the interaction. This paradigm can be useful for measuring both performance-based and relation-based trust, as well as a mix of the two, in situations in which the participant has no control over the robot's behaviors.

The fourth main measure is following a robot's advice. Like the behavioral change measure, this trust measure is used when participants cannot control the robot's behavior. The robot makes suggestions or offers advice, and participants have the option whether or not to follow it. This is often used when the robot and human are teammates, or share a common goal. Trust is measured by whether or not, or how often, the participant follows the robot's advice. This is an especially useful measure for robot-human teams, and can be useful for measuring performance-based trust, relation-based trust, or a mix of the two.

In the following section, we step through each paper in our survey that used an objective measure of trust. We categorize them as measuring either performance- or relation-based trust, or

a mix of the two, and then for performance-based and mixed measures, we further categorize them by which type of objective trust measure they use. Within those categories, we describe what each measure was, and how it indicated the participant's trust in the robot.

5.2 Performance-Based Objective Measures

5.2.1 Task Intervention

The majority of objective trust measures that have been used in recent years in HRI have attempted to assess the participants' trust in terms of the robot's performance. Task intervention scenarios measure trust based on how much of the task the human allows the robot to do, and how often the human takes over doing the task from the robot. This method is utilized by Pedersen et al. (2018) and Chen et al. (2018). The former had participants interact with a self-driving car, and trust was measured by the amount of times the participant took over driving from the car. In the latter, participants worked with a robot that was clearing a table. Trust was measured as the amount of times the participant cleared objects off of the table themselves. In both tasks, fewer interventions indicated more trust in the robot.

5.2.2 Task Delegation

In task delegation scenarios, participants choose which agent does which task; unlike in the intervention scenarios, agents cannot only do some of the task some of the time. In Xie et al. (2019), participants saw the robot perform one task, and then were asked if that robot or a human should perform a different task. In Sanders et al. (2019), participants were told to pretend they were a supervisor who needed to hire agents for two different jobs. For both jobs, they were asked to choose between hiring a robot or a human. Rossi et al. (2018a) presented an emergency situation and asked participants if they would rather take care of it themselves, delegate fixing it to the robot, or work together to fix it. Task intervention and task delegation scenario are both fairly direct measures of whether or not a person trusts a robot to perform a task successfully and satisfactorily.

5.2.3 Behavioral Change

A more indirect way to measure trust in a robot's performance is to analyze how participants' behaviors change based on how the robot acts. Jayaraman et al. (2018) looked at street crossing behavior as participants acted as pedestrians while an autonomous vehicle (AV) came towards them. The authors considered how much space the participants left between themselves and the AV, how long they waited before crossing the street, and how often they jaywalked to indicate how much the participants trusted the AV's behavior. In Weigelin et al. (2018), human dyads used a robotic arm to perform kinesthetic therapy tasks. The authors analyzed videos of the interactions for signs of distress, which they took to indicate that the subjects did not trust the robot. In behavioral change scenarios, while the authors did not ask participants about the robot's performance directly, they could extrapolate their feelings of trust based on participants' behavioral cues.

5.2.4 Following Advice

The most commonly used objective performance trust measure is whether or not a person follows the task performance advice that a robot gives. In Gombolay et al. (2018), a robot gave nurses suggestions about how they should distribute nursing station assignments across staff. Wang et al. (2018) had participants complete a virtual reconnaissance task, and provided a robot that would tell the person if it believed the building was safe to enter or not. Participants had the choice of whether or not they listened to the robot's safety assessment. Christensen et al. (2019) had participants navigate blindly through the maze while a robot provided instructions about which direction to turn; however, it eventually became obvious that the robot was leading the participant in a circle. The amount of time participants continued to listen to the robot's instructions indicated how much the participants trusted the robot. In Xu and Howard (2018), participants had to report the number of toothpicks that were briefly shown on a screen, and were told that they were competing against another team. The participant's robot teammate advised the participant about how many toothpicks it believed there to be, and participants had the option to change their answer to the robot's answer. Finally, Geiskkovitch et al. (2019) had children interact with two different robots, one of whom labeled a common object correctly and the other incorrectly. Then each robot held up a different unfamiliar object, and provided the same name for it. The child was asked which of the

two objects matched the name. The one that the child chose indicated which robot it trusted to provide the correct information, based likely on its performance in previous rounds. In all of these studies, the participants had the option to ignore the robot's advice or suggestions and continue to perform the task their own way. However, if they trusted the robot's performance, they should follow the advice that it provided in order to successfully complete the task.

5.3 Relation-Based Objective Measures

Of the surveyed literature, Behrens et al. (2018) was the only study that used a solely relation-based objective trust measure to indicate a participant's relation-based trust in a robot. Participants were first shown an image of a robot and heard it speak, and were asked to indicate what information from a provided list they would be willing to share with the robot. In a second study, participants interacted in person with a robot, and were asked to share with it an embarrassing personal story and login credentials for a website. Trust was measured in both studies as the amount of information the person was willing to share with the robot. There was no task that the robot needed to perform, and therefore the participants did not need to rely on the robot's behavior. Instead, they needed to trust it in a social manner to not share their vulnerable information with a third party. This can be considered a behavioral change scenario, because participants would choose to change their behavior and tell the robot their vulnerable information if they trusted it.

5.4 Mixed Objective Measures

5.4.1 Task Delegation

There are a number of studies that measure trust primarily in a performance-based way, but that also depend on a relation-based aspect of trust as well. In the surveyed studies, none of the papers that used a mixed performance-based and relation-based objective trust measure utilized a task intervention scenario. Novitzky et al. (2018) and Robinette et al. (2019) used a trust delegation scenario. In both studies, trust was determined by a participant choosing which robot they would want as their teammate. In this scenario, participants watched two different robots perform a task, and were then asked which of the two robots they would want as their teammate if they had to perform the same task. Because the task was a competitive one, the participants needed to trust

that whichever robot they delegated the task to would help them succeed and perform better than the other team. However, there is also a relation-based aspect to choosing a teammate. A person may not choose someone or something that they do not trust to have adequate social skills if they need to work closely with it.

5.4.2 Behavioral Change

Sebo et al. (2019) took the approach of the participant competing against a robot rather than with it as its teammate. The participant and the robot played a video game against one another, and at the start the robot promised not to utilize a power-up that would be detrimental to the human player as long as the human made the same agreement. The robot then violated that agreement and used the power-up, and trust was measured by seeing if the person retaliated by also using the power-up. The person could trust the robot socially to not break their pact again, or it could trust that its performance would be better if it did not use the power-up again. The person's behavior in the game indicated how much they trusted the robot.

In another study in which trust was measured based on whether or not participants changed their behavior based on a robot's answers, Salomons et al. (2018) asked people to play a modified version of Asch's (1956) conformity study. Participants had to submit an answer to the question about which image matched a given ambiguous word, and then would hear the answers of a group of robots. The participants could then change their answer to the robot's before submitting their final answer. Because the words to which they needed to match images were ambiguous, participants who switched their answer to the robot's may have trusted the robot's performance and understanding of the words more than their own. In line with the Asch studies, however, there can be a relation-based trust in conforming to the group with whom one is a part.

5.4.3 Following Advice

As discussed previously, there are some studies in which following a robot's advice depends almost exclusively on a person's trust in the robot's performance. However, there are also times in which following the robot's advice requires relational trust as well. Herse et al. (2018) had participants decide whether or not they trusted the restaurant recommendation that a robot offered them. The performance of the robot mattered because there is likely a level of poor recommendation in which

a person will never accept the suggestion. But trusting the recommendation also implies that the robot understood something about the person's preferences, and social norms about the types of restaurants that should be suggested to people (i.e., when asked for a restaurant recommendation, a person is unlikely to seriously answer with McDonald's). Ghazali et al. (2018) used another food-based scenario. Participants were told to make a drink for an alien, with a robot available to give advice on what ingredients to use. Similar to the restaurant scenario, there is a performance-based trust aspect in that one must trust it not to suggest an undrinkable item, like rocks. And there is the trust that it does not want the alien, as another social being, to suffer, so it will not suggest something feasible but inappropriate for a drink, such as sardines.

6 Comparing within Studies

In the above sections, we divided trust in robots into two factors, performance-based trust and relation-based trust (as well as a mix of the two). Table 1 shows an overview of the provided definitions of trust, trust questions, and objective measures of trust seen in the recent HRI literature categorized into these types of trust.

Ideally, in a given paper, the trust category of the definition, question, and objective measure (if used) should match; the study would then be consistent in how it views trust, how that view plays into the factors about trust it is studying, and how its results indicate that the participant trusted the robot. The above studies that used an objective measure of trust are fairly evenly split between studies that are investigating performance-based factors that influence trust, relation-based factors, and mixed performance-based and relation-based factors. In this section, we categorize papers based on the type of trust that their questions of interest reflected, and discuss the subsequent trust types into which their trust definition and objective measure fall. This section therefore only discusses the surveyed literature that included an objective measure of trust.

6.1 Performance-Based Papers

The five studies that focused on questions about the robots performance all use objective measures that solely indicated the participants' trust in the reliability and capability of the robot's performance (Weigelin et al., 2018; Xu & Howard, 2018; Pedersen et al., 2018; Jayaraman et al.,

2018; Geiskkovitch, 2019). Of these five, Geiskkovitch et al. (2019) and Weigelin et al. (2018) use definitions of trust that specifically acknowledge reliability (Rotter, 1971) of the system (Weigelin et al., 2018). Xu and Howard (2018) used a definition of trust in which the trustee mitigates the risk of the truster (Robinette et al., 2017). Similarly, Jayaraman et al.'s (2018) definition of trust is a willingness to be vulnerable to the actions of a robotic system. While these definitions do not explicitly reference reliability or capability, they are implied in the assumption that the performance will determine whether or not a vulnerable agent is at risk. Pedersen et al. (2018) did not provide an operational definition of trust, but participants in their study believed themselves to be physically vulnerable because they believed they would be shocked if the AV did not perform properly, so they needed to trust it to be reliable. Current questions that are solely about performance-based trust, therefore, have objective measures and trust definitions that align with the goal of study performance-based factors quite well.

6.2 Relation-Based Papers

As indicated above, Behrens et al. (2018) was the only study to use an objective trust measure that got purely at a relation-based trust, rather than a relation-based and performance-based trust or just a performance-based trust. The authors were asking about how gender, a relation-based factor, influenced trust; however, they did not provide an operational trust definition. Two of the four of the studies that investigated relation-based factors of trust used objective measures that indicated a mix of both performance-based and relation-based trust (Herse et al., 2018; Ghazali et al., 2018). In these papers, Herse et al. (2018) used a definition of trust that is based on performance and reliability (Lewis et al., 2018; Moray & Inagaki, 1999). Ghazali et al.'s (2018) study did not explicitly state their operational trust definition. Sanders et al. (2019) used an objective measure of trust that seems to indicate trust primarily in the robot's performance. They reference Lee and See (2004) and Hancock et al.'s (2011a) in their definition of trust; thus, their operational definition involves uncertainty, vulnerability, and reliance that one agent's actions will not be used to harm another. This definition does not clearly fall into being either a performance-based or relation-based type of trust definition. For relation-based trust papers, therefore, it is less clear than relation-based trust is actually being measured.

6.3 Mixed Papers

Studies that asked about both relation-based and performance-based trust factors are split between using objective measures that are just performance-based and ones that are both relation-based and performance-based. Novitzky et al. (2018) and Robinette et al. (2019) both used a performance-based/relation-based measure, and defined trust as the trustee mitigating risk for the truster (Wagner & Arkin, 2011). Similarly, Sebo et al. (2019) and Salomons et al. (2018) had measures that implied both types of trust. The former referenced Mayer's (1995) trust definition about an agent being willing to be vulnerable to another's actions without necessarily monitoring said actions; the latter defined truth as "how reliable other sources are believed to be" (Salomons et al., 2018, p. 189).

Christensen et al. (2019) had a question that was both relation-based and performance-based, and an objective measure and trust definition that were just performance-based. They referenced Hancock et al.'s (2011b) trust definition of being reliable and predictable. Xie et al. (2019) and Gombolay et al. (2018) had performance-based measures, and questions and definitions that were both relation-based and performance-based. The former defined trust as a summary of past experiences to predict future behavior in a vulnerable scenario (Soh et al., 2018; Chen et al., 2018), and the latter as uncertainty and vulnerability (Lee & See, 2004). Wang et al. (2018) did not provide an operational definition of trust. Objective trust measures that are performance-based only potentially miss something important when they are being used to explore how both relation-based and performance-based factors affect trust.

7 Discussion

Our survey of the current literature in HRI on human trust in robots reveals a fairly wide-spread conflation between performance-based trust and relation-based trust. Definitions and questions are not always clear about the type of trust in which the researchers are interested, which results in objective trust measures not having a clear indication of the type of trust they are measuring. For example, Ghazali et al. (2018) characterize trust to be necessary for people to "feel safe to rely on social robots for physical or even emotional support" (p. 2). Their objective measure, however, was whether or not people took the suggestion of a robot for what to put in a drink the participant

was supposed to make for an alien. This does not seem to be an indicator for whether or not the person felt like they could rely on the robot for physical or emotional support.

There is also a disconnect between the types of questions that people are asking and the way that they are attempting to answer those questions. Papers that focus on relation-based questions are mostly either only using subjective measures or an objective measure that mixes performance-based trust and relation-based trust. Sanders et al. (2019) showed participants videos of humans and robots completing different tasks, and had the participants imagine that they were the supervisor of the task and to choose which agent they would want to complete it. Varying the type of agent is a relation-based factor, but choosing an agent to hire is more likely to get at trusting their performance. Conversely, papers that focus on performance-based trust rarely use validated questionnaires when collecting subjective data.

There is a wide and varying field of questions about relation-based trust, but there seems little interest in measuring relation-based trust questions objectively. Relation-based trust is almost exclusively measured by (subjective) questionnaires. If we want to understand how we trust social robots in social contexts and interactions, we need to develop ways to measure relation-based trust objectively. As the studies presented in this survey and others have shown, varying levels of trust can affect the way people actually interact with a robot. However, their objective task behavior may not match their subjective reflections about themselves when they answer a questionnaire. If robots are going to be in our society as social agents, we need to see how people actually react to them, not just how they believe they will react. As we have outlined here, the types of questions that researchers are asking about trust do not necessarily line up with the manner with which they are defining trust, nor the way they are measuring trust objectively. For the HRI community to make headway, we need a more formalized and universal approach to trust, so that the definitions, questions, and measurements can align for a given study. To do this, we first need an agreed upon set of trust definitions.

Note that there does not necessarily need to be only one definition of trust; as we indicated, trust in HRI can be split at least into two broad categories, though there are likely more nuanced ways to divide it. From these formal definitions, we will have a clearer path to creating a regular paradigm or paradigms that can measure trust objectively, allowing us to compare results across studies. We can perhaps borrow heavily from fields that have been studying trust for much longer,

such as psychology and human factors. Behavioral game theory could also be a promising direction; it is a tool used in psychology and behavioral economics which utilizes social, cooperative games to objectively measure the trust a person has in the other agent playing with them. The manner in which people play is indicative of their level of trust. Similarly, we need to be clear about the types of trust which our questionnaires are asking. A more formalized path forward will allow for the field's greater advancement. Based on these considerations, we make a few explicit recommendations.

7.1 Recommendation for Formalization of Trust

We propose the following formal definition of trust. A *truster* is the person doing the trusting; a *trustee* is the person or system that is being trusted; *entrusting with* is to assign the responsibility of doing something (to someone). As an example situation, take the sentence "The robot is going to take out the trash in the kitchen after dinner." The robot is the trustee who was entrusted with taking out the trash, and the person who owns or is in charge of the robot is the truster. We have a context (the kitchen), a time (after dinner) and a behavior (take out trash). Hence, we can construe trust as a four-place relation such that:

entrust(the-robot,take-out(trash),kitchen,after-dinner)

Bringing in explicitly the truster A, i.e., the person that trusts the trustee B with regard to some behavior X in context Y at time t, turns trust into a five-place relation:

entrust(A,B,X,Y,t) or **trusts-that**(A,B,ϕ) where ϕ = X in Y at t

Note that A or B can either be persons or machines, although some would argue that machines cannot really trust. We can then ask what it would take for a machine or person to trust another machine or person with regard to some behavior X in context Y at time t. From the answer we can ascertain that if those conditions of trust were met, the person/machine would indeed entrust the other person/machine with doing X in context Y at time t.

7.2 Recommendations for Experimental Paradigms

Surprisingly, none of the recent HRI trust studies used any established games like the Prisoner's Dilemma or Dictator Game to model a human's trust in a robot teammate. These are games which have a clear way to maximize personal reward. Most participants in these games, however, choose to pursue a different strategy, implying that they bring forth social heuristics and expectations when

interacting with the other player (Murnighan & Wang, 2016). Therefore, the manner in which an agent plays one of these games can indicate how much they trust the agent with whom they are interacting because they are putting themselves in a vulnerable situation by not going for a selfish strategy that would maximize their own score and ruin the other player's. They have to trust that if they play in that manner, the other play will as well. These games are used to measure trust in human-human interactions (Alarcon, Lyons, Christensen, Bowers, Klosterman, & Capiola, 2018), and have been used before to measure trust in HRI (DeSteno, Brezeal, Frank, Pizarro, Baumann, Dickens, & Lee, 2012). However, this type of trust measure has not appeared in HRI literature in recent years; it may be a promising path forward that will allow for measuring relational trust objectively.

Though Behrens et al. (2018) was the only study to use a solely relation-based objective trust measure, the relation-based papers were more likely than the performance-based papers to use a validated subjective trust questionnaire. The performance-based papers relied primarily on ad hoc questionnaires if they used a subjective measure. Ideally, if a study uses a subjective questionnaire, it will rely on one that has been validated. For relation-based trust papers, this may mean leaning away from surveys that measure trust in the robot's performance, and more towards trust as a social agent (i.e., Ullman & Malle, 2018).

7.3 Recommendations for Trust Questions

Nearly all of the papers about trust in HRI are about trusting a robot to do a specific task in a specific context. However, human trust is more general than that. For example, John and Mary are friends, Mary will probably trust John to pick up her mother at the airport, even if she has never experienced John driving to the airport or meeting her mother. As humans, we transfer and generalize trust to novel situations. This allows for an ease in interacting with people that we should aim to have with our robots, especially social robots. It would be tedious if a person who owns a household service robot needed to see it perform each task that it could possibly do before entrusting it with that task. Therefore, we need to understand how human trust in robots transfer to novel situations. Soh et al. (2018) and Shu et al. (2018) have begun doing this type of research, but significantly more needs to be done in order for us to be able to fully utilize people's ability to transfer trust in our robotic systems.

8 Conclusion

In this survey, we have provided an overview of the current state of measuring and understanding trust in HRI. We have examined the common definitions and their relevance to the field, and the questions that researchers are currently interested in studying. We have also laid out the ways in which those trust questions are being measured objectively, and what types of trust those objective measures may imply. What we have shown is that there is great interest in questions that harken back to a kind of relation-based trust, even though there are virtually no studies that objectively measure a purely relation-based trust. As a result, we have very little understanding of how people may trust robots to perform social tasks that do not have a clear performance goal, something that we will need to understand as robots become more prevalent social agents. We have also highlighted discrepancies between questions that are being studied and how trust is defined and measured. This results in a lack of clarity in what is actually meant by "trust." For the advancement of the field, we recommend adopting a more formalized definition of trust, a generalized objective trust paradigm, especially for measuring relation-based trust, and a focus on how people may transfer trust in a robot from one task to a general series of tasks.

9 Acknowledgements

This work was funded by AFOSR grant FA9550-18-1-0465.

References

[1] Alarcon, G. M., Lyons, J. B., Christensen, J. C., Bowers, M. A., Klosterman, S. L., & Capiola, A. (2018). The role of propensity to trust and the five factor model across the trust process. *Journal of Research in Personality, 75*, 69-82.

[2] Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological monographs: General and applied, 70* (9), 1.

[3] Behrens, S. I., Egsvang, A. K. K., Hansen, M., & Møllegård-Schroll, A. M. (2018, March). Gendered Robot Voices and Their Influence on Trust. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 63-64). ACM.

[4] Byrne, K., & Marín, C. (2018, June). Human Trust in Robots When Performing a Service. In *2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)* (pp. 9-14). IEEE.

[5] Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model* (Vol. 18). John Wiley & Sons.

[6] Chen, M., Nikolaidis, S., Soh, H., Hsu, D., & Srinivasa, S. (2018, February). Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/ IEEE International Conference on Human-Robot Interaction* (pp. 307-315). ACM.

[7] Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction, 31* (10), 692-702.

[8] Christensen, A. B., Dam, C. R., Rasle, C., Bauer, J. E., Mohamed, R. A., & Jensen, L. C. (2019, March). Reducing Overtrust in Failing Robotic Systems. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*(pp. 542-543). IEEE.

[9] Correia, F., Guerra, C., Mascarenhas, S., Melo, F. S., & Paiva, A. (2018a, July). Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*(pp. 507-513). International Foundation for Autonomous Agents and Multiagent Systems.

[10] Correia, F., Mascarenhas, S., Prada, R., Melo, F. S., & Paiva, A. (2018b, February). Group-based emotions in teams of humans and robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 261-269). ACM.

[11] DeSteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L., & Lee, J. J. (2012). Detecting the trustworthiness of novel partners in economic exchange. *Psychological science, 23*(12), 1549-1556.

[12] Dunning, D., & Fetchenhauer, D. (2011). *Understanding the psychology of trust.* Psychology Press.

[13] Fischer, K., Weigelin, H. M., & Bodenhagen, L. (2018). Increasing trust in human–robot medical interactions: effects of transparency and adaptability. *Paladyn, Journal of Behavioral Robotics, 9*(1), 95-109.

[14] Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007, May). Measurement of trust in human-robot collaboration. In *2007 International Symposium on Collaborative Technologies and Systems* (pp. 106-114). IEEE.

[15] Geiskkovitch, D. Y., Thiessen, R., Young, J. E., & Glenwright, M. R. (2019, March). What? That's Not a Chair!: How Robot Informational Errors Affect Children's Trust Towards Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 48-56). IEEE.

[16] Ghazali, A. S., Ham, J., Barakova, E. I., & Markopoulos, P. (2018). Effects of robot facial characteristics and gender in persuasive human-robot interaction. *Frontiers in Robotics and AI, 5*, 73.

[17] Gombolay, M., Yang, X. J., Hayes, B., Seo, N., Liu, Z., Wadhwania, S., ... & Shah, J. (2018). Robotic assistance in the coordination of patient care. *The International Journal of Robotics Research, 37*(10), 1300-1316.

[18] Groom, V., & Nass, C. (2007). Can robots be teammates?: Benchmarks in human–robot teams. *Interaction Studies, 8*(3), 483-500.

[19] Hancock, P. A., Billings, D. R., & Schaefer, K. E. (2011a). Can you trust your robot?. *Ergonomics in Design, 19*(3), 24-29.

[20] Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011b). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors, 53*(5), 517-527.

[21] Haspiel, J., Du, N., Meyerson, J., Robert Jr, L. P., Tilbury, D., Yang, X. J., & Pradhan, A. K. (2018, March). Explanations and Expectations: Trust Building in Automated Vehicles.

In *Companion of the 2018 ACM/IEEE International Conference on Human- Robot Interaction* (pp. 119-120). ACM.

[22] Heerink, M., Krose, B., Evers, V., & Wielinga, B. (2009, September). Measuring acceptance of an assistive social robot: a suggested toolkit. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 528-533). IEEE.

[23] Herse, S., Vitale, J., Tonkin, M., Ebrahimian, D., Ojha, S., Johnston, B., ... & Williams, M. A. (2018, August). Do You Trust Me, Blindly? Factors Influencing Trust Towards a Robot Recommender System. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 7-14). IEEE.

[24] Hodges, B. H. (2014). Rethinking conformity and imitation: Divergence, convergence, and social understanding. *Frontiers in psychology, 5*, 726.

[25] Hodges, B. H., & Geyer, A. L. (2006). A nonconformist account of the Asch experiments: Values, pragmatics, and moral dilemmas. *Personality and Social Psychology Review, 10*(1), 2-19.

[26] Jayaraman, S. K., Creech, C., Robert Jr, L. P., Tilbury, D. M., Yang, X. J., Pradhan, A. K., & Tsui, K. M. (2018, March). Trust in AV: An Uncertainty Reduction Model of AV-Pedestrian Interactions. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 133-134). ACM.

[27] Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53-71.

[28] Kraus, M., Kraus, J., Baumann, M., & Minker, W. (2018). Effects of Gender Stereotypes on Trust and Likability in Spoken Human-Robot Interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC-2018).

[29] Larzelere, R. E., & Huston, T. L. (1980). The dyadic trust scale: Toward understanding interpersonal trust in close relationships. *Journal of Marriage and the Family,* 595-604.

[30] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors, 46*(1), 50-80.

[31] Lewis, M., Sycara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In *Foundations of Trusted Autonomy* (pp. 135-159). Springer, Cham.

[32] Lyons, J. B., & Guznov, S. Y. (2018). Individual differences in human–machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science*, 1-19.

[33] Madsen, M., & Gregor, S. (2000, December). Measuring human-computer trust. In *11th australasian conference on information systems* (Vol. 53, pp. 6-8).

[34] Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of applied psychology, 84*(1), 123.

[35] Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review, 20*(3), 709-734.

[36] McCroskey, J. C., & Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement. *Communications Monographs, 66*(1), 90-103.

[37] Merritt, S. M., LaChapell, J., & Lee, D. (2012). *The perfect automation schema: Measure development and validation*. Technical report submitted to the Air Force Research Laboratory, Human Effectiveness Directorate.

[38] Moray, N., & Inagaki, T. (1999). Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control, 21*(4-5), 203-211.

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

[40] Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics, 37*(11), 1905-1922.

[41] Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39*(3), 429-460.

[42] Murnighan, J. K., & Wang, L. (2016). The social world as an experimental game. *Organizational Behavior and Human Decision Processes, 136*, 80-94.

[43] Newaz, F., & Saplacan, D. (2018, September). Exploring the role of feedback on trust for the robots used in homes of the elderly. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction* (pp. 681-685). ACM.

[44] Novitzky, M., Robinette, P., Benjamin, M. R., Gleason, D. K., Fitzgerald, C., & Schmidt, H. (2018, March). Preliminary interactions of human-robot trust, cognitive load, and robot intelligence levels in a competitive game. In *Companion of the 2018 ACM/ IEEE International Conference on Human-Robot Interaction* (pp. 203-204). ACM.

[45] Pedersen, B. K. M. K., Andersen, K. E., Köslich, S., Weigelin, B. C., & Kuusinen, K. (2018, March). Simulations and Self-Driving Cars: A Study of Trust and Consequences. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 205-206). ACM.

[46] Robinette, P., Howard, A. M., & Wagner, A. R. (2017). Effect of robot performance on human–robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems, 47*(4), 425-436.

[47] Robinette, P., Novitzky, M., Fitzgerald, C., Benjamin, M. R., & Schmidt, H. (2019, March). Exploring Human-Robot Trust During Teaming in a Real-World Testbed. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 592-593). IEEE.

[48] Rossi, A., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2018a). The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario. *Paladyn, Journal of Behavioral Robotics, 9*(1), 137-154.

[49] Rossi, A., Holthaus, P., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2018b, December). Getting to know Pepper: Effects of people's awareness of a robot's capabilities on their trust in the robot. In *Proceedings of the 6th International Conference on Human-Agent Interaction* (pp. 246-252). ACM.

[50] Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust 1. *Journal of personality, 35(*4), 651-665.

[51] Rotter, J. B. (1971). Generalized expectancies for interpersonal trust. *American psychologist, 26*(5), 443.

[52] Salomons, N., van der Linden, M., Strohkorb Sebo, S., & Scassellati, B. (2018, February). Humans conform to robots: Disambiguating trust, truth, and conformity. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 187-195). ACM.

[53] Sanders, T., Kaplan, A., Koch, R., Schwartz, M., & Hancock, P. A. (2019). The Relationship Between Trust and Use Choice in Human-Robot Interaction. *Human factors*, 0018720818816838.

[54] Schaefer, K. E. (2013). The perception and measurement of human–robot trust (Doctoral dissertation). University of Central Florida, Orlando.

[55] Sebo, S. S., Krishnamurthi, P., & Scassellati, B. (2019, March). "I Don't Believe You": Investigating the Effects of Robot Trust Violation and Repair. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction* (HRI)(pp. 57-65). IEEE.

[56] Shu, P., Min, C., Bodala, I., Nikolaidis, S., Hsu, D., & Soh, H. (2018, March). Human trust in robot capabilities across tasks. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 241-242). ACM.

[57] Soh, H., Shu, P., Chen, M., & Hsu, D. (2018). The Transfer of Human Trust in Robot Capabilities across Tasks. *RSS*.

[58] Tay, B., Jung, Y., & Park, T. (2014). When stereotypes meet robots: the double- edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior*, 38, 75-84.

[59] Ullman, D., & Malle, B. F. (2018, March). What does it mean to trust a robot?: Steps toward a multidimensional measure of trust. In *Companion of the 2018 ACM/ IEEE International Conference on Human-Robot Interaction* (pp. 263-264). ACM.

[60] van Straten, C. L., Peter, J., Kühne, R., de Jong, C., & Barco, A. (2018, December). Technological and Interpersonal Trust in Child-Robot Interaction: An Exploratory Study. In *Proceedings of the 6th International Conference on Human- Agent Interaction* (pp. 253-259). ACM.

[61] Volante, W. G., Sosna, J., Kessler, T., Sanders, T., & Hancock, P. A. (2018). Social Conformity Effects on Trust in Simulation-Based Human-Robot Interaction. *Human factors*, 0018720818811190.

[62] Wagner, A. R., & Arkin, R. C. (2011, July). Recognizing situations that demand trust. In *2011 RO-MAN* (pp. 7-14). IEEE.

[63] Wang, N., Pynadath, D. V., Rovira, E., Barnes, M. J., & Hill, S. G. (2018, April). Is it my looks? or something i said? the impact of explanations, embodiment, and expectations on trust and performance in human-robot teams. In *International Conference on Persuasive Technology* (pp. 56-69). Springer, Cham.

[64] Weigelin, B. C., Mathiesen, M., Nielsen, C., Fischer, K., & Nielsen, J. (2018, August). Trust in medical human-robot interactions based on kinesthetic guidance. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 901-908). IEEE.

[65] Wheeless, L. R., & Grotz, J. (1977). The measurement of trust and its relationship to self-disclosure. *Human Communication Research, 3*(3), 250-257.

[66] Xie, Y., Bodala, I. P., Ong, D. C., Hsu, D., & Soh, H. (2019, March). Robot Capability and Intention in Trust-Based Decisions Across Tasks. In *2019 14th ACM/ IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 39-47). IEEE.

[67] Xu, J., & Howard, A. (2018, August). The Impact of First Impressions on Human- Robot Trust During Problem-Solving Scenarios. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 435-441). IEEE.

[68] Yamagishi, T. (1998). The structure of trust: The evolutionary games of mind and society. *Tokyo: University of Tokyo Pres.*

[69] You, S., & Robert Jr, L. P. (2018, February). Human-robot similarity and willingness to work with a robotic co-worker. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 251-260). ACM.

Table 1: Comparison within papers.

Performance-Based Papers

Citation	Definition Reference	Factors	Subjective Measure Reference	Objective Measure	Objective Measure Trust Type
Byrne & Marín (2018)	Schaefer (2013); Hancock et al. (2011b)	Task Completion	Ad hoc.		
Chen et al. (2018)	No ref. Keywords: "perceived robot ability"	Taking trust into account	Ad hoc.	Number of times participant intervened in robot's task.	Performance; task intervention
Geiskkovitch et al. (2019)	Rotter (1971)	Factual mistake		Percentage of trials in which the child chooses to side with the previously correct robot.	Performance; following advice
Jayaraman et al. (2018)	No ref. Keywords: "willingness to be vulnerable to actions"	Driving Behavior	Muir (1987)	Participant street crossing behavior	Performance; behavioral change
Pedersen et al. (2018)	No definition.	Real world consequences		If participant assumed control of AV.	Performance; task intervention
Rossi et al. (2018a)	Lee & See (2004)	Severity and timing of performance errors		If participant delegated fixing an emergency situation to the robot.	Performance; task delegation
Rossi et al. (2018b)	Lee & See (2004)	Capability awareness	Ad hoc.		
Shu et al. (2018)	Castelfranchi & Falcone (2010)	Generalize trust across tasks	Ad hoc.		
Soh et al. (2018)	Castelfranchi & Falcone (2010)	Generalize trust across tasks	Muir (1994); Muir & Moray (1996)		
Weigelin et al. (2018)	No ref. Keywords: "safety and reliability of the system"	Performance		Video analysis of behavioral signs of distress.	Performance; behavioral change
Xu & Howard (2018)	Robinette et al. (2017)	First impressions	Ad hoc.	Whether or not participant followed robot's advice.	Performance; following advice

Relation-Based Papers

Behrens et al. (2018)	No definition.	Robot gender	Ad hoc.	How much information participant shares with robot.	Relation; behavioral change
Correia et al. (2018b)	No definition.	Group emotions	Ad hoc.		
Ghazali et al. (2018)	Rotter (1967)	Gender similarity; facial expressions	Jian et al. (2000); Tay et al. (2014); Heerink et al. (2009); Wheeless & Grotz (1977)	How many times participant asked the robot for help.	Mixed; following advice
Herse et al. (2018)	Lewis et al. (2018); Moray & Inagaki (1999)	Embodiment; preference consideration	McCroskey & Teven (1999)	Participant acceptance of robot's suggestion.	Mixed; following advice
Kraus et al. (2018)	Lee & See (2004); Hancock et al. (2011b)	Robot gender; task gender stereotype	Unspecified questionnaire.		
Lyons & Guznov (2018)	Lee & See (2004)	Individual biases	Merritt et al. (2012); Mayer et al. (1995)		
Sanders et al. (2019)	Lee & See (2004); Hancock et al. (2011a)	Type of agent	Jian et al. (2000)	Which agent the participant would choose to complete the task.	Performance; task delegation
You & Robert (2018)	Groom & Nass (2007); Hancock et al. (2011b)	Similarity	Jian et al. (2000)		

Mixed Papers

Christensen et al. (2019)	Hancock et al. (2011b)	Vocal warnings of uncertainty		Amount of time spent following robot's advice.	Performance; following advice
Correia et al. (2018a)	Hancock et al. (2011a)	Mistake justification	Schaefer (2013)		
Fischer et al. (2018)	Choi & Ji (2015)	Transparency; adaptation	Ad hoc.		
Gombolay et al. (2018)	Lee & See (2004)	Embodiment; suggestion quality	Jian et al. (2000)	Whether or not participant followed the robot's advice.	Performance; following advice
Haspiel et al. (2018)	No ref. Keywords: "openness to being subjected to actions"	Explanation timing	Ad hoc.		
Newaz & Saplacan (2018)	Madsen & Gregor (2000)	Feedback	Ad hoc.		
Novitzky et al. (2018)	Wagner & Arkin (2011)	Competitiveness	Ad hoc.	Which robot participant would choose to defend their flag in a game.	Mixed; task delegation
Robinette et al. (2019)	Wagner & Arkin (2011)	Competitiveness		Which robot participant would choose as a teammate.	Mixed; task delegation
Salomons et al. (2018)	Hodges (2004); Hodges & Geyer (2006)	Conformity		Whether or not participant conformed to robots in critical round.	Mixed; behavior change
Sebo et al. (2019)	Mayer et al. (1995)	Trust violation framing; trust repair strategy	Larzelere & Huston (1980)	Participant choice of retaliating power-up usage.	Mixed; behavior change
van Straten et al. (2018)	Lee & See (2004); Madsen & Gregor (2000); You & Robert (2018)	Technological vs interpersonal trust	Ad hoc.		
Volante et al. (2018)	Freedy et al. (2007)	Robot's performance; group perception	Schaefer (2013)		
Wang et al. (2018)	No definition.	Explanations	Mayer et al. (1995)	Percentage of times participant complied with robot's suggestion.	Performance; following advice
Xie et al. (2019)	Soh et al. (2018); Chen et al. (2018)	Type of agent; intention; capability	Schaefer (2013)	Whether or not participant delegated task to the robot.	Performance; task delegation