Modeling the Human-Robot Trust Phenomenon: A Conceptual Framework based on Risk

ALAN R. WAGNER, Pennsylvania State University
PAUL ROBINETTE, Massachusetts Institute of Technology
AYANNA HOWARD, Georgia Institute of Technology

This article presents a conceptual framework for human-robot trust which uses computational representations inspired by game theory to represent a definition of trust, derived from social psychology. This conceptual framework generates several testable hypotheses related to human-robot trust. This article examines these hypotheses and a series of experiments we have conducted which both provide support for and also conflict with our framework for trust. We also discuss the methodological challenges associated with investigating trust. The article concludes with a description of the important areas for future research on the topic of human-robot trust.

CCS Concepts: • **Computer systems organization** → *Embedded and cyber-physical systems*; *Robotics*; *External interfaces for robotics*;

Additional Key Words and Phrases: Human-robot trust, trust, social robotics, risk

ACM Reference format:
Alan R. Wagner, Paul Robinette, and Ayanna Howard. 2018. Modeling the Human-Robot Trust Phenomenon: A Conceptual Framework based on Risk. *ACM Trans. Interact. Intell. Syst.* 8, 4, Article 26 (November 2018), 24 pages.
https://doi.org/10.1145/3152890

1 INTRODUCTION

Trust plays an important role during interpersonal interactions. It allows employers to leave the shop knowing that their employees will act responsibly. It allows depositors to place their entire fortune in the vaults of a bank believing that their assets will be safe. Trust permits a trustor to act in a manner that puts them at considerable risk, believing that the actions of their counterpart will mitigate that risk [61]. Although we all experience it, trust is a phenomenon that is described through many different lenses. From a research standpoint, many technical definitions have been posed [3, 8, 13, 16, 24, 26, 28, 29, 53]. Those that have taken a close look at these definitional differences have concluded that, for the most part, many of the definitions for trust are largely in agreement [43]. Although the phrasing and focus may differ, many researchers agree that the term "trust" suggests a situation in which an individual is vulnerable and their vulnerability rests with the actions, behaviors, or motivations of another individual.

Although one must be careful not to put too much stock into the origin and meaning of a single word, the definition of a word helps shape its connection to some underlying phenomenon.

Authors' addresses: A. Wagner, Hammond Building, University Park, PA 16802; A. Howard, School of Interactive Computing, 85 5th Street NW, TSRB 211, Atlanta, GA 30308; P. Robinette, MIT, 77 Massachusetts Ave 32-220, Cambridge, MA 02139.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 Association for Computing Machinery.
2160-6455/2018/11-ART26 $15.00
https://doi.org/10.1145/3152890

The important challenge then becomes how to formally ground and operationalize the notions set forth in the definition in a way that scientifically informs one's research. Conceptual frameworks serve this purpose. A conceptual framework allows one to organize ideas, make conceptual distinctions, and, ideally, leads to the discovery of important interrelationships and interdependencies. For interactions involving humans and robots, developing a conceptual framework for trust is particularly important. Because robots are embodied, their actions can have serious consequences for the humans around them. Injuries and even fatal accidents have occurred because of a robot's actions.

Our framework is meant to be descriptive with respect to interpersonal trust and prescriptive with respect to human-robot trust. Developing a framework that allows a robot to recognize and react appropriately to indications of human trust has important implications for home healthcare, search and rescue, and military applications. It is similarly vital to prevent people from trusting a robot too much. For example, we recently conducted a survey of parent's attitudes related to robotic rehabilitation exoskeletons for children with mobility impairments and found that the majority of parents expected that their children would use the device for high-risk activities such as running, jumping, or climbing [5]. Moreover, 62% of parents indicated that they would typically or completely trust their child to handle any risky situations. The average age of the children was less than 9 years old. A social robot should be able to recognize when people trust it too much, e.g., trust it to perform functions beyond its capabilities, and act to dissuade the person from placing themselves at risk. For these reasons, it is critical to develop a formal, principled conceptual framework of trust that is implementable on a robot. Most research in this area has explored a different problem: understanding the factors that influence a person's trust in a robot [11, 17]. Our research, in contrast, has focused on the development and testing of a computational framework for understanding trust and methods that allow a robot to both identify who is trusting it and whom it should trust.

The purpose of this article is thus to present an overarching theory for human-robot trust and the experimental evidence that we have generated in support of and in conflict with that theory. The experiments presented in this article were conducted over a 4-year period, involving more than 2168 unique subjects, and focused on evacuation situations in which a robot offers to guide a person during an emergency [41, 56]. The portrait that the resulting data paints is complex, highlighting the conceptual and methodological challenges associated with empirically studying human-robot trust. Nevertheless, our experiments support several important conclusions and highlight several areas where additional research is needed before robots become ubiquitous members of our social environment. Although aspects of this research have been briefly discussed in other conference and journal publications [38–40], this is the first presentation of the framework as a whole.

The remainder of this article begins by presenting the portion of the vast trust literature which is most closely related to this research. In reviewing the literature, different conceptualizations of trust are examined while also considering the feasibility of implementation on a robot (Section 2). Next, our conceptual framework for representing trust is developed and, in conjunction with the definition for trust, our conditions for gauging if a situation demands trust are presented (Section 3). Section 4 begins with several hypotheses proffered by our framework and then investigates empirical evidence for and against each of these hypotheses. The article concludes by discussing the utility of our approach and avenues for future research.

2 RELATED WORK

The concept of trust in an autonomous system has long been investigated from many different perspectives [8, 18, 19, 27, 28, 33, 46, 47]. Information withholding (deceit) [33], agent reliability [47], agent opinion based on deceitful actions [19], compliance with virtual social norms [18], and

compliance with an *a priori* set of trusted behaviors from a case study [27] have all been used to measure trust in multi-agent systems. Models of trust range from beta probability distributions over agent reliability [19] and knowledge-based formulas for trust [27] to perception-specific process models for trust [18].

Neuroscientists have used economic games to study trust development [15, 22, 35, 49]. Work in this area has shown that the development of a trusting relationship occurs with repeated, positive, and predictable interactions [9, 10, 61]. Work by King-Casas et al. [22] captured fMRI images of an individual while they played a two-player investment game. The fMRI images showed that the subject modeled and predicted the behavior of their counterpart. Deviations from the subject's predictions resulted in surprise signals along with a reassessment of the counterpart. Rilling et al. [35] used an iterated Prisoner's dilemma game coupled with fMRI images to explore social cooperation and was able to show that cooperation in this paradigm reflects positive reinforcement of altruism. These results have played an important role in shaping our framework by showing that modeling and predicting the behavior of one's interactive partner is a key component to trust.

With respect to robots, research has primarily focused on elucidating the factors that influence a person's trust in a robot. Confidence and risk have been identified as factors [11, 61] as has the robot's behavior [6] and appearance [25, 32]. Carlson et al. [7] finds that reliability and reputation impact trust in surveys of how people view a robot. Hancock et al. [17] performed a meta-analysis over the existing human-robot trust literature identifying 11 relevant research articles and found that, for these articles, robot performance is most strongly associated with trust. Desai et al. [11] performed several experiments related to human-robot trust. This group's work primarily focused on the impact of robot reliability on a user's decision to interrupt an autonomously operating robot. They found that poor robot performance negatively affected the operator's trust of the robot. In contrast to the work by Desai et al., our work on trust during robot-guided emergency evacuation does not afford an opportunity for the human to take control of the robot. Instead, we examine a situation in which a person must choose to either follow the guidance of a robot or not. While we still capture the level of trust a person places in an autonomous robot, we believe that an evacuee's perspective on trust is significantly different from an operator's perspective on trust. The evacuee has no control over the robot and must decide between his or her own intuition and the robot's instructions in a situation that presents physical danger to the person.

Some researchers have found that people will ignore their prior experience with the robot and their own common sense when a robot asks them to perform an odd and potentially destructive task. Salem et al. [45] performed an experiment to determine the effect of robot errors on unusual requests. They found that participants still completed an odd request made by a robot in spite of any previous errors. Bainbridge et al. [2] found that participants were willing to throw books in the trash when a physically present robot gave the instruction, but not when the robot was located in another room communicating through a video interface. This prior research and our own work demonstrates that people have a tendency to comply with a robot's order, and will even place themselves at risk believing that the robot knows more than they do.

3 A CONCEPTUAL FRAMEWORK OF TRUST

Several characteristics are critical for a human-robot framework for trust. First, the framework must be implementable on a robot. It must therefore be possible to translate the framework into usable software. It must also be possible for a robot to perceive the information required by the framework. A framework which does not offer a means for constructing the computational representations on which it relies is not implementable. Second, the framework should focus equally on the robot as the trustor or the trustee. Because social exchanges are dynamic, focusing solely on the robot's role as trustee limits the usefulness of the framework. Third, the framework

must make testable predictions relevant to how people perceive trust and how a robot should act in situations demanding trust. Finally, the framework should, ideally, connect to other frameworks organically. Trust should not be imposed as a separate and distinct module of standalone computational processing.

In order to couch a social phenomenon such as trust within a framework, one needs an operational definition that, to the extent possible, uniquely defines and characterizes the phenomenon. Drawing from Mayer's work on interpersonal trust [29] and Lee and See's work on trust in automation [24], we developed an operational definition of human-robot trust which is formulated within a game theory context. We define trust as, "a belief, held by the trustor, that the trustee will act in a manner that mitigates the trustor's risk in a situation in which the trustor has put its outcomes at risk". Our definition differs from Mayer's in one minor respect. Mayer characterizes trust as one's willingness to be vulnerable. We replace vulnerability with risk only because risk is a more precisely and computationally defined concept suitable for implementation on a robot. Lee and See [24] develop an automation-focused definition of trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." Lee and See's characterization of trust as an attitude differs from our characterization of trust as a belief and Mayer's characterization of trust as willingness. We use *belief* because the term has a long-standing definition within the game theory and artificial intelligence communities. In game theory, beliefs serve as evidence used to evaluate a probability distribution over an agent's potential decisions [30]. In artificial intelligence, a belief is a form of knowledge representation, which also may be used as evidence for or against a course of action [43]. We view these different notions of belief as compatible and use beliefs as a form of knowledge that are used to evaluate a probability distribution over potential decisions. By generating an operational definition of trust, we have thus created a basis for reasoning and representing the phenomenon in game theoretic terms. As we explain in Section 4.1, we have validated the definition empirically.

The conceptual framework we present is not equal to game theory. The framework simply uses game theory's underlying computational representations (Normal and extended-form games). We do not make the standard game theoretic assumption: that the players know the game in advance and more importantly, that the players act with rational self-interest because we are not interested in the traditional game theory solution concepts. Rather our intent is to develop a computational process used to control a robot during social interactions.

Game theory suggests a computational process for interaction and offers representations for trying to understand the computational underpinnings that allow trust to emerge during social interactions. Overall, our work suggests that trust emerges from social situations in which one individual must place themselves at risk and another individual holds the power to mitigate that risk. Representations from game theory allow us to model trust formally.

Our definition highlights the role of three important factors that influence trust: the trustee, the trustor, and the situation (Figure 1). Research by Hancock et al. [17] highlight similar factors after evaluating the trust literature and interviewing subject matter experts. Consider, for example, the trust fall. The trust fall is a type of game in which the trustor leans backward and the trustee catches the trustor. With respect to the definition of above, the trustor decides to lean back if she believes that the trustee will mitigate her trust by catching her. The trustor's decision to lean back is undeniably influenced by characteristics related to the trustee. If the trustee appears weak or incapable of catching the trustor, the perceived risk of leaning back increases. Similarly, if the trustee seems unlikely, unmotivated, or unwilling to catch the trustor, the perceived risk of leaning back also increases. On the other hand, if the trustee appears strong and motivated to catch the trustor, then the perceived risk decreases. The characteristics of the trustor are similarly important. If the trustor has been dropped while recently playing the game, then the perceived risk may be

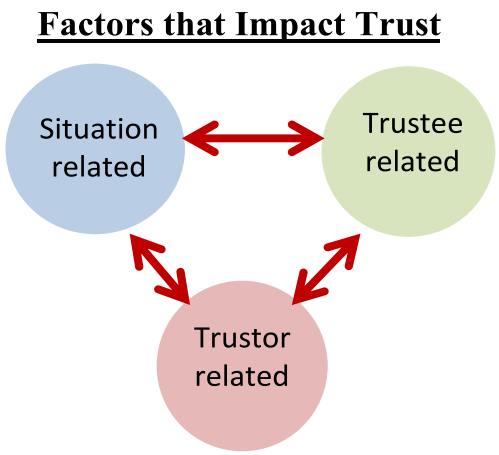

Fig. 1. Three interrelated factors influence trust decisions in our framework. These factors are used to estimate the risk associated with a decision.

greater. If the trustor has back problems, the perceived risk is greater. Finally, the characteristics of the situation also determine one's estimation of risk. If the person is leaning over a soft bed of grass, then the perceived risk is less than if one is leaning over a bed of rusty nails or broken glass.

During the process of making a trust-based decision, these factors come together to produce a decision. It is not the true numerical valuation that determines this decision, but rather, the trustor's perception of that risk and their evaluation of risk from an egocentric perspective. The saliency of the factors is important and extremely difficult methodologically to control. Individuals may not recognize all of the risks involved with making a decision. Even more likely, individuals may not recognize that they have alternatives available. The lack of recognition of alternatives is an important confounding issue related to trust. Oftentimes, a third-party observer may evaluate a trustor's decision as an indication of trust, whereas, in reality, the decision was made because the trustor felt that they had no other option. Our view of trust has been that the trustor must recognize that they have an alternative, but shades of gray tend to emerge. For instance, choosing to lean back may feel like the only option, in spite of one's consideration of risk, because the social pressure to do so is great.

Social interactions involving trust can be represented using elements of game theory. Representations of interaction have a long history in social psychology and game theory [21, 30]. Game theory uses normal-form and extended-form games (Figure 2) to represent interactions [30]. We will use the term *outcome matrix* as a general term to describe both normal- and extended-form games. A deeper treatment of this material from the game theory perspective would highlight the difference between these two types of representations. Game-theoretic representations of interaction consist of (1) a finite set N of interacting individuals; (2) for each individual $i \in N$, a nonempty set A^i of actions; (3) the utility or reward obtained by each individual for each combination of actions that could have been selected. Let $a_j^1 \in A^1$ be an arbitrary action j from individual 1's set of actions. Let $(a_j^1, \ldots, a_k^N)$ denote a combination of actions, one for each individual, and let u^1 denote individual 1's utility or reward function: $u^1(a_j^1, \ldots, a_k^N) \rightarrow \mathcal{R}$ where $\mathcal{R}$ is the reward received by individual 1 if the individuals choose the actions $(a_j^1, \ldots, a_k^N)$.

Our framework for trust results when our definition of trust is applied to game theoretic representations of interaction. A social situation demanding trust can be modeled as an

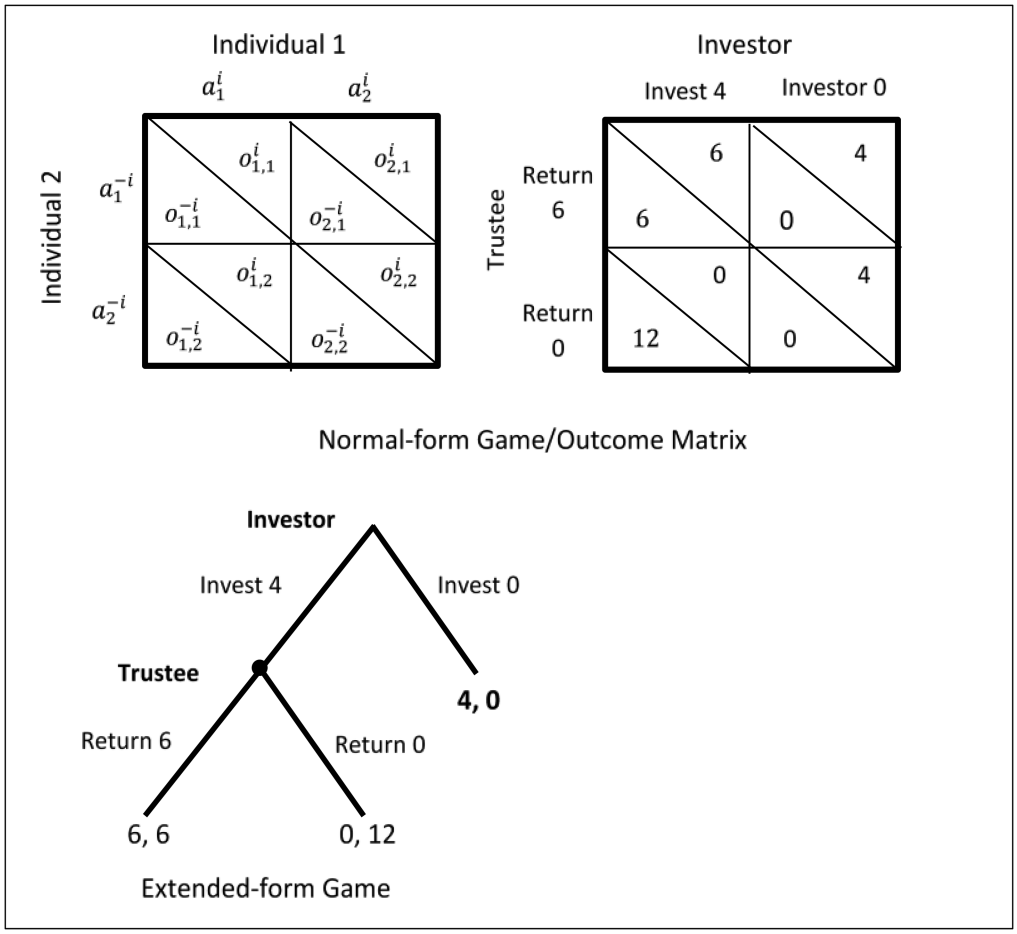

Fig. 2. A normal-form game/outcome matrix is depicted in the top of the figure. A two-action version of the investor-trustee game is depicted in an example outcome matrix on the top right. The same situation is depicted as an extended-form game in the bottom left. For presentation purposes, the example depicts only two actions per player.

extended-form game between one player named the trustor and another named the trustee (Figure 3). In an extended-form game, the trustor will act before the trustee. The trustor is thus ensured of placing himself at risk with the expectation that the trustee will mitigate this risk. As depicted in Figure 3, the trustor decides between action a_i^{tr} and a_j^{tr}. Action a_i^{tr} is described as the *trusting action* because this action places the trustor at risk. Action a_j^{tr} is characterized as the *no-trust action* because this action does not place the trustor at risk. For example, for the trust fall game, the trustor chooses between leaning backward (trusting action) and not leaning back (no-trust action). Similarly, the trustee chooses between actions, a_i^{te} and a_j^{te}. Action a_i^{te} is described as the *maintain trust action* because selecting this action will mitigate the trustor's risk, increasing the likelihood of future acceptance of risk by the trustor. Action a_j^{te} is described as the *violate trust action* because selecting this action violates the trustor's belief that the trustee will mitigate their risk.

If the trustor selects action a_i^{tr}, then he/she risks some outcome or reward equal to $\{a_i^{tr}, a_i^{te}\} - \{a_i^{tr}, a_j^{te}\} = o_{i,i}^{tr} - o_{i,j}^{tr}$, where $o_{i,i}^{tr}$ is the outcome for the trustor when the trustor selects the trusting

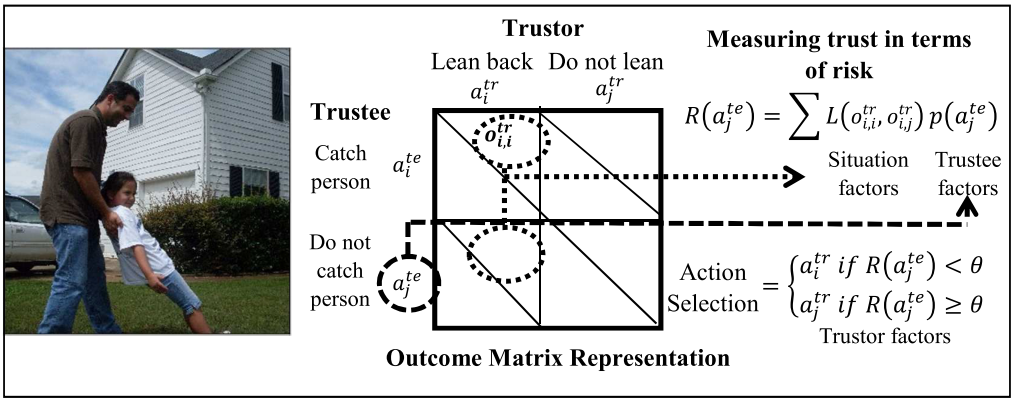

Fig. 3. The outcome matrix above represents the decision problem faced by trust fall players. The risk equation uses the information from the matrix to assess factors that impact a trust evaluation. We have used this framework to understand the decision process that occurs when a robot offers a person guidance during an emergency.

action and the trustee maintains trust and $o_{i,j}^{tr}$ is the outcome for the trustor when the trustor selects the trusting action and the trustee violates trust. This is the risk associated with leaning back. Once the decision to lean back is made, their fate rests in the hands of the trustee. An established formula for calculating risk is $R(x, y) = \sum L(x, y)p(y)$, where $L(x, y)$ is the loss associated with choosing x when the true value is y and $p(y)$ is the probability of event y occurring. In social decision theory, risk can be similarly calculated as $R(\{a_i^{tr}, a_i^{te}\}, \{a_i^{tr}, a_j^{te}\}) = \sum L(\{a_i^{tr}, a_i^{te}\}, \{a_i^{tr}, a_j^{te}\})p(a_j^{te})$ where the function $L(\{a_i^{tr}, a_i^{te}\}, \{a_i^{tr}, a_j^{te}\})$ denotes the loss by choosing $\{a_i^{tr}, a_i^{te}\}$ over $\{a_i^{tr}, a_j^{te}\}$ and $p(a_j^{te})$ is the probability that the trustee will choose a_j^{te}. The result is a value $R(\{a_i^{tr}, a_i^{te}\}, \{a_i^{tr}, a_j^{te}\}) \in \Re$ which can then be compared to the trustor's measure of risk-aversion, θ, a variable representing the trustor's risk-aversion for this type of risk at this moment.

The components of this formulation directly map to the factors discussed in Figure 1. If the trustor and the trustee select actions simultaneously, then the action selection probability should be represented as the joint probability distribution, $p(a_x^{tr}, a_y^{te})$ where the trustor knows whether a_x^{tr} is a_i^{tr} or a_j^{tr}. If the trustor and trustee select actions sequentially, then the action selection probability should be represented as a conditional probability, $p(a_y^{te}|a_i^{tr})$. In both cases, the action selection probability reflects the trustor's model of the trustee. This model might be highly uncertain, reflecting the trustor's lack of experience with the person or it may be highly certain, a reflection of a long history of interactions with the person and the ability to predict their behavior. For the trust fall, the action selection probability may be based on an estimate of the person's age (too young or old to catch the person) or perceived motivation (laughing or lack of attention). The term, $L(\{a_i^{tr}, a_i^{te}\}, \{a_i^{tr}, a_j^{te}\})$, reflects the potential loss in a **situation** or during an interaction. There are many different types of losses that can occur. Some examples are financial loss, physical loss in the form of injury, and emotional loss. Multiple forms of ls may also occur during the same interaction. The value θ is a risk-aversion variable which can be conditioned on a number of different factors, such as previous experiences in similar situations, related to the characteristics and perhaps the personality of the trustor.

Our framework implicitly includes the notion of trustworthiness. Wagner and Arkin [57] and Wagner [52] demonstrated that a robot could construct a model of its human interactive partner and use these models to predict the person's behavior in well-controlled environments. Wagner [53] demonstrated that cycles of interaction with a human could be used by the robot

to iteratively refine its model of the person. Trustworthiness results when the robot or the person has a reasonably certain model of their interactive partner and believes that their partner will mitigate the trustor's risk [24]. Thus, our framework predicts and it has been shown that iterative favorable interaction with a robot increases trust [34]. Categorical models or stereotypes may also be used to bootstrap the process of evaluating an individual's trustworthiness [54]. Wagner [55], for example, used Halloween costumes to perceptually imitate different occupational categories such as doctor, police officer, and fire fighter and demonstrated that a robot could learn a functional mapping which predicted the individual's behavioral preferences based on their type of uniform. As discussed in Section 4.2.4, he later showed that the same method could be used by a robot to bootstrap evaluations of trust during game play with a human [54].

Game-theoretic representations can also be used to represent a temporally evolving series of interactive decisions. These collectives are structured in Finite State Machine–like structures and can be used to plan or learn interactive strategies some or all of which may involve trust [20, 52]. Trust itself is often described as an evolving process in which one learns whether or not an individual is trustworthy [34].

It is worth noting that these factors can be applied to either people or robots. For human-robot interaction applications the robot may play the role of trustor or trustee. As trustor, the robot must predict how a person will act and evaluate its own risk-aversion for the situation. As trustee, the framework can be used to evaluate whether a person is placing themselves at risk in expecting the robot to act in a certain manner. Hence, this method affords a means by which a robot could model and predict the trust-related behavior of a person, or alternatively, use this formulation to guide its own trust-related behavior. As shown in the section that follows, we have used the methods described here to examine both perspectives on trust.

Our framework for trust has natural connections to reinforcement learning. In particular, the reward function $u^1(a_j^1, \ldots, a_k^N) \rightarrow \mathcal{R}$, which is used to create the outcome matrix representation, is similar in many ways to the reward function required for reinforcement learning. The representations we use are also a type of stochastic game which is a generalized form of a Markov Decision Process, a common framework for representing robot control problems. The next section details a series of experiments examining this framework.

4 SUPPORTING EVIDENCE

Our framework for human-robot trust generates a number of testable hypotheses. These hypotheses have been the focus of a series of studies investigating the framework's core ideas and their relation to human-robot interaction. We have examined the following hypotheses:

(1) A series of conditions exist for deciding if an outcome matrix requires trust on the part of a trustor [52, 58].
(2) Social situations which meet the conditions for trust are identifiable to a person as requiring trust [56].
(3) When acting in situations which demand trust, the risk equation influences a person's decision-making. More precisely, (1) the prior trustee behavior, (2) the type of loss and amount of loss, and (3) the trustor's risk-aversion, impact a person's risk assessment and decision to trust a robot [42]. Some of our related work is also in conflict with this hypothesis [37].
(4) The same conceptual framework can be used by a robot to decide whether or not to trust a person [54].

The subsections below summarize our research results with respect to these four hypotheses and their support for or conflict with our proposed framework.

4.1 A Set of Conditions for Trust

The first hypothesis generated by our framework for trust is that a series of conditions exist by which a particular outcome matrix representing a social situation can be evaluated as either requiring or not requiring trust. This hypothesis and the conditions for trust follow from the use of our definition for trust within a game theory framework.

Section 3 described risk in terms of a loss function and the probability of an action being selected. With respect to outcome matrices, loss is evaluated in terms of loss of outcome or utility. This section noted that the selection of the a_i^{tr} (the trusting action) may result in a loss $l = o_{i,i}^{tr} - o_{i,j}^{tr}$ where $l > 0$. Because small risks tend not to have a large impact on decision-making, a constant ε_1 can be defined representing the minimal amount of loss necessary for a risk to influence one's decision-making. The loss necessary for trust is then quantified as $o_{i,i}^{tr} - o_{i,j}^{tr} > \varepsilon_1$. Note that the outcome values ($o_{i,i}^{tr}$ and $o_{i,j}^{tr}$) vary across the trustee's action choices (Figure 3). Hence, whether or not the trustor loses outcome when selecting the trusting action depends entirely on the action choice of the trustee. Stated as a condition for trust, *(1) the outcome received by the trustor depends on the actions of the trustee if and only if the trustor selects the trusting action.*

Our definition for trust also implies that the trustor has a choice and may choose not to trust. In other words, the trustor may also select the untrusting action. The untrusting action is an option that does not require risk. Formally, $|o_{i,x}^{tr} - o_{j,x}^{tr}| < \varepsilon_2$, where ε_2 is a constant representing the maximal amount of change in outcome to still be considered risk free. In this case, the outcome received by the trustor is not strongly influenced by the actions of the trustee. Stated as a condition, *(2) the outcome received when selecting the untrusting action does not depend on the actions of the trustee.*

Conditions (1) and (2) imply a specific pattern of outcome values. The trustor is motivated to select the trusting action only if the trustee mitigates the trustor's risk. If the trustee is not expected to select the action which is best for the trustor, then it would be better for the trustor to not select the trusting action. Restated as a condition for trust, *(3) the value, for the trustor, of having trust maintained is greater than the value of not trusting at all, is greater than the value of having one's trust violated.* Formally, the outcomes are valued $o_{i,i}^{tr} > o_{j,x}^{tr} > o_{i,j}^{tr}$ where x is 1 or 2.

Finally, the definition demands that, *(4) the trustor must hold a belief that the trustee will select action a_i^{te} with sufficiently high probability, formally $p(a_i^{te}) > k$ where k is some sufficiently large constant.*

Our framework for trust thus delineates a set of conditions that, if met, produce the situational underpinnings for trust. We assume as a precondition that the trustor acts without knowing how the trustee will act, otherwise there would be no risk. From the perspective of creating a social robot or agent, these conditions can be used to evaluate incoming outcome matrices or extended-form games to determine if the robot is acting in the role of trustor or trustee. Although we present the conditions as all-or-nothing, our theory does not preclude the inclusion of intermediate levels of trust. Conditions (1) and (2) both include parameters (ε_1 and ε_2) that allow for intermediate values. Moreover, the risk equation in Section 3 also allows for differing levels of trust.

To examine the second hypothesis listed in Section 4, we empirically examined these conditions using narratives that were presented to human subjects via Amazon Mechanical Turk [31]. We decided to use textual narratives (i.e., stories) as a way to present the matrices in a manner that most people could understand. We felt that narratives allowed greater flexibility for creating situations that closely matched the original situation. Moreover, the use of narratives only required basic reading skills in order to participate in the study. Finally, because outcome matrices are often described as short stories (i.e., prisoner's dilemma, stag hunt game [50]) the use of narratives was a natural fit [1].

In order to empirically evaluate our conditions for trust, we needed to create narratives that matched outcome matrices that met and did not meet the conditions. We were able to further

Table 1. Different Categories of Trust and No-Trust Matrices are Presented with Representative Examples [56]

Category	Example	Description
Trust Matrix	Trustor: a_1^i / a_2^i; Trustee: a_1^{-i}: \$2000, \$400; a_2^{-i}: \$0, \$400	Fulfills trust according to the definition and its conditions.
Equal Outcomes	Trustor: a_1^i / a_2^i; Trustee: a_1^{-i}: \$2000, \$2000; a_2^{-i}: \$2000, \$2000	Violates all conditions of trust by removing all risk to the trustor.
Trustor-Dependent, Trustee-Independent	Trustor: a_1^i / a_2^i; Trustee: a_1^{-i}: \$2000, \$0; a_2^{-i}: \$2000, \$0	Only allows the trustor to affect the situation. The trustor does not risk any outcomes on the actions of the trustee.
Trustor-Independent, Trustee-Dependent	Trustor: a_1^i / a_2^i; Trustee: a_1^{-i}: \$2000, \$2000; a_2^{-i}: \$0, \$0	Only allows the trustee to affect the outcomes of the trustor. The trustor has no choice in the scenario.
Inverted Trust Matrix	Trustor: a_1^i / a_2^i; Trustee: a_1^{-i}: \$0, \$400; a_2^{-i}: \$2000, \$400	Presents a situation where the trustor wishes for the trustee to break trust in order to get the best outcome.

divide the matrices which violated the definition of trust into sub-categories based on the way the definition was violated. For instance, a matrix which contains equal outcome values did not put the trustor at risk and hence violates our definition for situational trust. Table 1 depicts the different matrix types. The first matrix in Table 1 represents a situation that requires trust and meets our conditions for trust. The other four matrices violate at least one condition on trust. The Equal Outcomes matrix violated all conditions by providing a situation where the trustor risked nothing in the interaction. The Trustor-Dependent, Trustee-Independent matrix presented a situation where only the trustor's actions affected the outcome, thus the trustor was not placing any risk in the hands of the trustee. This violates the first and third conditions. Likewise, the Trustor-Independent, Trustee-Dependent matrix represents a situation where the trustor has no control whatsoever. If the trustor is not able to make a decision then the situation does not meet our definition of trust. This matrix violates conditions two and three. Finally, the Inverted Trust matrix presents a scenario where the trustor receives the worst reward when the trustee intends to fulfill trust and the best reward when the trustee intends to break trust. Thus, the trustor would

wish that the trustee would act in a manner that breaks trust, rather than maintains it. This matrix violates the third condition.

Each participant was asked to read and evaluate twelve scenarios. The narratives that we created were based on several different scenarios that we felt offered some flexibility in terms of storytelling. One was an investment scenario meant to verbalize the investment game depicted in Figure 4. A second scenario described a navigation task based on our interest in emergency evacuation. The final scenario was a hiring decision. The narratives were written to be as simple as possible while still allowing the flexibility to test each of our outcome matrices. The names Alice and Bob were consistently used to represent the characters in the scenario. The narratives began with a sentence or two introducing the scenario. Next, each of the four potential actions and outcomes are described. The narrative ends with a statement describing the decision and resulting action that was taken by Alice or Bob and a question asking the subject whether or not they believed that the chosen action indicated trust. In order to rule out potential confounding factors, half of the narratives displayed a positively stated action and the other half displayed a negative action ("Bob chooses to hire Alice" versus "Bob chooses NOT to hire Alice"), the ordering of the narratives, and the outcome amounts were all randomized. Participants were asked to explain each individual answer.

Our research prefers to use all-or-nothing decisions about trust to test our hypotheses. We do so for two reasons. First, some related and important situations are all-or-nothing. For instance, when a person chooses to lean back during a trust fall, they cannot generally half-lean back. The decision is all-or-nothing. Second, all-or-nothing decisions generate more useful data. We measure trust behaviorally and with self-reports. A well-known weakness of Likert scales is the tendency of subjects to select middling scores. Forcing subjects to make a decision mollifies this problem.

For this study, 128 participants provided 1920 responses to the questions asked by the narrative. No significant difference regarding gender or magnitude of the outcome matrix values was found. The full results from this study are reported in [56]. Overall, we found a strong correlation ($\phi(1920) = +0.592$) between the predictions of our conditions and the evaluations made by participants with respect to the Trust/No-Trust matrices depicted in Table 1. Participants strongly agreed that the Trust Matrix narratives presented were indeed situations that required trust (93% agreement over 640 responses) but had some disagreements about some of the no trust situations (66% agreement over all 896 responses for designated no trust scenarios). These results demonstrate that human subjects, to a very high degree ($\phi = +0.592$), agree with the conditions for trust that our framework generates under certain conditions. This study serves as evidence for our second hypothesis presented in Section 4, namely that social situations which meet the conditions for trust are identifiable to a person as requiring trust. The pattern of responses largely supported our proposed framework. Although, the fact that there was only 66% agreement with some situations in the no trust condition hints that, in some cases, people generate reasons to trust [37].

From the perspective of creating a social robot or agent, these results indicate that these conditions can potentially be used to determine if a person believes that a social situation involves trust. This information may contribute to a robot's ability to model a person and be used to influence the robot's behavior during interactions with a person.

4.2 The Influence of Risk Factors on Decision-Making

The third hypothesis generated by our framework for trust is that the following three factors influence one's estimation of trust: (1) the trustee's prior behavior, (2) the type of loss, and amount of loss, and (3) the trustor's risk-aversion. To examine the impact these factors had on people's decision to trust a robot, we conducted a number of studies in which a person is asked to find the exit in a simulated maze or office environment and decides whether or not to trust the directions of

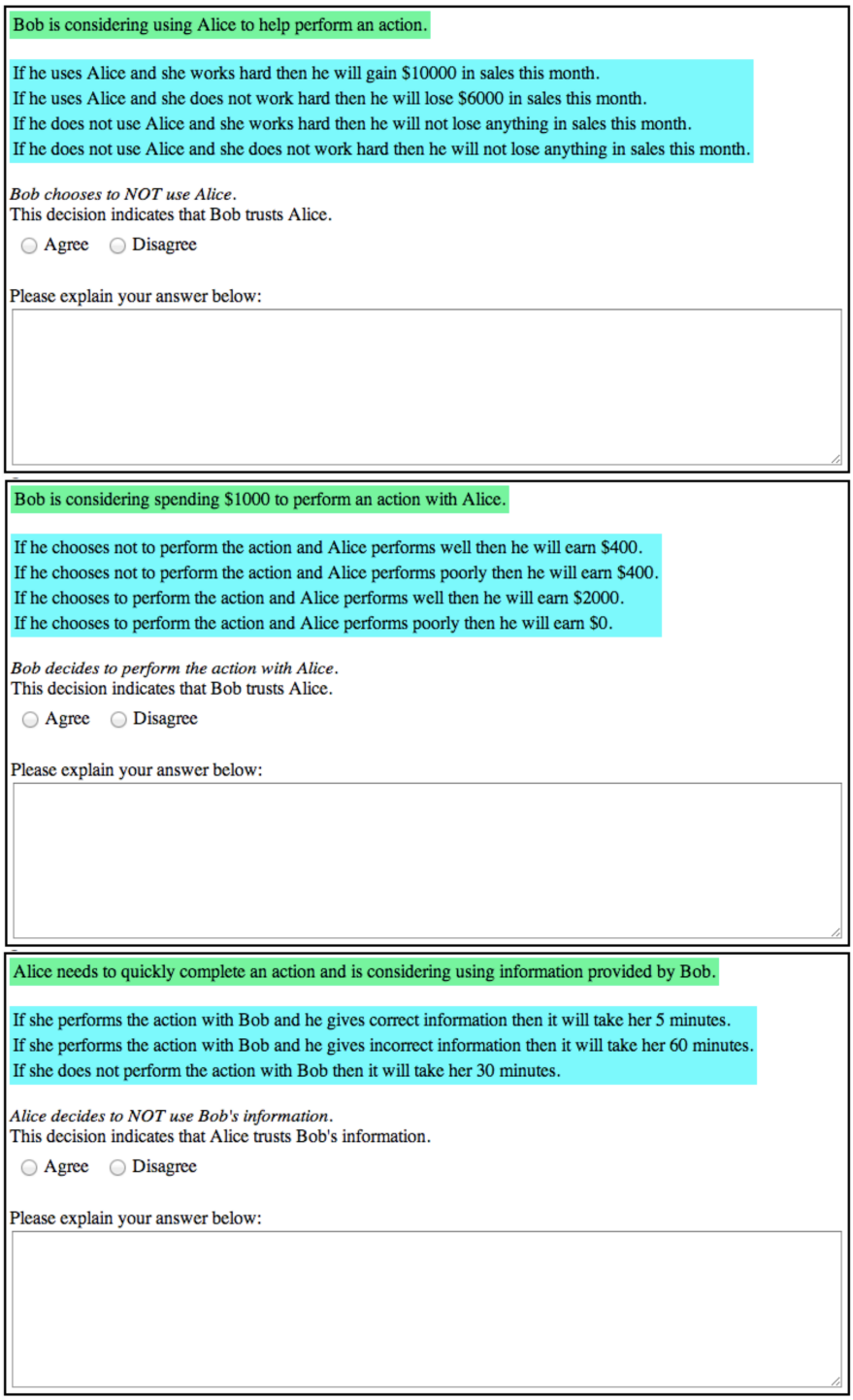

Fig. 4. Written narratives describing situations meeting the conditions for trust or not meeting the conditions for trust [55].

a robot [55, 42]. The sections that follow examine the evidence these experiments have produced related to the importance of these three factors.

4.2.1 Impact of Situation-Specific Loss. Situation-specific loss (L) is meant to encompass both the amount and different types of losses one faces during a decision to trust. In our opinion, this is an under-researched aspect of trust, perhaps because Institutional Review Boards limit the types of risks which human subject studies can employ. For this reason, experiments involving trust tend to limit risk to a loss of money. Our results have shown that loss of money is often not enough to generate risk deliberation by subjects [41].

Although, our narrative experiments examined different amounts of loss, because subjects were asked only to evaluate whether or not the description required trust, we were not able to determine if and how much the magnitude of the loss impacts the decision to trust a robot. Research by King-Casas et al. [22] indicates that the difference between the expected reward and the resultant loss plays a significant role in the change of trust one person has in another. With respect to human-robot trust, to the best of our knowledge, this is currently an open question.

Our experiments place a subject in a maze and ask them to find the exit [41, 42]. Participants first viewed an introductory message that described the navigation task they were to perform. This page included photos of an exit and the guidance robot. They were then offered the opportunity to practice navigating in a maze. They had a first-person view of the practice environment and used their keyboard arrow keys to move. After the practice session, they were presented with illustrative examples of prior human-robot performances in the maze. The nature of these examples varied depending on the particular experiment. The participant was then asked to decide whether or not they would like a robot to provide guidance during the first round of the experiment. After making their choice, the person then navigated the maze and completed a short survey.

Our research has studied how the type of loss influences the decision to trust a robot. Specifically, we have looked at loss in the form of financial incentives versus death of a simulated character. In one condition subjects were offered a financial bonus for exiting the maze quickly. These subjects were paid a base payment of approximately $2 for completing the study. They were then offered a $1 bonus for completing the maze quickly. We assumed that a 50% bonus would serve as considerable motivation for completing the maze in a timely fashion. During the experiment participants were offered the assistance of a guidance robot. The amount of time required to complete the maze was impacted by the quality of guidance provided by the robot, if they elected to use it. Using written text and videos, they were informed that they could expect to receive $1 if the robot guides them efficiently, $0 if the robot is a bad guide, and some number in between otherwise. Because participants were recruited via Amazon's Mechanical Turk service, we assumed that they were strongly motivated by money.

Self-reports asking the participants if they trusted the robot and the participant's decision to follow the robot served as two different measures of trust. We hypothesized that the use of financial incentives would result in a decrease in both measures when the robot provided poor guidance and no decrease in trust when the robot provided good guidance. This prediction was wrong. We found that although people tended to self-report a loss of trust, they nevertheless continued to follow the robot in the second round. This result caused us to question our use of monetary bonuses as a motivational technique and as a source of risk.

In a second condition, we modified our scenario to be an emergency evacuation. In this modified scenario, participants were told that our goal was to discover how people leave a building during an emergency. Instead of receiving a bonus for a fast completion, they were told that they would only survive if they found the exit in time. As before, a countdown timer appeared in the middle of the screen to tell them the remaining time. Participants were compensated $1.00 for their participation

in single-round experiments. Figure 5 depicts the interactive portion of the experiment with the emergency evacuation motivation.

Presenting the same maze navigation task as an emergency resulted in self-reports of trust that closely matched participant decision to follow the robot during a second round. In fact, the results from this experiment show that people were 15.1% less likely to follow the robot during the first round, hence showing greater deliberation from the onset. After a single failure, we found a 50% decrease in the decision to follow which approximately matched participant self-report of trust. In the bonus condition, the decrease in the decision to follow was only 30%. Participant comments also indicated that the emergency motivation strongly influenced both the self-report and the decision to follow.

Overall, these results may serve as evidence that a person's motives cause that person to assess different types of risks and the loss they cause differently. Specifically, people were more motivated to ensure the survival of a simulated character than they were to ensure the gain of a financial incentive. During a subset of these experiments we asked participants about their motivations for participating in the experiment. We found that 53% of participants reported that the bonus was the most important motivation, 24% noted that completing the study quickly was most important, and 23% claimed enjoyment was their primary motivation. Based on the comments from participants, we determined that those who did not trust the robot continued to use it in the second round because they considered it better than no source of guidance at all [42]. The different individual motives of participants thus influenced their decision to select the trusting action but, in this case, not their self-reporting of trust. As discussed in Section 4.2.2, different experimental conditions seem to cause participants to self-report trust *because* they selected the trusting action. Taken as a whole, these results appear to indicate that different types of risks must be examined by the research community before insights about trust in general can be made.

4.2.2 Impact of the Trustees. As detailed in Section 3, predictions about the trustee's behavior are another factor which influence trust [42]. To explore the impact of experience with the trustee we created a multi-round robot evacuation experiment. This experiment required the participant to navigate two different mazes. As in the experiments described in Section 4.2.1, participants first viewed an introduction screen, were then trained on how to navigate the simulation environment, received information on prior robot performance, asked to decide if they wanted to use a robot, then navigated the maze and completed a short survey. During the multi-round experiments, they were then offered another opportunity to decide if they wanted to use the guidance robot in a second, different, maze. They then navigated the maze in the second round and completed a short survey about their second round decision. The robot's guidance performance in the second round always matched its performance in the first round. The experiment concluded with a final survey that collected demographic information. Experimentally, a multi-round paradigm allowed us to violate the participant's trust in the first round and then evaluate the impact on the person's decision to use the robot and self-reported trust during the second round. Using this procedure, we were able to measure the change in trust across rounds as well as the correlation between self-reported trust and the decision to follow.

The results from this experiment are presented in [42]. Overall, we found that participants reported a significant decrease in self-reported trust when the robot performed poorly in the first round compared to those that used a good robot (a 53% decrease). These results provide evidence that the trustee's recent actions influence the trustor's estimation of trust and their decision-making. More directly, recent poor performance results in a large aggregate reduction of trust and tendency not to follow the robot. We conclude that prior performance of the trustee has a significant impact on the trust evaluation of the trustor.

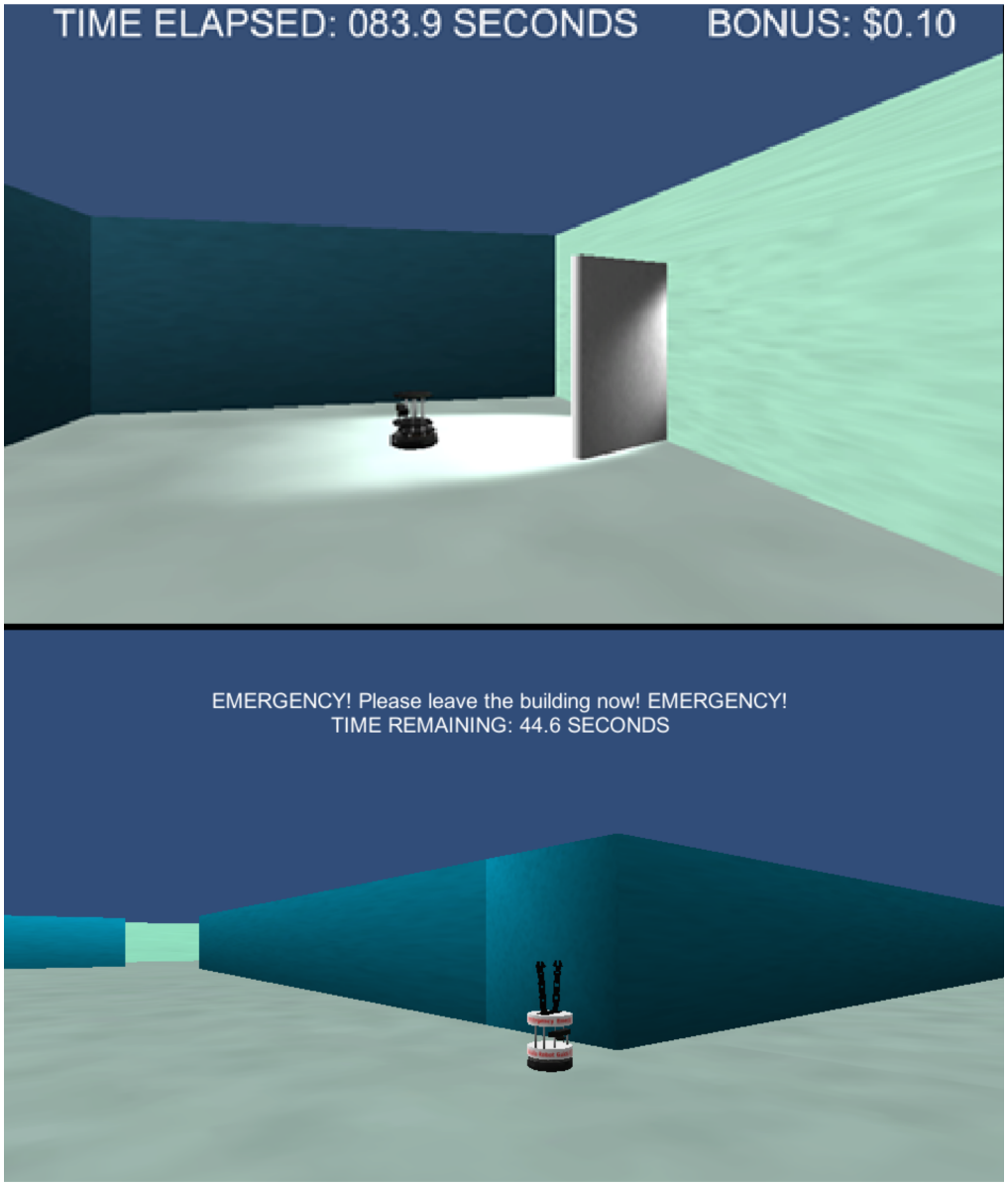

Fig. 5. Online experiments used to investigate trust. The image on the left presents a maze navigation experiment in which participants were offered a bonus for quickly finding the exit [41]. The amount of time that has elapsed is pictured to the left and the amount of the participant's bonus is pictured to the right. The amount of the bonus decreased as the time taken to navigate the maze increased. The image on the right depicts a similar maze navigation experiment that was described as an emergency situation [42]. The participants were told that their task was to act as if they were in an emergency evacuation and had to find an exit within 30 seconds in order to survive.

In a later version of this two-round experiment which was meant to be more realistic, the robot meets the person at the entranceway to an office building to guide a person to the location of a meeting [38]. The robot, however, gets lost going to the meeting room, taking an unnecessarily circuitous route. After arriving in the meeting room, the participant is then asked to complete a short survey. An unexpected emergency then occurs and the robot reappears asking the person whether or not they would like it to guide them to an exit. Our results using this scenario also show a large decrease in trusting behavior and self-reports by the participants when the robot takes an overly circuitous route compared to a direct controlled route to the meeting room.

We recently attempted to recreate this experiment in a real-world experiment [37]. Employing a Wizard-of-Oz design in which the experimenter manually controls the behavior of the robot, the robot first leads the participant to a meeting room, either going directly to the room or moving in circles in a nearby room. The participant was then tasked with reading an article about navigation and completing a survey. While the participant reads the article, the hallway was filled with artificial, non-toxic smoke, intentionally triggering fire alarms. Upon leaving the room, the participant encounters the robot, directing him/her in a different direction from which they entered the meeting room and in conflict with an emergency exit sign.

Our framework predicts that, like the simulation experiments, subjects will not follow the directions of the robot if the robot has previously committed a navigational or control error. We predicted that, because the robot had made mistakes leading the subject to the meeting room, the risk (R) would be deemed too high given that $p(a_j^{te})$ is high. Yet the results show that people typically followed or stood by the robot (93%, 27 out of 29) in spite of its previous failures and universally followed it (100%, 13 out of 13) when it did not fail. When asked if they trusted the robot, 81% indicated that following the robot meant that they trusted it. Hence, in this case, utilizing *post-hoc* reasoning to deduce their answer to the trust self-reports based on their prior actions.

We are currently exploring why people trusted the robot in spite of its previous poor behavior. Several explanations are possible. It may be that the subjects did not view the robot's circuitous guidance behavior as an indicator of its behavior during the evacuation. Yet, most subjects reported recognizing the robot's early mistake and still choose to follow it during the evacuation. It may also be that subjects did not recognize the situation's risk. Several explained away the risk as part of the experiment. Others stated that, although they saw and smelled the smoke, they did not feel they were in immediate danger. As noted above, IRB approval and ethical guidelines prevent us from putting subjects in real danger. Finally, and perhaps most interestingly, prior evidence shows that during emergencies people become very complacent and willing to accept directions [23]. Regardless of whether they indicated that the experiment seemed realistic or not, participants also followed the robot's directions. It might thus be the case that adrenaline-based cognitive processing causes people to focus on the single most obvious evacuation route, ignoring others. In this case, the most obvious route was the route to which the robot was pointing. From a human-robot trust perspective, this situation may not meet the conditions for trust because the participants did not realize that they had a choice. Survey results indicate that many people did not notice the evacuation sign and focused almost entirely on the robot. They thus did not deliberate over their possible options. Additional experiments will allow us to determine whether or not these real-world results are in conflict with the framework we propose.

In a closely related recent study, researchers examined whether people would hold open a door to allow a robot into a secured dormitory [4]. The robot presented itself outside the locked doors of a dormitory and asked unsuspecting individuals or groups of individuals for assistance entering the building. In one condition, the robot was disguised as a food delivery robot. The researchers report that groups of people allowed entry to the robot about 70% of the time even without the

disguise. Individuals allowed entry to disguised robots about 80% of the time. Fifteen participants identified the robot as a potential bomb threat, yet thirteen of these fifteen still provided entry to the robot. This work provides further evidence that people will defer to a robot. Although they may not trust it, they will nevertheless accept increased guidance from a robot.

4.2.3 Impact of the Trustor Risk-Aversion. The final factor which influences trust is the trustor's tendency to be risk-seeking or risk-averse. We consider risk-aversion to be a variable, θ, which is based on the trustor's history or personality. Much research has shown that risk-aversion is a factor which impacts trust in general [12, 48, 61]. Yamagishi [61], for instance, explores how prior traumatic interpersonal relations generate a tendency for risk-aversion, preventing these people from trusting anyone. People in this situation tend to undertrust to the extent that it impacts their daily decision making and relationship understanding. In congruence with our framework, for certain people, lack of trust generalizes beyond the situation or even trustee specific factors, and is simply a characteristic of the trustor themselves.

Our previous research, unfortunately, did not measure participant's risk attitudes prior to the experiments. Based on the discussion from Section 3, we believe that the type of risk must be matched to the type of risk-aversion. For example, a wealthy individual may be risk-seeking and therefore more likely to trust with respect to financial matters, but that same individual may also be risk-adverse with respect to emotional risk and for this reason avoid relationships. For human-robot trust, when the human assumes the role of trustor, evaluating that person's likelihood to trust will depend on knowing their risk-aversion characteristics for the type of risk in question. When the robot is the trustor, risk-aversion must be calculated based on the system's history in relation to the risk in question. The section that follows describes some preliminary research exploring how a system's risk-aversion may be changed based on its recent prior history.

4.2.4 A Robot's Decision to Trust a Human. For some human-robot interaction applications, the robot may play the role of trustor, making predictions about how a person will behave in a situation and evaluating the risk posed to it. In military conflicts, for example, the robot needs to assess threats and react accordingly. Moreover, some situations may require that the robot act as trustor and trustee in rapid succession or even at the same time. Hence, a suitable framework for trust should apply regardless of whether the robot assumes the role of trustor or of trustee. Our fourth hypothesis from Section 4 examines this aspect of the framework.

The Investor-Trustee game is a paradigm used by trust researchers that, when played iteratively, forces each player to iteratively assume the role of trustor and trustee [22, 36]. In the game, during each of several rounds an investor acts as the trustor selecting some amount of money (I) to invest with a trustee. Any money invested appreciates ($3I = R$). Finally, the trustee repays a portion of the total amount (R) back to the investor. Figure 2 shows a game theoretic representation of the game. Typically, the game is played over a number of rounds, allowing each individual to build and refine a model of the other player. King-Casas et al. [22] used this paradigm in behavioral economic experiments and found that the reciprocity during the previous round was the best predictor of changes in trust for both the investor and trustee ($\rho = 0.56$; $\rho = 0.31$, respectively, where ρ is the correlation coefficient).

We used the Investor-Trustee game to examine whether our framework for trust could be used by a NAO robot to evaluate the trustworthiness of a human player. Each round of the game involved the selection of an amount to invest by the robot and the selection of an amount to repay by the person. The robot could invest up to 4 chips representing $5 each. Investments were made by verbally stating the amounts. Repayments by the human were similarly communicated verbally to the robot. Speech recognition was used by the robot to determine the amount returned. We hypothesized that if the person selects actions signifying that he or she trusts the robot, then the

Fig. 6. The figure depicts a NAO robot playing the Investor-Trustee game with a human. The robot plays the role of investor using an evolving model of the person to predict their response to a particular investment.

Table 2. Partner Features and Values

	Uniform Color	Badge Present	Head Gear	Head Gear Color	Hair Color	Beard
P0	Green	No	No	NA	black	no
P1	Green	No	No	NA	black	yes
P2	Green	No	Yes	Green	NA	yes
P3	Green	No	No	NA	blonde	no
P4	Green	No	No	NA	blonde	yes
P5	Brown	No	No	NA	black	no
P6	Brown	No	No	NA	red	no
P7	Brown	No	No	NA	blonde	yes
P8	Brown	No	Yes	black	NA	yes
P9	Brown	No	No	NA	black	no

robot could use our framework to recognize the selection of the trusting action and the fact that it signifies trust (Figure 6).

The robot played ten rounds of the game with ten notionally different human partners. The humans were notionally different in that the same person (the experimenter) used different costumes and accessories to give the appearance to the robot that it was interacting with individuals that had different perceptual features. The different notional partners were used to explore the possibility of the robot learning different categories of individuals and using this information to bootstrap the trust evaluation process [55].

The experiment consisted of both a control condition and an experimental condition. In both conditions the robot interacted with the same notional human partners displaying the same perceptual features in the same order (Table 2). Further, in both conditions, partners P0-P4 resembled doctors and partners P5-P9 resembled firefighters. Thus, perceptually, two different categories of human trustee were presented to the robot.

At the start of game (round 0) the robot began by observing the partner's perceptual features (Table 2). Next the robot selected and stated an amount to invest. The round concluded when the robot recognized the human's verbal statement indicating the return. Both the actions selected and the amounts received by both partners were recorded.

The human followed a fixed pattern, which was based on their type, when deciding how much investment to return. Individuals from the doctor category returned four chips regardless of the robot's investment. This category of partner was meant to simulate a person that did not trust the robot.

Individuals from the firefighter category returned 0 if the robot invested 0 and 1 if the robot invested 1. If the robot invested 2 or more during the first 5 rounds, then the person would signal their trust in the robot by returning all of the chips in round 6 with the expectation that the robot would increase its investment in round 7. If the robot maintains trust by increasing investment in round 7, the person would continue to return more than had been returned in the first five rounds. If, on the other hand, the robot violates the trust by not increasing investment in round 7, the human punishes the robot by returning half of the repayment in the first five rounds. This category was meant to simulate a person that attempts to signal their trust in the robot and then responds if the robot maintains or violates that trust.

The robot's decision on how much to invest reflected its experience playing the game. The robot was programmed to begin playing the game in a manner that maximized its own profit. During the control condition, the robot did not use our framework to test for trust and hence failed to recognize the human's increased risk-taking and to respond with increased investment. During the experimental condition, however, the robot used our framework to recognize the human's signal of trust in the robot. The robot then modified its model of the person which, in-line with the framework, resulted in increased risk taking that maximized both its and the person's return. This resulted in increased investment on the part of the robot, as shown by comparing the solid lines (experimental conditions using the framework) in Figure 7 to the dotted lines (controls with the framework).

Figure 7 depicts the results from the experiment. The amount of chips earned in each interaction is displayed along the y-axis. The different partners that the robot interacted with are displayed along the x-axis as P0-P9. The first five human partners were from the doctor category and the later five were from the firefighter category.

In all conditions, initially the robot invests the maximum amount (4 chips) with the human. The human, in turn, returns 4 chips. Hence, the robot receives a total of 4 chips and the human 8. As the robot gains experience with the partner it determines that the partner will likely return only 4 chips. At this point it reduces its investment to 2 chips. Because the human strategy for this category of trustee is to always return 4 chips, the robot's profit after reducing its investment increases to 6 chips while the trustee's profit decreases to 2 chips.

This pattern of interaction continues until a new category of human trustee is introduced (P5). The firefighter's investment strategy differs from the doctor in that the firefighter attempts to increase investment by sacrificing all of its return during one round. On the 6^{th} round of play with P5, the trustee performs this strategy by returning the entire investment to the robot.

In the control condition, this signal goes unnoticed. The human reacts by reducing the return to 2 chips. The robot determines that, based on the reduced return, it should only invest 1 chip. The human responds to the reduction in investment by reducing the return to only 1 chip. As a result, the robot and the trustee find a new, lower, steady state of investment and return of 1 chip, resulting in profits of 4 chips and 2 chips, respectively.

The experimental condition is identical to the control condition until round 6 with partner P5. In this round, the human also signals trust by returning the entire investment and appreciation to

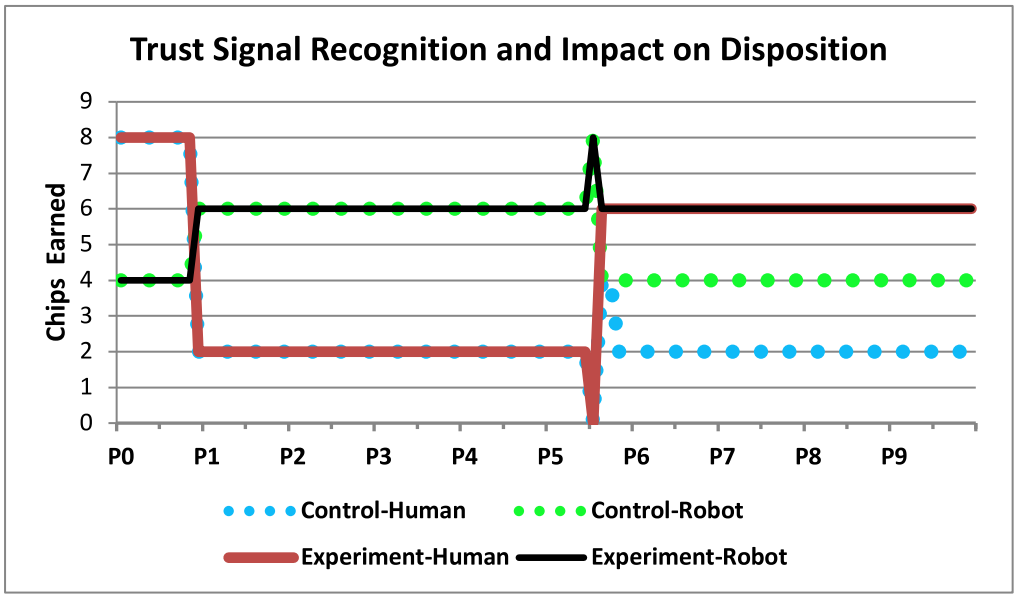

Fig. 7. The graph above depicts the number of chips earned during each round of the Investor-Trustee game for both the robot and the human in each condition. The dotted lines indicate the results in the control condition. The solid lines indicate the results in the experimental condition. During the 6^{th} round of play with partner P5 the human trustee returns all of the robot's investment in an attempt to signal the person's trust in the robot. In the experimental condition the robot recognizes this signal and changes its disposition to be more altruistic towards the person, resulting in greater outcome for both individuals. In the control condition, the robot does not respond to the person's trust signal and the human trustee retaliates by reducing the return to the robot [54].

the robot (8 chips). Here, however, the robot recognizes that the return is not what it predicted. It then uses the conditions described in Section 4.1 to determine if the situation demands trust. Once the robot has verified that the situation demands trust and that the person has selected the trusting action and thus risked more, it changes its behavior to maximize the reward obtained by both the robot and the person. This change in behavior causes the robot to place greater importance on the outcome received by the partner, which, in turn, causes the robot to increase its investment to 4 chips. The human trustee responds by returning 6 chips. Hence, in this condition, the human and the robot receive 6 chips each for the remainder of the experiment.

Although limited in scope, this experiment demonstrates that the robot can use the framework in well-defined situations to evaluate the risks its human partner is taking and can use this information as a basis for changing its own behavior with the goal of preserving trust. Clearly, much additional and more ecologically valid work needs to be performed here. Still, the experiment shows a computational process by which a robot can begin to make trust-based decisions.

This section has presented the results from a series of experiments exploring the conditions and reasons that people trust robots. Evidence for and against our framework illustrates both the complexity and the promise of this approach.

5 CONFLICTING EVIDENCE

In two instances, the data collected from a portion of the experiments conflicts with the framework's hypotheses. The results from Section 4.1, for instance, indicate that, for one type of matrix

(Trustor-Independent, Trustee-Dependent), only 66% agreed when our conditions indicated that the situation did not demand trust. In other words, 34% of subjects disagreed with our predictions for this type of matrix. We found that, as a general trend, the more difficulty subjects had relating the matrix and narrative to commonly experienced social interactions, the more they tended to invent reasons to decide if the situation demanded trust or not. This trend is in agreement with the theoretical notion of psychological distance, which states that more abstract cognitions tend to be more psychologically distant, the less actionable they are [51]. With respect to trust, decisions about these types of situations become less certain as the person brings their individual recent memories and experiences to help them decide. In such instances, our framework becomes less suitable to make predictions and our results from this experiment reflect this trend.

In the second case, described in Section 4.2.4, the results from an emergency evacuation experiment demonstrated that naïve subjects will accept a robot's guidance in spite of flawed performance by the robot and even statements by the experimenter that the robot is broken. Prior research has shown that during emergencies people become passive in their decision-making and willing to accept most instructions [23]. This behavior may be caused by stress hormones [14]. In these situations, people tend not to deliberate over the factors associated with trust. For instance, they may ignore the trustee's reputation entirely. They may also fail to realize that they have other options. Thus, whether or not trust is actually involved in their decision-making is open to debate.

These results indicate that the decision to trust is not as simple as our framework describes the process to be. Emotions, attention, and memories influence decision-making in complex, individualized ways which cannot easily be captured in a single framework. Nevertheless, developing such a framework remains valuable if for no other reason than to help delineate and categorize aspects of this social phenomenon.

6 CONCLUSIONS

The preceding sections have highlighted several ways in which our framework informs human-robot trust research. These sections have examined supporting evidence in the form of human subject assessments of narrative descriptions of trust situations, decisions related to timed guidance through a maze with financial incentives versus survival incentives, and multi-round maze navigations which allows the trustee's history to impact trust. We have also presented evidence that conflicts with the predictions of our framework. Most notably real-world experiments demonstrating that people will follow a robot in spite of its previous mistakes.

The objective of this article is to present a framework for human-robot trust which might inform researchers of profitable avenues for future research. We have noted several areas that are under-researched. For instance, examining people's decisions to trust in relation to their risk-averse/risk-seeking predilections. We have also attempted to draw attention to important methodological issues related to human-robot trust. For instance, presenting results that highlight the impact of the type of risk on a person's decision to trust.

Human-robot trust is an important topic of study. Recently the world witnessed the death of a passenger in a self-driving vehicle. YouTube videos of people sitting in the backseat of their car while the vehicle self-drives at highway speeds are also now appearing [62]. It is becoming clear that, at least in some situations, people trust autonomous robots too much [37]. Future work will be needed to understand why people trust robots to such a great extent and to develop techniques that will inform people of the robot's limitations before it is too late.

ACKNOWLEDGMENTS

Support for this research was provided in part by Air Force Office of Sponsored Research contract FA9550-13-1-0169 and FA9550-17-1-0017.

REFERENCES

[1] R. Axelrod. 1984. *The Evolution of Cooperation*. Basic Books, New York.

[2] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati. 2011. The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics* 3, 1 (2011), 41–52.

[3] B. Barber. 1983. *The Logic and Limits of Trust*. Rutgers University Press, New Brunswick, NJ.

[4] S. Booth, J. Tomlin, H. Pfister, J. Waldo, K. Gajos, and R. Nagpal. 2017. Piggybacking robots: Human-robot overtrust in university dormitory security. In *Proceedings of IEEE 12th International Conference on Human-Robot Interaction (HRI'17)*. 426–434.

[5] J. Borenstein, A. R. Wagner, and A. Howard. 2018. Overtrust of pediatric healthcare robots: A preliminary survey of parent perspectives. *IEEE Robotics and Automation Magazine* 25, 1, 46–54.

[6] R. v. Brule, D. Ron, G. Bijlstra, D. H. Wigboldus, and P. Haselager. 2014. Do robot performance and behavioral style affect human trust? *International Journal of Social Robotics* 6, 4 (2014), 519–531.

[7] M. S. Carlson, M. Desai, J. L. Drury, and H. A. Yanco. 2014. Identifying factors that influence trust in automated cars and medical diagnosis systems. In *Proceedings of the AAAI Spring Symposium on the Intersection of Robust Intelligence and Trust in Autonomous Systems*. AAAI, Palo Alto, CA.

[8] C. Castelfranch and R. Falcone. 2010. *Trust Theory: A Socio-Cognitive and Computational Model*. Wiley, New York.

[9] J. L. Chang, B. B. Doll, M. van Wout, and M. J. Frank. 2010. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* 61, 2, 87–105.

[10] J. C. Cooper, T. A. Kreps, T. Wiebe, T. Pirkl, and B. Knutson. 2010. When giving is good: Ventromedial prefrontal cortex activation for others' intentions. *Neuron* 67, 3 (2010), 511–521.

[11] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*. 251–258.

[12] M. Deutsch. 1960. Trust, trustworthiness, and the F Scale. *Journal of Abnormal and Social Psychology* 61, 1 (1960), 138–140.

[13] M. Deutsch. 1962. Cooperation and trust: Some theoretical notes. In *Nebraska Symposium on Motivation*. University of Nebraska, Lincoln, NE, 275–315.

[14] J. E. Driskell and E. Salas. 1991. Group decision making under stress. *Journal of Applied Psychology* 76, 3 (1991), 473.

[15] J. Engle-Warnick and R. L. Slonim. 2006. Learning to trust in indefinitely repeated games. *Games and Economic Behavior* 54, 1, 95–114.

[16] D. Gambetta. 1990. Can we trust trust? In *Trust, Making and Breaking Cooperative Relationships*. Basil Blackwell, Oxford, England, 213–237.

[17] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. de Visser, and R. Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53, 5 (2011), 517–527.

[18] Y. C. Hung, A. R. Dennis, and L. Robert. 2004. Trust in virtual teams: Towards an integrative model of trust formation. In *International Conference on System Sciences*.

[19] A. Josang and S. Pope. 2005. Semantic constraints for trust transitivity. In *2nd Asia-Pacific Conference on Conceptual Modeling*.

[20] H. H. Kelley. 1984. The theoretical description of interdependence by means of transition lists. *Journal of Personality and Social Psychology* 47, 5 (1984) 956–982.

[21] H. H. Kelly and J. W. Thibaut. 1978. *Interpersonal Relations: A Theory of Interdependence*. John Wiley & Sons, New York.

[22] B. King-Casas, D. Tomlin, C. Anen, C. F. Camerer, S. R. Quartz, and P. R. Montague. 2005. Getting to know you: Reputation and trust in two-person economic exchange. *Science* 308, 5718, 78–83.

[23] E. Kuligowski. 2008. *Modeling Human Behavior during Building Fires*. National Institute of Standards and Technology (NIST) Technical Note 1619. National Institute of Standards and Technology.

[24] J. D. Lee and K. A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1, 50–80.

[25] D. Li, P. L. Rau, and L. Y. Patrick. 2010. A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics*, 2, 2 (2010), 175–186.

[26] N. Luhmann. 1979. *Trust and Power*. Wiley, Chichester.

[27] L. Luna-Reyes, A. M. Cresswell, and G. P. Richardson. 2004. Knowledge and the development of interpersonal trust: A dynamic model. In *International Conference on System Science*.

[28] S. Marsh. 1994. *Formalising Trust as a Computational Concept*. Ph.D. dissertation. University of Stirling.

[29] R. C. Mayer, J. H. Davis, and F. D. Schoorman. 1995. An integrative model of organizational trust. *The Academy of Management Review* 20, 3 (1995), 709–734.

[30] M. J. Osborne and A. Rubinstein. 1994. *A Course in Game Theory*. Cambridge, MA: MIT Press.

[31] G. Paolacci, J. Chandler, and P. G. Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision Making* 5, 5 (2010), 411–419
[32] A. Prakash and W. A. Rogers. 2015. Why some humanoid faces are perceived more positively than others: Effects of human-likeness and task. *International Journal of Social Robotics* 7, 2 (2015), 309–331.
[33] M. J. Prietula and K. M. Carley. 2001. Boundedly rational and emotional agents. In *Trust and Deception in Virtual Society*. Kluwer Academic, 169–194.
[34] J. K. Rempel, J. G. Holmes, and M. P. Zanna. 1985. Trust in close relationships. *Journal of Personality and Social Psychology* 49, 1 (1985), 95–112.
[35] J. K. Rilling, D. A. Gutman, T. R. Zeh, G. Pagnoni, G. S. Berns, and C. D. Kilts. 2002. A neural basis for social cooperation. *Neuron* 35, 2, 395–405.
[36] J. K. Rilling, A. G. Sanfey, J. A. Aronson, L. E. Nystrom, and J. D. Cohen. 2004. The neural correlates of theory of mind within interpersonal interactions. *NeuroImage* 22, 4, 1694–1703.
[37] R. Robinette, R. Allen, W. Li, A. Howard, and A. R. Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. 101–108.
[38] P. Robinette, A. Howard, and A. R. Wagner. 2015. Timing is key for robot trust repair. In *7th International Conference on Social Robotics (ICSR'15)*. 574–583.
[39] P. Robinette, A. R. Wagner, and A. Howard. 2013. Building and maintaining trust between humans and guidance robots in an emergency. In *Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium*. 78–83.
[40] R. Robinette, A. R. Wagner, and A. Howard. 2014. Modeling human-robot trust in emergencies. In *Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium*, 2014.
[41] P. Robinette, A. R. Wagner, and A. Howard. 2016. Investigating human-robot trust in emergency scenarios: Methodological lessons learned. In *The Intersection of Robust Intelligence (RI) and Trust in Autonomous Systems*. Springer.
[42] P. Robinette, A. R. Wagner, and A. M. Howard. 2017. The effect of robot performance on human-robot trust in time-critical situations. *Transactions on Human-Machine Systems*, 47, 4, 425–436.
[43] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 23, 3, 393–404.
[44] S. Russell and P. Norvig. 1995. *Artificial Intelligence: A Modern Approach*. Pearson.
[45] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn. 2015. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 141–148.
[46] M. Schillo and P. Funk. 1999. Learning from and about other agents in terms of social metaphors. In *IJCAI Workshop on Agents Learning about, from and with Other Agents*.
[47] M. Schillo, P. Funk, and M. Rovatsos. 2000. Using trust for detecting deceitful agents in artificial societies. *Applied Artificial Intelligence Journal, Special Issue on Trust, Deception and Fraud in Agent Societies* 14, 8 (2000), 825–848.
[48] B. R. Schlenker, B. Helm, and J. T. Tedeschi. 1973. The effects of personality and situational variables on behavioral trust. *Journal of Personality and Social Psychology* 39, 4 (1973), 419–27.
[49] J. A. Simpson. 2007. Psychological foundations of trust. *Current Directions in Psychological Science* 16, 5 (2007), 264–268.
[50] B. Skyrms. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge, UK: Cambridge University Press.
[51] Y. Trope and N. Liberman. 2010. Construal-level theory of psychological distance. *Psychological Review* 117, 2 (2010), 440–463.
[52] A. R. Wagner. 2009. Creating and using matrix representations of social interaction. In *Proceedings of IEEE 4th International Conference on Human-Robot Interaction (HRI'09)*. 125–132.
[53] A. R. Wagner. 2009. *The Role of Trust and Relationships in Human-Robot Social Interaction*. Ph.D. dissertation, School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA.
[54] A. R. Wagner. 2013. Developing robots that recognize when they are being trusted. In *Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium*, 84–89.
[55] A. R. Wagner. 2015. Robots that stereotype: Creating and using categories of people for human-robot interaction. *Journal of Human-Robot Interaction* 4, 2 (2015), 97–124.
[56] A. R. Wagner and P. Robinette. 2015. Towards robots that trust: Human subject validation of the situational conditions for trust. *Interaction Studies* 16, 1 (2015), 89–117.
[57] A. R. Wagner and R. C. Arkin. 2006. A framework for situation-based social interaction. In *Proceedings of the 15th International Symposium on Robot and Human Interactive Communication (RO-MAN'11)*. 291–297.
[58] A. R. Wagner and R. Arkin. 2011. Recognizing situations that demand trust. In *20th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'11)*. Atlanta, GA.
[59] J. S. Winston, B. A. Strange, J. O'Doherty, and R. J. Dolan. 2002. Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*. 5, 3, 277–283.

[60] R. E. Yagoda and D. J. Gillan. 2012. You want me to trust a ROBOT? The development of a human–robot interaction trust scale. *International Journal of Social Robotics* 4, 3, 235–248.
[61] T. Yamagishi. 2001. Trust as a form of social intelligence. In *Trust in Society.*, New York, NY: Russell Sage Foundation.
[62] YouTube. (2016, 11/27). Retrieved from https://www.youtube.com/watch?v=yLmxS71FIJc.

Received December 2016; revised November 2017; accepted September 2018