Towards Robots that Trust: Human Subject Validation of the Situational Conditions for Trust

Alan R. Wagner
Georgia Tech Research Institute
250 14th Street NW
Atlanta, GA 30332-0822
alan.wagner@gtri.gatech.edu

Paul Robinette
Georgia Tech Research Institute
250 14th Street NW
Atlanta, GA 30332-0822
probinette3@gatech.edu

Abstract

This article investigates the challenge of developing a robot capable of determining if a social situation demands trust. Solving this challenge may allow a robot to react when a person over or under trusts the system. Prior work in this area has focused on understanding the factors that influence a person's trust of a robot (Hancock, et al., 2011). In contrast, by using game-theoretic representations to frame the problem, we are able to develop a set of conditions for determining if an interactive situation demands trust. In two separate experiments, human subjects were asked to evaluate either written narratives or mazes in terms of whether or not they require trust. The results indicate a $\phi_1 = +0.592$ and $\phi_2 = +0.406$ correlation respectively between the subjects' evaluations and the condition's predictions. This is a strong correlation for a study involving human subjects.

Keywords: trust, game theory, autonomous system, social robot, human-robot interaction.

1 Introduction

Trust underpins many interpersonal interactions. It allows a person to act in a manner that puts them at considerable risk, believing that the actions of another individual will mitigate that risk (Yamagishi, 2001). Hence, trust plays an important role in one's most critical social decisions. We contend that the same is true for a social robot. For a robot interacting with humans, an understanding of 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 resulted from a robot's actions (Economist, 2006). Moreover, people may come to place too much trust in the robots around them. A social robot should be able to recognize such a situation and act to dissuade the person from placing themselves at risk. For these reasons, it is critical to develop a formal, principled conceptualization of trust that is implementable on a robot.

Most research in this area has focused on a different problem: understanding the factors that influence a person's trust in a robot (Desai, Kaniarasu, Medvedev, Steinfeld, & Yanco, 2013; Hancock, et al., 2011). This article, in contrast, investigates the challenge of determining if a social situation demands trust. Our overarching goal is to create robots that use trust to influence their interactions with a person. Recently we introduced a formal, principled conceptualization of trust that allowed us to derive conditions that indicate if a social situation demands trust (Wagner, 2009). In this paper, we present experimental evidence connecting our conditions for trust to the view of trust held by people.

It is important that such a connection is made. In order for a conceptualization of trust to be useful it must relate to the judgments of people. If a strong relationship is found between the conditions we are investigating and a person's evaluation of trust then our concept of trust could serve as a powerful tool for predicting a person's behavior or for guiding a robot's social behavior. It could also allow the robot to reason about whether or not it is trusting or should trust a person or other robot. Understanding if and how our concept of trust relates to people may also provide considerable insight into the phenomenon of trust itself.

Many conceptualizations of trust exist. In their review of computational trust and reputation models Sabater and Sierra state, "... current (trust and reputation) models are focused on specific scenarios with very delimited tasks to be performed by the agents" and "A plethora of computational trust and reputation models have appeared in the last years, each one with its own characteristics and using different technical solutions" (Sabater & Sierra, 2005). A critical aspect of our approach is that we couch the phenomenon of trust as one piece of a larger framework. This allows us to develop a computational understanding of trust from the top-down, starting with a definition of the phenomenon to produce algorithms that are independent of the scenario itself. We believe that this top-down approach will allow us to relate trust to other facets of social behavior important for a robot. In prior research, we have used the same framework to explore deception and the value and limitations of stereotypes to a social robot (Wagner & Arkin, 2011; Wagner, 2012).

The remainder of this article begins by presenting the portion of the vast trust literature that is most closely related to this research. In reviewing the literature, different conceptualizations of trust are examined from the perspective of feasibility of

implementation on a robot (Section 2). Next, a framework for representing trust is examined and conditions for gauging if a situation demands trust are developed (Sections 3 and 4). The hypotheses and methodology are then presented (Sections 5 and 6) followed by detailed descriptions of human subject experiments and their results. The paper concludes with a discussion section and conclusions.

2 Related Work

The word "trust" as used in everyday language, has numerous definitions (Deutsch, 1973; Luhmann, 1979; Barber, 1983). Deutsch, a psychologist, describes trust as a facet of human personality (Deutsch, 1962). He claims that trust is the result of a choice among behaviors in a specific situation. Niklas Luhmann, another early trust researcher, provides a sociological perspective (Luhmann, 1979). Luhmann defines trust as a means for reducing the social complexity and risk of daily life. Gambetta describes trust as a probability (Gambetta, 1990). Specifically, he claims that, "trust is a particular level of subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such action and in a context in which it affects his own action" (Gambetta, 1990, p. 216). Rousseau et al. have examined the definitional differences of trust from a variety of sources (Rousseau, Sitkin, Burt, & Camerer, 1998) and concluded that trust researchers generally agree on the conditions necessary for trust, namely risk and interdependence.

Lee and See review many definitions of trust and conclude that trust is "*the attitude that an agent will help achieve an individual's goals in a situation characterized by*

uncertainty and vulnerability" (Lee & See, 2004). We use Lee and See's definition of trust to generate a more conceptually precise and operational description of trust (Wagner, 2009). Trust is defined in terms of two individuals—a trustor and a trustee. The trustor is the individual at risk. The trustee represents the individual in which trust is placed. 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"***.

Several other notions of trust have been used in conjunction with autonomous systems. Information withholding (deceit) (Prietula & Carley, 2001), agent reliability (Schillo, Funk, & Rovatsos, 2000), agent opinion based on deceitful actions (Josang & Pope, 2005), compliance with virtual social norms (Hung, Dennis, & Robert, 2004), and compliance with an a priori set of trusted behaviors from a case study (Luna-Reyes, Cresswell, & Richardson, 2004) have all been used to measure trust. Models of trust range from beta probability distributions over agent reliability (Josang & Pope, 2005), to knowledge-based formulas for trust (Luna-Reyes, Cresswell, & Richardson, 2004), to perception-specific process models for trust (Hung, Dennis, & Robert, 2004). Castelfranchi and Falcone explore trust and its many definitions from a cognitive and computational perspective (Castelfranch & Falcone, 2010). They reject simplistic definitions of trust based on economic reductionism or probability and instead offer an elaborate model of trust that touches on numerous elements of cognition such as beliefs, mental states, motivations, etc.

Neuroscientists have also studied trust. Work in this area has shown that the development of a trusting relationship occurs with repeated, positive, and predictable interactions (Yamagishi, 2001; Cooper, Kreps, Wiebe, Pirkl, & Knutson, 2010) and that

activation of the amygdala is correlated to high evaluations of distrust (Adolphs, Tranel, & Damasio, 1998; Winston, Strange, O'Doherty, & Dolan, 2002). Researchers have used fMRI's to gather data while subjects play multi-round investment games to explore trust (King-Casas, et al., 2005). In these games, the investor selects some amount of money to invest, the money appreciates, and the trustee repays a self-determined portion of the money back to the investor. Previous reciprocity has been shown to be 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). The use of multi-round investor games has become an established method for investigating trust (Engle-Warnick & Slonim, 2006; Rilling, et al., 2002). Such games operationalize trust in terms of monetary exchange that allows for a simple quantitative analysis. Moreover, investment games afford a means for putting subjects at risk without the threat of physical harm. Finally, the use of a game makes the experiments repeatable.

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 (Desai, Kaniarasu, Medvedev, Steinfeld, & Yanco, 2013). A meta-analysis of these factors concludes that a robot's task performance has the greatest impact on trust (Hancock, et al., 2011). Our prior work in this area has focused on the development of a framework that allows a robot to recognize and reason about whether or not an interactive situation requires trust on the part of the robot, or the person (Wagner, 2009; Robinette, Wagner, & Howard, 2014).

3 Representing Interaction

Social psychologists define *social interaction* as influence—verbal, physical, or emotional—by one individual on another (Sears, Peplau, & Taylor, 1991). Representations for interaction have a long history in social psychology and game theory (Kelley & Thibaut, 1978; Osborne & Rubinstein, 1994). Interdependence theory, a type of social exchange theory, is a psychological theory developed as a means for understanding and analyzing interpersonal situations and interaction (Kelley, et al., 2003). Game theory also explores interaction (Osborne & Rubinstein, 1994). Game theory focuses on the formal consideration of strategic interactions, such as the existence of equilibriums and economic applications. Game theory, like interdependence theory, has long used the outcome matrix (normal-form game) to represent interaction (Kelley & Thibaut, 1978; Osborne & Rubinstein, 1994). An outcome matrix represents an interaction by expressing the utilities afforded to each interacting individual with respect to each pair of potential behaviors chosen by the individuals (Figure 1). Extended-form games can be used to represent interactions in a manner that highlights the turn-taking aspects of the same situation.

The outcome matrix is a standard computational representation for interactions and social situations (Kelley & Thibaut, 1978). Here the term “interaction” is meant to describe a discrete event in which two or more individuals select defined social behaviors as part of a social environment or context. A social situation, on the other hand, is used to describe an abstract class of interactions that have similar outcome values. Social situations ignore who is interacting and the actions from which they are choosing. For example, an investment game type of interaction would involve particular actors, say Alice and Bob, who are deciding among distinct amounts to invest and return. A social situation representing the

same game, on the other hand, would include the same pattern of utilities without describing the individuals or the actions. Conceptually a social situation can be used to explore aspects of an entire class of interactions independent of a particular scenario.

Outcome matrices can be described formally (Figure 1). The notation presented here draws heavily from game theory (Osborne & Rubinstein, 1994). An outcome matrix consists of 1) a finite set N of interacting individuals; 2) for each individual $i \in N$ a

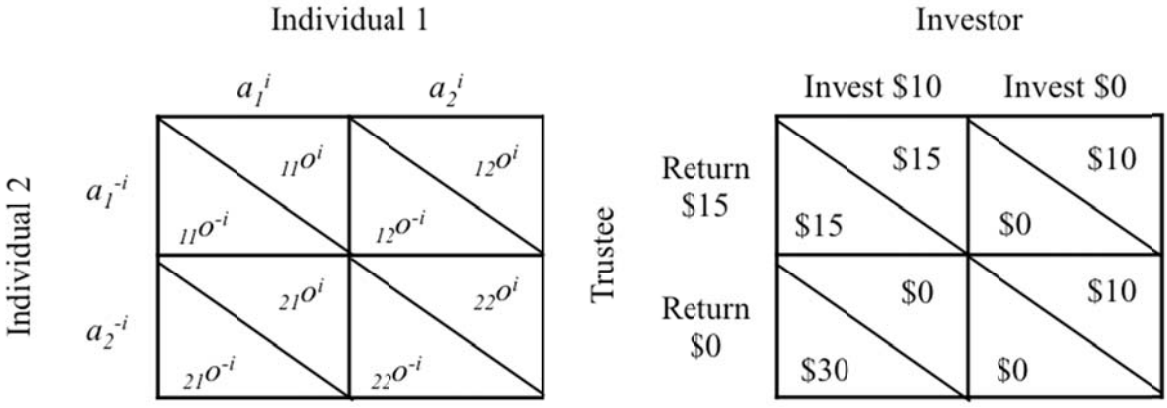


Figure 1: An example outcome matrix is depicted formally and as an investment game.

nonempty set A^i of actions; 3) the utility obtained by each individual for each combination of actions that could have been selected. The term O is used to denote an outcome matrix. A particular outcome within a matrix can be expressed as a function of an outcome matrix and an action pair, thus $O^i\left(a_2^i, a_1^{-i}\right) = {}_{12}o^i$. The variable o denotes an outcome value. The term ${}_{12}o^i$ denotes that it is individual *i*'s outcome from the first row and second column of the matrix. The superscript -*i* is used to express individual *i*'s partner. Thus, for example, A^i denotes the action set of individual i and A^{-i} denotes the action set of individual *i*'s interactive partner.

Some researchers have argued that human interaction cannot be reduced to the simple utilities represented in a matrix format (Shafir & LeBoeuf, 2002). We contend only that the use of game theoretic representations is a useful place to begin exploring social phenomena such as trust and that such a representation is easily implementable on a robot.

4 Recognizing Situations that Demand Trust

Given a means for representing social situations and a definition of trust we considered the following research question: does a series of conditions exist that would classify an interaction in terms of trust? What are these conditions? By classification we mean to make a definitive true/false statement concerning whether or not the selection of a particular action in an outcome matrix would require trust. This is not to claim that trust itself is binary. Rather, our intention is to explore whether the decision to act in a particular manner demands trust. In one-shot situations involving trust, the decision to act or not to act in a particular manner reflects a binary decision related to trust. One-shot interactions, in contrast to multi-shot interactions, do not rely on prior experience with the interactive partner. The conditions that we develop below specify whether a decision about a course of action requires trust on the part of the trustor. In related work we delineate a method for evaluating the amount of trust needed to select an action (Wagner, 2009).

Consider some examples. The trust fall is a game played to build trust in which the trustor leans backward and the trustee arrests the trustor's fall. In this situation the trustor may choose to lean back or not to lean back. Similarly the trustee must choose whether or not to catch the trustor. If the trustor chooses not to lean back then he or she risks nothing. On the other hand, once the trustor chooses to lean backward the risk they place themselves

in depends on the action of the trustee. Consider an investment game, such as the one described in Section 2, as another example. The trustor (investor) chooses whether or not to invest. The trustee must decide whether or not to return the investment. The action of investing puts the trustor at risk of losing money. Moreover, the trustor's risk can be mitigated by the actions of the trustee. The trustee's choice of action will determine if the person receives a return or loses their investment. Both situations can be represented using an extended or normal form game. In the investment game, investment amounts can be discretized into finite amounts or a continuous game can be employed.

To derive conditions for trust, let a_1^i represent a trusting action (investing with the trustee) and a_2^i represent an untrusting action (not investing) for the trustor. The definition for trust implies a specific temporal pattern for trusting interaction. Because the definition requires risk on the part of the trustor, the trustor cannot know with certainty which action the trustee will select. It therefore follows that *1) the trustee does not act before the trustor*. This order is described with the condition in outcome matrix notation as $i \Longrightarrow -i$ indicating that individual *i* acts before individual -*i*. Alternatively, an extended-form game can be used to indicate action selection order.

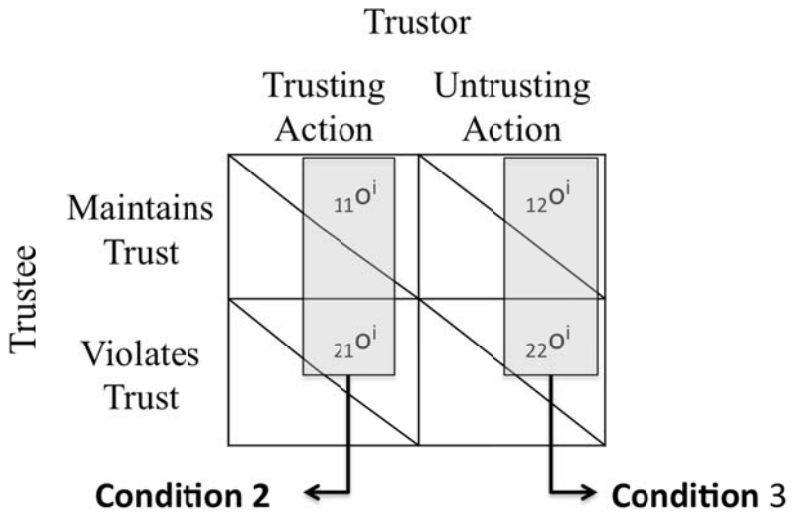


Figure 2 The figure depicts the outcome values that are compared for trust conditions 2 and 3.

The definition for trust indicates that situations involving trust put the trustor at risk. Risk can be modeled as the potential loss of outcome. Hence, selection of the a_1^i (trusting action) results in a possible loss $l = {}_{11}o^i - {}_{21}o^i$ where $l > 0$. Because small risks tend not to have a large impact on decision-making, we can define a constant ε_1 representing the minimal amount of loss necessary for a risk to influence one's decision (Adolphs, Tranel, & Damasio, 1998). The loss necessary for trust then is quantified as ${}_{11}o^i - {}_{21}o^i > \varepsilon_1$. Note that the outcome values (${}_{11}o^i$ and ${}_{21}o^i$) vary across the trustee's action choices (Figure 2). Hence, whether or not the trustor loses outcome when selecting the trusting action depends entirely on the action choice of the trustee. In related work we quantify the amount of trust as proportional to the loss, $T \propto l$ (Wagner, 2009). Stated as a condition for trust, *2) the outcome received by the trustor depends on the actions of the trustee if and only if the trustor selects the trusting action.*

The definition 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. From the discussion above, an untrusting action is an option that does not require risk. Formally, $\left|{}_{12}o^{i} - {}_{22}o^{i}\right| < \varepsilon_2$, 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, *3) the outcome received when selecting the untrusting action does not depend on the actions of the trustee.*

Conditions 2 and 3 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, *4) the value, for the trustor, of fulfilled trust is greater than the value of not trusting at all, is greater than the value of having one's trust broken.* Formally, the outcomes are valued ${}_{11}o^{i} > {}_{x2}o^{i} > {}_{21}o^{i}$ where *x* is 1 or 2.

Finally, the definition demands that, *5) the trustor must hold a belief that the trustee will select action a_1^{-i} with sufficiently high probability,* formally $p^{i}(a_1^{-i}) > k$ where *k* is some sufficiently large constant.

Figure 3 presents these conditions with respect to the investment game described in Section 2. We assume that the first condition is met. The matrix in Figure 3 results in outcome values $l = {}_{11}o^{i} - {}_{21}o^{i} = 15 - 0 = 15$. The second condition considers $15 - 0 > \varepsilon_1$. Thus, in this example, action a_1^{i} depends on the action of the trustee if $\varepsilon_1 < 15$. The values assigned to the constants $\varepsilon_1, \varepsilon_2, k$ are likely to be trustor specific. In related work we

have explored the possibility of relating these constants to the trustor's prior experience with different types of trustees (Wagner, 2013) For instance, a lower value of ε_1 reflects a more risk-averse trustor. A lower value of ε_2, on the other hand, reflects a tighter threshold associated with the untrusting action. For this example, the third condition results in values, $|10 - 10| < \varepsilon_2$. Here, the action a_2^i does not depend on the actions of the partner for constant $\varepsilon_2 > 0$. The final condition results in values $15 > \{10,10\} > 0$. Hence, for the

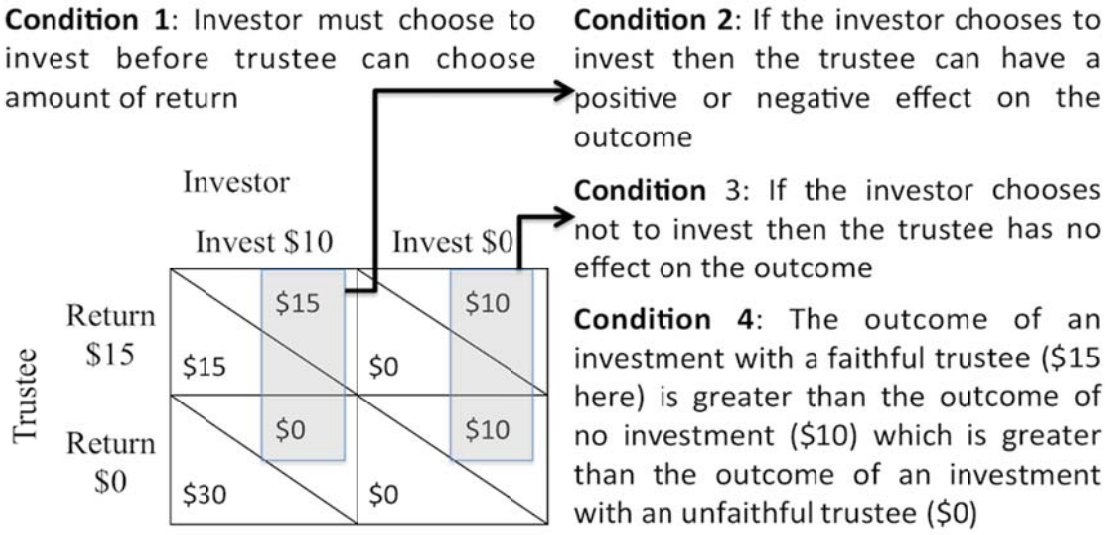


Figure 3 An example of an investment game with annotations to indicate how the trust conditions are applied is depicted. In this example, an investor can choose to invest in a trustee or not. If he invests then the trustee will receive three times the investment and will be able to decide how much to return to the investor. For the purposes of illustration, only the cases where the trustee would return an even split or no money at all are considered.

investor, the selection of action a_1^i involves risk that can be mitigated by the actions of the trustee and the selection of action a_2^i does not involve risk that is mitigated by the actions of the trustee.

The first four conditions describe the **situational conditions** necessary for trust. We argue that the situational condition must be satisfied in order for an interaction to require

trust on the part of the trustor. Testing a situation for these conditions therefore determines if an interactive situation requires trust.

The final condition for trust is based on the trustor's model of the trustee. In a sense, it captures the trustworthiness of the trustee. It requires that the trustor predict the likelihood that the trustee will select the action which mitigates the trustor's risk, formally $p^i(a_1^{-i})$. In prior work, we have shown that predictive models of the trustee can be built from experience (Wagner, 2009) or bootstrapped by stereotyping (Wagner, 2013). This final condition addresses the temporal aspects of trust such as reputation building and confidence.

Many trust models focus on characteristics of the trustee, such as reputation (Schillo, Funk, & Rovatsos, 2000). Again consider the trust fall. A model of the trustee would predict how likely the person is to catch the trustor's fall. Yet it would not fully capture the risk associated with the fall. Such risk might be a product of the environment. For example, if the trust fall is performed over broken glass the risk is significantly greater then over grass, regardless of the reputation of the trustee. Or it might be a product of the trustor, for example, if the trustor is elderly. In either case, the outcome values associated with the matrix capture this risk whereas a model of the trustee in itself does not.

These situational conditions for trust are a small part of a larger framework. We contend that the conditions can be used to evaluate an outcome matrix representing a particular interaction. We have explored methods for creating outcome matrices in prior work (Wagner, 2009). We believe that these conditions offer a method for determining if an interaction demands trust. In order to justify this belief we attempt to show that the matrices

predicted to require trust by our conditions positively correlate to those deemed to require trust by people.

5 Experimental Hypotheses

The purpose of these experiments is to evaluate the extent to which the conditions for situational trust that were derived from our definition of trust correlates to the classifications made by human subjects. With respect to this purpose, the following hypotheses are enumerated:

1: Situations that meet our conditions for trust correlate strongly to those selected by participants to demand trust. With respect to the behavioral sciences, a correlation of $\phi > 0.30$ is considered strong and a correlation of $0.20 < \phi < 0.30$ is considered moderate (Hemphill, 2003).

2: The correlation between our conditions and the participant's evaluations is true regardless of the trustor's gender, the type of scenario, the magnitude of outcome values, whether the action was stated in a positive manner or negative manner, or whether the trustee or robot performed well or poorly. These variables represent possible confounding factors.

3: The correlation between our conditions and participant's evaluations is true regardless of whether the subjects are reading narratives about the actions of others or selecting behaviors with a robotic teammate.

6 Methodology

Two different sets of complimentary experiments were conducted. The first experiment required participants to read a fictional narrative about two people and to decide whether or not the selection of a particular action demanded trust. The second experiment placed participants in a maze and allowed them to decide whether or not to use a robot as a guide through the maze.

This article focuses on one-shot interactions. In contrast to multi-shot interactions, one-shot interactions require a person to make a social decision without any prior experience with their interactive partner. For most humans, one-shot interactions are very common. Emergency situations also tend to be one-shot interactions. This is an area of interest for us. In other work, we have explored the possibility of robots guiding humans during an emergency (Robinette, Wagner, & Howard, 2014) and defined situations where a human may not trust a robot during an emergency (Robinette, Wagner, & Howard, 2013).

Crowdsourcing was used to collect data for both experiments. Crowdsourcing is a method for collecting data from a relatively large, diverse set of people (Paolacci, Chandler, & Ipeirotis, 2010). Crowdsourcing sites, like Amazon's Mechanical Turk, post potential jobs for crowdworkers, manage worker payment, and track worker reputation. The use of crowdworkers offers a quick and efficient complement to traditional laboratory experiments. Moreover, the population of workers that provide the data tends to be somewhat more diverse than traditional American university undergraduates. In order to ensure the best possible data, individuals were required to have a 95% acceptance rate for their past work and were only allowed to participate once.

6.1. Trust versus No-Trust Matrices

The conditions for trust (Figure 3) can be used to indicate if a situation demands trust or not. Because there are many different types of matrices that do not meet the conditions for trust, we created six categories to capture different classes of matrices that do not meet the conditions.

Table 1 Different categories of trust and no-trust matrices are presented with representative examples.

Category of Matrix

Category	Abbreviation	Example
Trust Matrix	*TR*	Trustor: a_1^i, a_2^i / Trustee: a_1^{-i} (\$2000, \$400), a_2^{-i} (\$0, \$400)
Equal Outcomes	*EO*	Trustor: a_1^i, a_2^i / Trustee: a_1^{-i} (\$2000, \$2000), a_2^{-i} (\$2000, \$2000)
Trustor-Dependent, Trustee-Independent	*DI*	Trustor: a_1^i, a_2^i / Trustee: a_1^{-i} (\$2000, \$0), a_2^{-i} (\$2000, \$0)
Trustor-Independent, Trustee-Dependent	*ID*	Trustor: a_1^i, a_2^i / Trustee: a_1^{-i} (\$2000, \$2000), a_2^{-i} (\$0, \$0)
Inverted Trust Matrix	*INV*	Trustor: a_1^i, a_2^i / Trustee: a_1^{-i} (\$0, \$400), a_2^{-i} (\$2000, \$400)

The **Equal Outcome (EO)** category of matrix represents matrices in which all of the outcome values are equal. This type of matrix violates the second and fourth conditions from the conditions for situational trust (see Table 1 row 2 for an example). For this matrix, the outcome received by the trustor does not depend on the action choice of either the trustor or the trustee. As a result, this type of matrix explores whether or not the participants view trust as inherently part of a particular scenario. Results indicating that participants regard this type of matrix as requiring trust may reflect the view that outcome values do not matter when evaluating trust.

The **Trustor-Dependent, Trustee-Independent (DI)** category represents matrices in which the action selected by the trustee has no impact on the outcome received by the trustor (Table 1 row 3). In this case, the trustor's outcome depends only on the trustor's action selection. Hence the trustor has complete control. With respect to the conditions for trust, this violates the second condition because $\varepsilon_1 = 0$ and the fourth condition because ${}_{11}o^i = {}_{21}o^i < {}_{x2}o^i$ or ${}_{11}o^i = {}_{21}o^i > {}_{x2}o^i$ depending on the exact matrix. Matrices from this category were created either with greater outcome assigned to the trusting action (a_1^i) or, alternatively, greater outcome assigned to the untrusting action (a_2^i). These matrices are abbreviated as DI-1 and DI-2, respectively.

The **Trustor-Independent, Trustee-Dependent (ID)** category represents matrices which place total control with the trustee. For this category, the trustor's outcomes depend only on the action selected by the trustee. Intuitively, this type of matrix presents the trustor with only risky actions. No untrusting action is offered. For this category, the third condition is violated because $\varepsilon_2 \neq 0$. The fourth condition is also violated because ${}_{11}o^i =$

${}_{12}o^i < {}_{21}o^i = {}_{22}o^i$ or ${}_{11}o^i = {}_{12}o^i > {}_{21}o^i = {}_{22}o^i$, depending on the exact matrix. Matrices from this category could also be created in two different ways. Greater outcome for the trustor could either be assigned to the trustee's action that maintained trust (a_1^{-i}) or, alternatively, to the action that violated trust (a_2^{-i}). These matrices are abbreviated as ID-1 and ID-2 respectively.

Finally, the **Inverted Trust (INV)** category of matrices presents the subject with a situation in which the outcome received by the trustor is greater if the trustee violates the trust. These matrices only violate the fourth condition because ${}_{21}o^i > {}_{x2}o^i > {}_{11}o^i$.

7 The Narrative Experiment

In order to test the hypotheses presented in Section 5 we designed two human subject experiments that asked people to evaluate different situations in terms of trust. The first experiment required participants to read written narratives describing a situation representing an outcome matrix. Written narratives are a flexible way to present a wide variety of different trust situations to the human participants in a way that would be easily understood. Still, the narratives differed from real situations and interactions in that participants acted as observers judging the actions of others. Three general scenarios were used:

1. An investment scenario
2. A navigation guidance scenario
3. An employee hiring scenario

These scenarios were designed to be simple and understandable to non-academics but also sufficiently adaptable to represent the wide variety of outcome matrices from the trust and no-trust categories described in Table 1. Best practices in research design were used to create the narratives (Mitchell & Jolley, 1992).

The study design was informed by several pilot studies. These pilot studies indicated that keywords such as “invest”, “follow”, or “hire” biased some participants to conclude that all scenarios where the trustor decided to invest in, follow, or hire the trustee required trust, regardless of the actual outcomes values. This well-known bias is called the anchoring bias and describes the human tendency to focus heavily on the first piece of information when making decisions (Tversky & Kahneman, 1974). In response, we modified the scenarios to use abstract terms such as “perform an action” rather than “invest in.” These modifications made the scenarios vague with respect to the action that was performed but did not change scenarios in any other way. Examples of the narratives are presented in Figure 4.

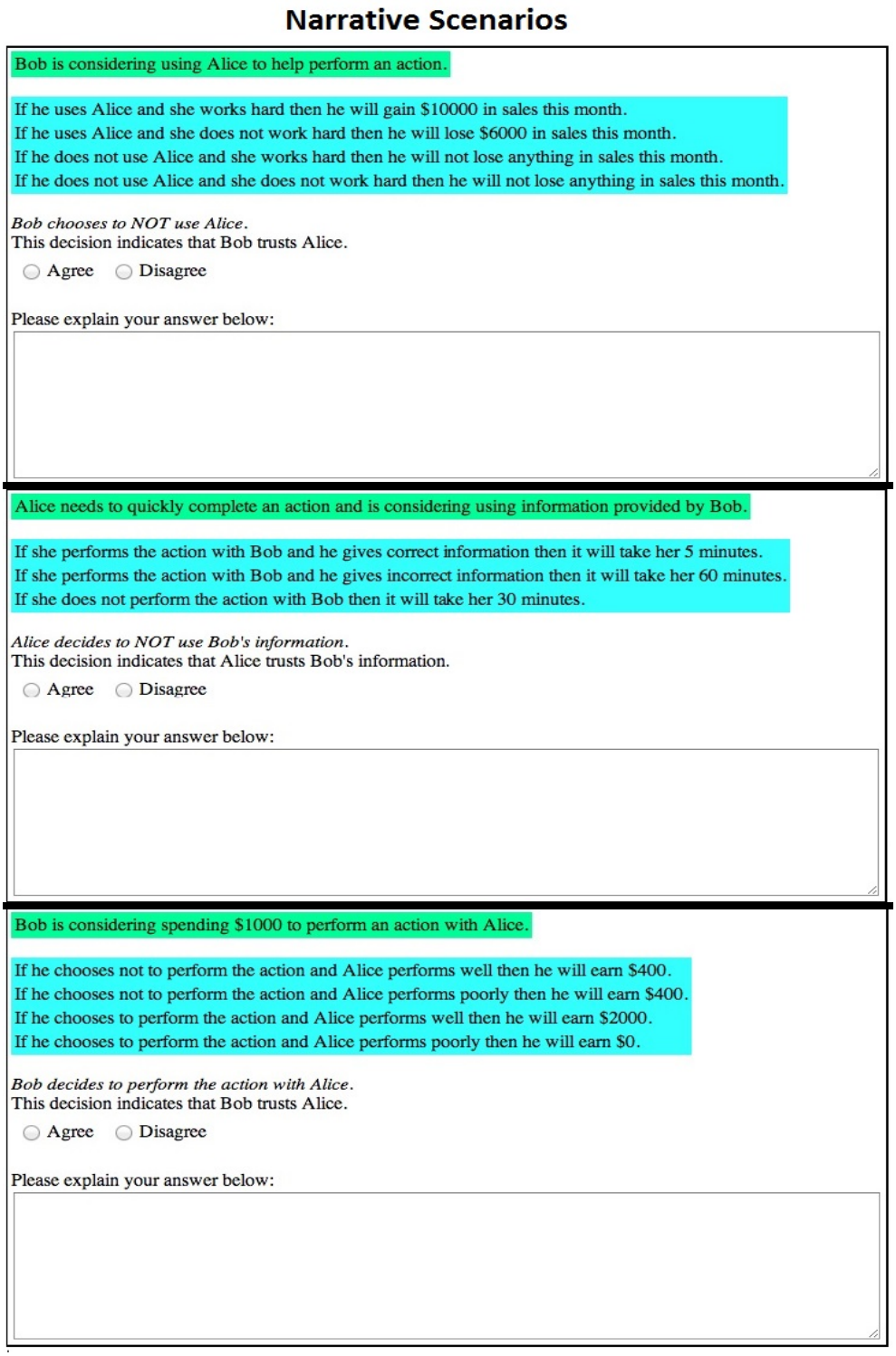


Figure 4 Examples of the three different scenario narratives are depicted. The highlighting and format are the same as those presented to the participants.

The experiment consisted of four separate runs. Each run compared a set of narratives that were based on matrices that met the conditions for trust to narratives based on one category of matrices that did not meet the conditions for trust (rows 2-5 from Table 1). In runs 1 and 4 half of the narratives were based on trust matrices and half were based on no-trust matrices (Table 2). Run 1 compared narratives based on trust matrices with narratives based on the Equal Outcomes Matrices. Run 4 contrasted trust matrices with narratives based on Inverted Trust matrices. Runs 2 and 3 presented trust matrices in one-third of the narratives and no-trust matrices in the other two-thirds. Run 2 presented the Trustor-Dependent, Trustee-Independent category of matrices in the following manner: one-third of the matrices required trust, one-third of the matrices rewarded action a_1^i regardless of the partner's action and the remaining one-third rewarded action a_2^i regardless of the partner's action. Run 3 presented the Trustor-Independent, Trustee-Dependent matrices in the same manner. Runs 2 and 3 eliminated the Employee Hiring Interaction so as to make the total number of questions consistent across all runs. Table 2 depicts a breakdown for all runs of the experiment.

Table 2 A breakdown of the different matrix categories for each run of the experiment. The total number of narratives for each run was 384.

Conditions for the Narrative Experiment

Run	**Trust Narratives Answered**	**No-Trust Narratives Answered**		**Matrix Types**	**Hiring narrative**
1	192	192		Trust and EO	Yes
2	128	256 total	128 greater outcome a_1^{-i}	Trust and DI	No
			128 greater outcome a_2^{-i}		
3	128	256 total	128 greater outcome a_1^{-i}	Trust and ID	No
			128 greater outcome a_2^{-i}		
4	192	192		Trust and INV	Yes

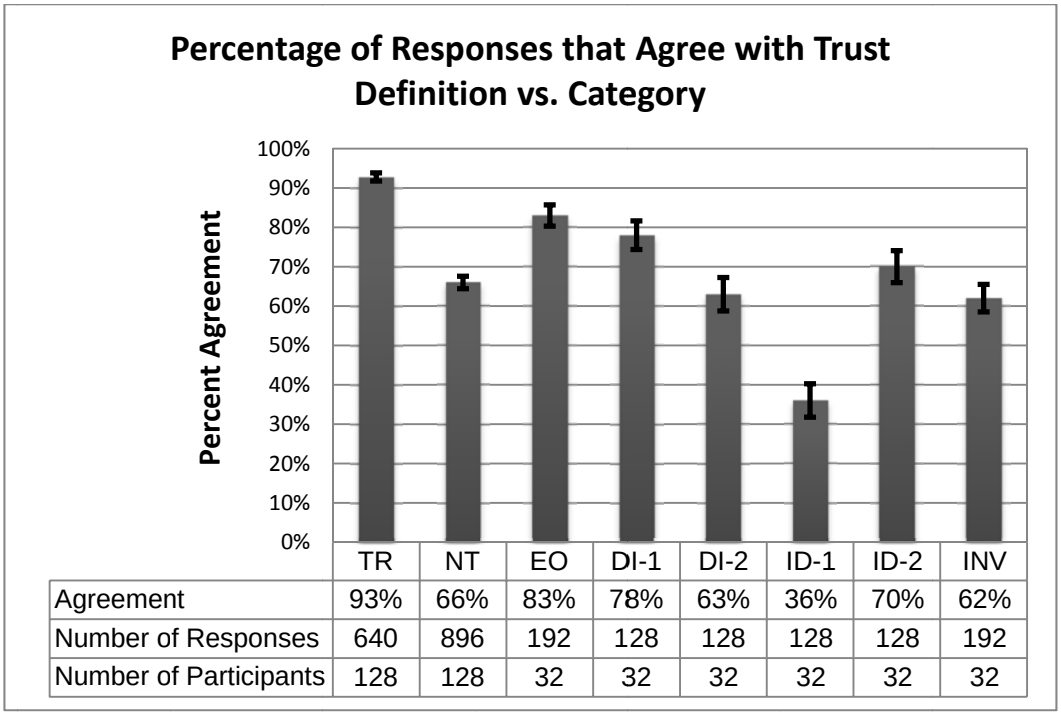
Participants began the study by reading a consent form. They were then directed to read twelve different narratives. Each narrative involved a trustor and a trustee. The names "Alice" and "Bob" were used in all narratives. Half of the narratives seen by a participant had Bob as the trustor and Alice as the trustee and the remainder were reversed in order to test for gender bias. Each narrative ended with an action chosen by the character in the narrative (see Figure 4 for examples), e.g. "Bob decides to perform the action with Alice." Following this action was the statement "This decision indicates that Bob trusts Alice." Participants were then asked to agree or disagree with this statement for each narrative and state their reasons for their answer. Their choice and statement for each narrative was recorded. The surveys were conducted over the internet through a web browser. No participant was allowed to participate in the study more than once. Participants who completed the survey were paid $1.67. IRB approval was obtained for this study.

In order to ensure that the relationship between the trust/no-trust conditions and the data collected was not spurious, we randomized with respect to several potential confounds. The gender of the trustor was randomized and recorded as a possible confound. Whether an action was positively stated (e.g. Bob hires Alice) or a negatively stated (e.g. Alice does not follow Bob's directions) was also randomized and recorded. Finally, the value of the outcomes were randomly chosen to be either x or $2x$ and recorded to determine participants' sensitivity to the magnitude of the values.

A total of 48 narratives were generated to represent all possible combinations of these variables for each run. Each participant read 12 different narratives. The order of the narratives was randomized. Eight different participants were asked about each specific combination of the variables.

7.1. Results

The results from the experiment indicate a strong, positive correlation between the matrices deemed to require trust by the human subjects and the predictions of the situational trust algorithm, $\phi(1536) = +0.592, p < 0.01$. These results are based on answers from a total of 128 different participants who read and responded to 1536 narratives. For data involving human subjects, this represents a strong, positive correlation (Hemphill, 2003) and supports our first hypothesis. A total of 77.4% of responses agreed with the algorithm's predictions across all surveys and outcome matrix categories.



	TR	NT	EO	DI-1	DI-2	ID-1	ID-2	INV
Agreement	93%	66%	83%	78%	63%	36%	70%	62%
Number of Responses	640	896	192	128	128	128	128	192
Number of Participants	128	128	32	32	32	32	32	32

Figure 5 The results from our narrative experiment are depicted above. The abbreviations correspond to those presented in Table 1. The result for all trust matrices in all rounds is denoted as TR. The result for all no-trust matrices is labeled NT. Thus NT is an aggregate of EO, DI-1, DI-2, ID-1, ID-2, and INV. Hence, round one included TR and EO, round two included TR, DI-1, and DI-2, round three was TR, ID-1, and ID-2, and round four consisted of TR and INV. The error bars denote 95% confidence intervals.

Analyzing the results by category produces a clearer picture of the participants' agreement with the algorithm and definition. In narratives predicted by the algorithm to require trust, 92.8% of responses agreed (out of 640 total responses from 128 different participants), over the course of all four runs (Figure 5 TR overall). Looking at the runs individually, for the trust matrices agreement ranged from 96.9% (run 4) to 87.5% (run 3). Tests for statistical significant across each pairwise combination of runs indicated only a single significant difference (two sample t-test, $p < 0.01$), between runs 3 and 4. Thus, our conditions for trust consistently had a high degree of agreement with participants' selections.

Table 3 A sample of comments from the participants is presented below.

Representative Comments from Trust Matrix Participants

He stood to lose $400 if he he [*sic*] trusted her and was wrong. Since he chose to work with her and put that money at risk, he must trust her.
This is completely trust. He runs the risk of losing everything yet bets it all on her competence.
This does indeed indicate trust. With Bob deciding to perform the action, he is putting trust in her that she will perform well. There is a lot at stake by performing the action with Alice. There is indeed risk.
If she didn't, then she wouldn't run the risk of doubling the amount of time the action could take her. She clearly trusts him.

The participant's comments also tended to indicate that they recognized the connection between risk and trust in the narratives (Table 3). For example, when the trustor chooses to perform an action requiring trust, participants often commented that the trustor must believe that the trustee would mitigate the trustor's risk (our language). Likewise, when the trustor chooses not to perform the trusting action, participants noted that the trustor must have felt that he had a better chance on his own. Both of these responses

strongly agree with the definition of trust. There was little consensus in the comments of the 7.8% of responses that disagreed with the condition's prediction that the narrative requires trust.

There was slightly less agreement when the participants were presented with narratives deemed by the conditions not to require trust. Each run examined a different category of matrix that did not require trust. Figure 5 presents the results (EO, DI-1, DI-2, ID-1, ID-2, and INV). The percent agreement in the case of the no-trust matrices ranged from 83.3% (Equal Outcomes) to 35.9% (Trustor-Independent, Trustee-Dependent Matrices-Action 1 Rewarded). Hence, the type of no-trust matrix faced by the participant impacted one's agreement with the conditions. The percent agreement over all no-trust narratives was 66.4%.

For the equal outcomes category, there was 83.3% agreement with the conditions. Participants evaluating this category of narrative predominately confirmed that if all outcomes are equal then any decision made by the trustor did not require trust. A small minority of participants, however, indicated that performing any action with the trustee requires trust, even if there is no risk or reason for performing the action.

For the Trustor-Dependent, Trustee-Independent category of outcome matrices, the strength of the results depended on which action was rewarded. In matrices where the trusting action (a_1^i) produced a greater reward, 78.1% of responses agreed that the narrative did not involve trust. Yet, when the untrusting action (a_2^i) produced a greater reward, 63.3% of responses agreed with the no-trust prediction, in spite of the fact that these narratives violated the same conditions. Although both results agree with the hypothesis,

we speculate that the decrease in agreement reflects the oddity of a narrative in which not trusting someone results in maximal reward. In the investment scenario, for instance, the trustor decides not to invest and the trustee does generate a poor return, yet the amount received by the trustor is maximal. Participant's comments indicate that this type of no-trust matrix caused some people to reason that the trustee must have performed better than the narrative indicated.

The Trustor-Independent, Trustee-Dependent category showed similar disparity. Narratives where the trustor received greater reward when the trustee violated the trust (a_2^{-i}) resulted in 69.5% of responses stating no-trust. Yet, only 35.9% of responses indicated no-trust when the trustor received greater reward if the trustee maintained trust (a_1^{-i}). According to the comments, trust can occur even if the trustor's choice has no bearing on the result. Many participants explained their answer by simply commenting "Bob trusts Alice's performance," "She is relying on Bob to perform well, whether she performs or not," and similar statements. The difference may have been compounded by the wording in the narratives. The a_1^{-i} action was referred to as "performs well" or "gives correct information" and a_2^{-i} was referred to as "performs poorly" or "gives incorrect information." The comments seemed to indicate that subjects had a difficult time imagining a scenario in which the trustee "performs poorly" and yet the trustor received the maximal reward.

For the Inverted Trust category, 62.0% of responses agreed that the narrative did not require trust. Some participants stated that they believed it to be impossible to trust an

individual to perform an action in an unfaithful manner and thus trust is not possible in this situation.

Not surprisingly, we found that participants tended to invent reasons that explained the trustor's choice of actions. If the trustor performed an action that was against his benefit (according to the outcome matrix) or did not perform an action that would be to his benefit, participants occasionally invented stories to justify the person's behavior. For example one participant stated, "Bob uses Alice's information since he trusts the information enough to thoroughly finish in 120 minutes. He'd rather take the time to correctly finish something over finishing it fast."

With respect to the potential confounding factors, we found that the results were not a reflection of the particular scenario as there was no statistical difference between the three scenarios, $\phi(1531) = -0.011, p > 0.05$. The phi coefficient, $\phi = \frac{n_{11} \cdot n_{00} - n_{10} \cdot n_{01}}{\sqrt{n_0 \cdot n_0 \cdot n_1 \cdot n_{1\cdot}}}$, calculated from a 2x2 contingency table, was used determine correlation (Runyon & Audrey, 1991). Further, the results were not impacted by the gender of the trustor or trustee, $\phi(1536) = -0.017, p = 0.50$ or the magnitude of the outcome values $\phi(1536) = +0.030, p > 0.05$. Statistically significant differences did result from positive and negative action labels; but the correlation of these labels was small and negative, $\phi(1536) = -0.104, p < 0.001$. We believe that the influence of positive and negative labels represents an experimental artifact the absence of which would have slightly strengthened our results. Overall, these results support our second hypothesis (Section 5).

The narrative experiment provides preliminary evidence that our conditions for situational trust correlate to the evaluations made by people. The results also show that this

correlation is not limited to a single scenario. Still, the use of narratives forced participants to reason about the interactions of fictional third parties. For this reason, we conducted a follow-up experiment in which the participants had to decide whether or not they trust a robot.

8 Robot Guidance Experiment

We conducted a robot guidance experiment as a follow-up to the narrative experiment. The guidance experiment placed participants in a simulated maze and tasked them with finding an exit. This scenario was motivated by our interest in developing robots that can provide guidance during a fire. For this reason, we developed a maze that was roughly similar to an office environment. Participants are placed in this environment and then given the option of using a guidance robot to assist them with navigating the maze. Regardless of their choice, once they completed the navigation task, they were then asked whether or not they trusted the robot and if their decision to use or not use the robot showed that they trusted the robot.

We created two types of maze, which were meant to correspond to the trust/no-trust matrices. In the trust matrix condition, the maze had several walls and barriers preventing participants from easily moving directly to the exit (Figure 6 top). The no-trust condition (Figure 6 bottom) was meant to correspond to the equal outcomes (EO) matrix from Table 1. The equal outcomes matrix reflects a risk-free situation in which the participant expects to receive the same outcome regardless of how either the robot or the person acts. Hence, for this condition, the exit was visible and directly in front of them, although it was located at a distance.

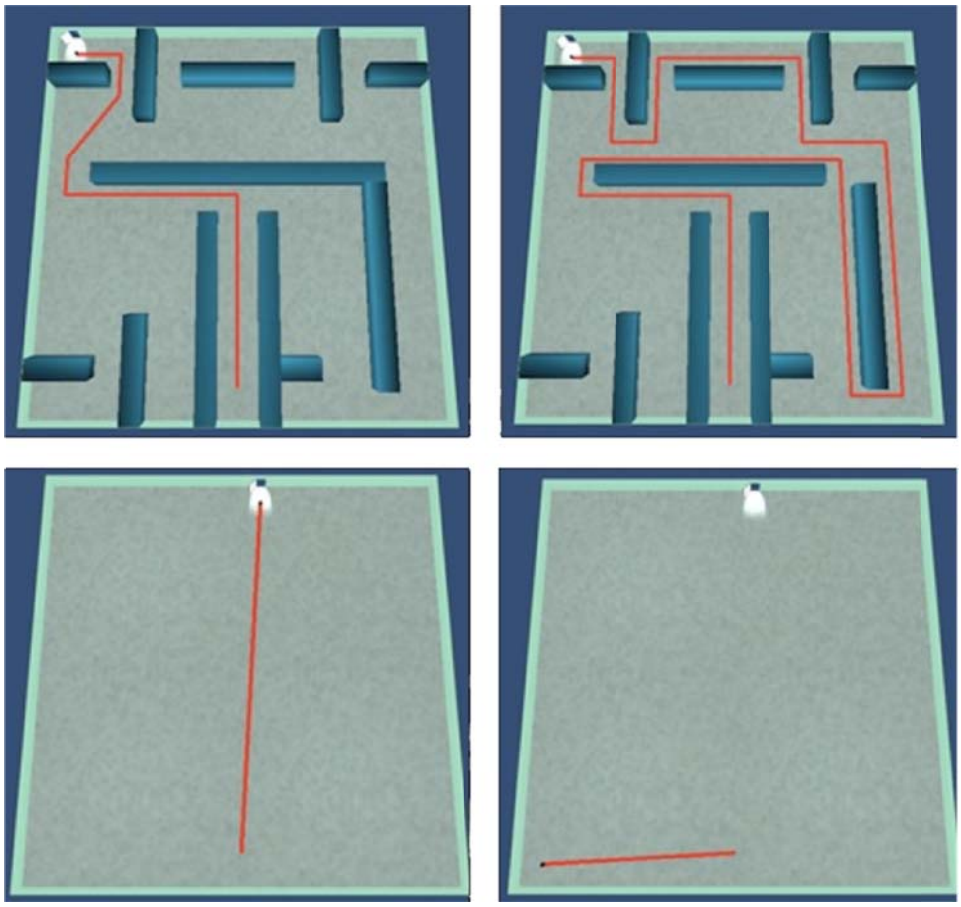


Figure 6 The top two images depict example mazes from the trust condition case and the bottom two images depict mazes for the no-trust condition. The images on the left illustrate the performance of good robots and the right images illustrate the performance of bad robots.

We hypothesized that self-reports of trust by the participants would correlate to the type of outcome matrix that the maze was based upon. Moreover, we hypothesized that the robot's performance (good versus bad) would not significantly impact the subject's trust self-report in the low risk maze.

Again Amazon Mechanical Turk crowdworkers were utilized as participants. A total of 120 people completed the experiment. Blender, an open source 3D modeling software package, and the Unity Game Engine were used to create the simulated robot and maze environment (Figure 6) in which the crowdworkers were placed. Each maze had only a single exit. Participants performed the experiment using the Unity Web Plugin and a web

browser. The robot that provided guidance was based on a Willow Garage TurtleBot 2 robot but also had two PhantomX Pincher AX-12 arms to garner attention. Results from our prior research indicate that this style of robot communicates directions that are easily understood by people (Robinette, Wagner, & Howard, 2014).

Each subject began the experiment by reading an introduction broadly describing the experiment and then consenting to participate. The introduction stated that we were testing how people leave buildings in emergencies and encouraged them to act as if they were in a real emergency. Extensive experimentation as part of our related research indicated that the use of this emergency scenario served as better motivation than the use of a monetary bonus (Robinette, Wagner, & Howard, 2014). Next, a short tutorial allowed the person to practice navigating a maze. They were then told that, because this was a simulated emergency, they would only have a certain amount of time to leave the building. No exact amount of time was provided. We noted that if they failed to locate the exit in time then their character would not be deemed to have survived.

In the trust condition, the pictures depicted example mazes along with survival rates. The survival rate for people that choose to use a good robot was listed as "typically survived," those that choose to use a bad robot stated as "typically did not survive," and did not choose to use the robot presented as "about half…survived." In the no-trust condition, all pictures presented a maze without barriers and with the exit clearly visible from the starting point. The survival rates in this condition noted that the person typically survived regardless of whether or not they choose to use the robot or whether the robot was good. Again this condition was meant to be risk-free. Figure 7 presents the introduction screen, the example screens for the trust and no-trust cases, and the survey page.

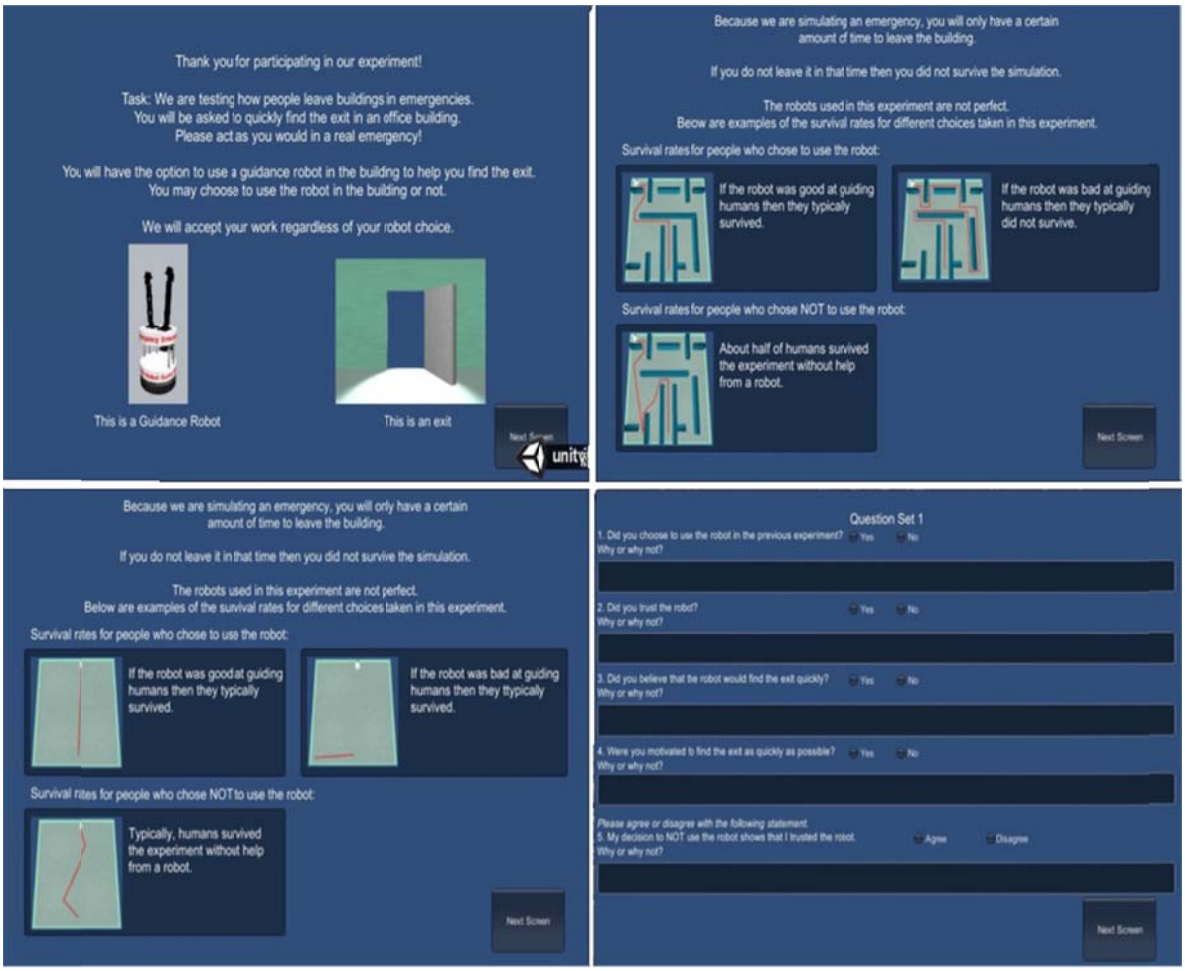


Figure 7 Screenshots depicting the guidance experiment's introduction screen (top left), trust condition example screen (top right), no-trust condition example screen (bottom left), and survey (bottom right). A participant would see either the trust condition examples or the no-trust condition example, but not both.

Next the participants were asked to choose whether or not they wanted to use the robot for guidance by pressing a button before the experiment started. In the no-trust case they were told that the maze would be the same as the one presented in the examples. In the trust case, they were informed that the maze would be different from the examples.

Once the button was pressed the software placed the participant in the maze. If they choose to use the robot, the participant was spawned in the maze with the guidance robot nearby. The robot would begin to move as soon as the participant moved. The participant could choose to navigate the maze with or without the robot's help. They were given 60

seconds to navigate the maze. Their remaining time was prominently displayed in the center of their screen (Figure 8). In both conditions, half of the robots provided good guidance and half provided poor guidance. Whether or not the participant pressed the button requesting the robot's guidance was recorded as their decision. We have not examined the extent to which participants actually followed the robot. Results from prior research indicate that subjects who choose to use a robot for guidance tend to continue to follow it regardless of its performance (Robinette, Wagner, & Howard, 2014).

Figure 8 A screenshot of the view of the environment that participants saw in the emergency condition.

Upon completing the maze participants were asked to take a short survey (Figure 7). Two questions focused on trust. The first trust-related question asked, "Did you trust the robot?" The second question asked participants whether they agreed or disagreed with the statement, "My decision to use the robot shows that I trusted the robot." If the participant

had chosen not to use the robot then the word “NOT” is inserted before the word “use.” Finally, a second set of survey questions collected demographic information.

8.1. Results

A total of 120 participants (mean age = 31.25, 40% female, 94.1% United States nationality) completed the experiment. The results from the experiment indicate a strong positive correlation between participant’s trust self-report and the predictions of the situational trust algorithm as to whether or not the maze required trust, $\phi(120) = +0.406, p < 0.001$ (Hemphill, 2003). The correlation is not as strong as in the narrative experiment (where $\phi = +0.592$). Our conditions for trust were met when the participants were presented with a trust maze and chose to use the robot. This occurred for 50 of the 120 subjects. When these conditions were met $74.0\% \pm 12.2\%$ of participants reported trust (Figure 9). For 70 of the 120 participants the conditions for trust were not met because either they were presented with a no-trust maze or they choose not to use the robot. In this case only $32.9\% \pm 11.0\%$ reported trust. This difference is statistically significant (Pearson’s chi-squared $\chi^2(1, 120) = 6.53, p < 0.001$). The Pearson’s chi-squared test, calculated as $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ where O and E are observed and expected value frequencies, is a standard test of significance for categorical data (Runyon & Audrey, 1991). The results were nearly identical regardless of which of the two self-report questions were used for the analysis. The result supports our hypothesis that participant responses correlate to the type of matrix. Overall, 76.7% of participants chose to use the robot for guidance.

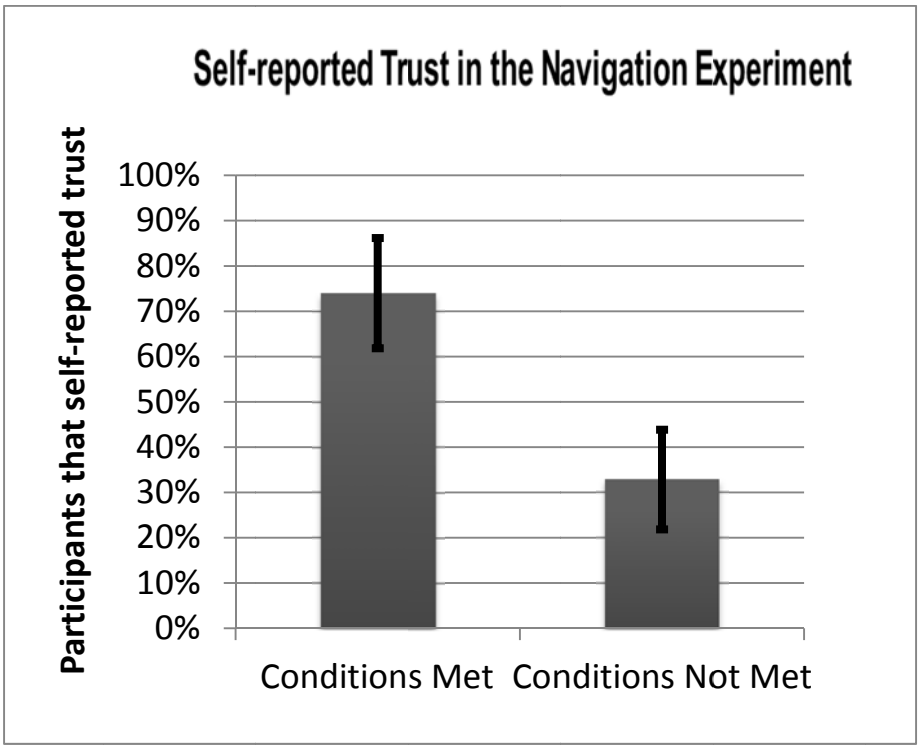


Figure 9 The results from the robot guidance experiment are depicted above. When the conditions for trust are met, participants were significantly more likely to self-report trust.

One potential confounding variable is the quality of guidance provided by the robot. Yet our overall results indicated that the quality of guidance did not significantly correlate to participant's reports of trust (Phi coefficient $(120) = +0.067$; Pearson's chi-squared $\chi^2(1, 120) = 0.53, p = 0.47$). This experiment thus provides additional evidence that our conditions for situational trust do correlate to the evaluations made by people.

9 Discussion

The results from the experiments support our primary hypothesis that outcome matrices that meet the conditions for trust are also deemed to require trust by people. In support of our third hypothesis from Section 5 we found this to be true regardless of whether the subjects are reading narratives about the actions of others or selecting behaviors with a robotic

teammate. Still, the results from the narrative experiment indicated a stronger correlation (+0.592) than the results from the guidance experiment (+0.406). Part of the reason for this difference may have been a social desirability bias present in the guidance experiment. Social desirability bias is a subject's tendency to respond in a manner that is socially desirable (Fisher, 1993). Social desirability may have influenced subjects to report that they trusted the robot regardless of how they actually felt. As evidence of this bias, nine subjects reported trusting the robot even though it headed in a direction that was clearly away from a visible exit and then stopped moving (Figure 6 bottom right depicts its path).

We assume that if Amazon Turk workers have an inherent bias for or against robots and technology, then the bias would equally influence results in both situations that met the conditions for trust and in situations that did not meet the conditions for trust. In other words, participants with a pro or anti-robot bias were approximately equally distributed across both conditions and, as such, participant bias would not account for the results. Because condition selection was random, and the sample was relatively large, we have no reason to believe, or rule out, the possibility that inherent participant bias may have influenced the results. A follow-on experiment could be conducted using the paradigm we present comparing pools of subjects with inherent pro and anti-robot biases. Although interesting, we have no immediate plans to perform such an experiment.

The experiments focused on the impact of the situation on trust. We have attempted to rule out the influence of several potential confounding variables such as the quality of guidance by the robot, the gender of the agents, the action selected by the trustor, and the context of the narrative (investment, navigation, employment related). The data indicates that these factors were not responsible for the results. We therefore conclude that the results

support our contention that facets of the situation, namely risk, strongly influence the trust decisions of human subjects. Moreover, this situational trust can be captured in a series of conditions implementable on a robot. We conclude by considering the importance and limitations of these results.

10 Conclusion

This article has investigated a set of conditions for determining if an interactive situation demands trust. Our focus has been to evaluate the extent to which these conditions correlate to the classifications made by people. Our motivation for doing so is to develop a general conceptualization of trust that could be used to guide the behavior of a robot. We have presented experiments that examine how people evaluate trust. Their evaluations were compared to the predictions of our conditions for trust. The results indicate a correlation between the predictions and the participants' evaluations. Overall, our data supports the contention that perceived risk is central to the trust phenomenon, even for interactions involving a robot. These results are in agreement with both the trust community (Gambetta, 1990; Sabater & Sierra, 2005) and the robotics community (Desai, Kaniarasu, Medvedev, Steinfeld, & Yanco, 2013).

We attempted to ensure the internal validity of the experiments by randomly selecting subjects, using control groups, addressing potential confounding variables, and limiting the potential for experimenter bias. The external validity of our results, however, is limited. Because the experimental environment was simulated and participants completed the task online, the study lacked the true visceral reaction of a real emergency. Additional

real-world studies are thus warranted. Further, although we tried to use a variety of different situations and contexts, fully capturing the many different environments in which trust occurs was impossible. The conclusions that we draw are necessarily based on a limited number of situations. Finally, although the study's participants were from a broad cross-section of the United States, they still represent a limited population. Hence, we are not claiming that the results represent a general truism about trust, rather, only that they serve as evidence of the connection and validity of our approach for conceptualizing and reasoning about trust with respect to outcome matrices.

In spite of these limitations, our technique is well suited for implementation and use on a robot. More importantly, however, this research should be viewed as one piece of a larger framework. This larger framework develops a means for reasoning about trust regardless of whether the robot assumes the role of trustor or of trustee, inherently includes a method for using stereotypes to bootstrap trust evaluations, naturally adjusts to factors such as experience and reputation, and can potentially be used to relate other social phenomena such as deception and fairness to trust (Wagner, 2009). Unfortunately, lack of space prevents us from addressing all of these aspects of trust. As a validation of our underlying conceptualization of trust, the work presented here potentially represents an important step towards developing a robot that can reason about whether or not it should trust a person or whether a person trusts it.

Our future research will explore ways to use these results to create robots that actively repair trust and prevent people from trusting a system too much. We are developing a number of strategies for trust repair that we are currently in the process of validating. Our future experiments will use these strategies in an evacuation scenario

involving real robots and people. We also plan to explore techniques that prevent people from trusting a robot too much. We believe that robots can be developed which recognize and warn a person when an unsafe level of trust is occurring.

Given the important role that autonomous systems will play in the future, and the risks that people will place in these machines, devising methods that effectively manage the trust between a person and a robot will be critical.

Acknowledgment

This work was funded by award #FA95501310169 from the Air Force Office of Sponsored Research.

References

Adolphs, R., Tranel, D., & Damasio, A. R. (1998). The human amygdala in social judgment. *Nature, 393*, 470-474.

Barber, B. (1983). *The Logic and Limits of Trust* . New Brunswick, New Jersey: Rutgers University Press.

Castelfranch, C., & Falcone, R. (2010). *Trust Theory: A Socio-Cognitive and Computational Model.* New York, NY: Wiley Publishers.

Cooper, J. C., Kreps, T. A., Wiebe, T., Pirkl, T., & Knutson, B. (2010). When Giving Is Good: Ventromedial Prefrontal Cortex Activation for Others' Intentions. *Neuron, 67*(3), 511–521.

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, (pp. 251-258). Tokyo, Japan.

Deutsch, M. (1962). Cooperation and Trust: Some Theoretical Notes. In M. R. Jones, *Nebraska Symposium on Motivation* (pp. 275-315). Lincoln, NB: University of Nebraska.

Deutsch, M. (1973). *The Resolution of Conflict: Constructive and Destructive Processes.* New Haven, CT: Yale University Press.

Economist. (2006). Trust me, I'm a robot. *The Economist*, pp. 18-19.

Engle-Warnick, J., & Slonim, R. L. (2006). Learning to trust in indefinitely repeated games. *Games and Economic Behavior*, 95-114.

Fisher, R. J. (1993). Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research, 20*(2), 303-315.

Gambetta, D. (1990). Can We Trust Trust? In D. Gambetta (Ed.), *Trust, Making and Breaking Cooperative Relationships* (pp. 213-237). Oxford England: Basil Blackwell.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., Visser, E. J., & Parasuraman, R. (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), 517-527.

Hemphill, J. F. (2003). Interpreting the Magnitudes of Correlation Coefficients. *American Psychologist, 58*(1), 78-80.
Hung, Y. C., Dennis, A. R., & Robert, L. (2004). Trust in Virtual Teams: Towards an Integrative Model of Trust Formation. *International Conference on System Sciences.* Hawaii.
Josang, A., & Pope, S. (2005). Semantic Constraints for Trust Transitivity. *Second Asia-Pacific Conference on Conceptual Modeling.* Newcastle, Australia.
Kelley, H. H., & Thibaut, J. W. (1978). *Interpersonal Relations: A Theory of Interdependence.* New York, NY: John Wiley & Sons.
Kelley, H. H., Holmes, J. G., Kerr, N. L., Reis, H. T., Rusbult, C. E., & Lange, P. A. (2003). *An Atlas of Interpersonal Situations.* New York, NY: Cambridge University Press.
King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to Know You: Reputation and Trust in Two-Person Economic Exchange . *Science*(308), 78-83.
Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, pp. 50-80.
Luhmann, N. (1979). *Trust and Power.* Chichester: Wiley Publishers.
Luna-Reyes, L., Cresswell, A. M., & Richardson, G. P. (2004). Knowledge and the Development of Interpersonal Trust: a Dynamic Model. *International Conference on System Science.* Hawai.
Mitchell, M., & Jolley, J. (1992). *Research Design Explained* (2nd Edition ed.). Orlando, FL: Harcourt Brace Jovanovich.
Osborne, M. J., & Rubinstein, A. (1994). *A Course in Game Theory.* Cambridge, MA: MIT Press.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*(5), 411-419.
Prietula, M. J., & Carley, K. M. (2001). Boundedly Rational and Emotional Agents. In C. Castelfranchi, & Y.-H. Tan, *Trust and Deception in Virtual Society* (pp. 169-194). Kluwer Academic Publishers.
Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A Neural Basis for Social Cooperation. *Neuron*, 395-405.
Robinette, P., Wagner, A. R., & Howard, A. (2013). Building and Maintaining Trust Between Humans and Guidance Robots in an Emergency. *AAAI Spring Symposium, Stanford University*, (pp. 78-83). Palo Alto.
Robinette, P., Wagner, A. R., & Howard, A. (2014). Assessment of Robot Guidance Modalities Conveying Instructions to Humans in Emergency Situations. *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 14).* Edinburgh, UK.
Robinette, P., Wagner, A. R., & Howard, A. (2014). The Effect of Robot Performance on Human-Robot Trust in Time-Critical Situation. *Human Factors*, under review.
Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so Different After All: A Cross-Discipline View of Trust. *Academy of Management Review, 23*, 393-404.
Runyon, R. P., & Audrey, H. (1991). *Fundamentals of Behavioral Statistics.* New York, NY: McGraw Hill.
Sabater, J., & Sierra, C. (2005). Review of Computational Trust and Reputation Models. *Artificial Intelligence Review, 24*, 33-60.
Schillo, M., Funk, P., & Rovatsos, M. (2000). Using Trust for Detecting Deceitful Agents in Artificial Societies. *Applied Artificial Intelligence Journal, Special Issue on Trust, Deception and Fraud in Agent Societies, .*
Sears, D. O., Peplau, L. A., & Taylor, S. E. (1991). *Social Psychology.* Englewood Cliffs, New Jersey: Prentice Hall.
Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology*, 491–517.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1130.
Wagner, A. R. (2009). Creating and Using Matrix Representations of Social Interaction. *Proceedings of the 4th International Conference on Human-Robot Interaction (HRI 2009).* San Diego, CA.
Wagner, A. R. (2009). *The Role of Trust and Relationships in Human-Robot Social Interaction.* Ph.D. diss., School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA.
Wagner, A. R. (2012). Using Cluster-based Stereotyping to Foster Human-Robot Cooperation. *Proceedings of IEEE International Conference on Intelligent Robots and Systems (IROS 2012)*, (pp. 1615-1622). Villamura, Portugal. doi:10.1109/IROS.2012.6385704

Wagner, A. R. (2013). Developing Robots that Recognize when they are being Trusted. *AAAI Spring Symposium.* Stanford CA.
Wagner, A. R., & Arkin, R. C. (2011). Acting Deceptively: Providing Robots with the Capacity for Deception. *The International Journal of Social Robotics, 3*, 5-26.
Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Nature Neuroscience. *5*, 277 - 283.
Yamagishi, T. (2001). Trust as a Form of Social Intelligence. In K. S. Cook, *Trust in Society.* New York, NY: Russell Sage Foundation.