# Recognizing Situations that Demand Trust

Alan R. Wagner, *Member, IEEE*, Ronald C. Arkin, *Fellow, IEEE*

***Abstract*—This article presents an investigation into the theoretical and computational aspects of trust as applied to robots. It begins with an in-depth review of the trust literature in search of a definition for trust suitable for implementation on a robot. Next we apply the definition to our interdependence framework for social action selection and develop an algorithm for determining if an interaction demands trust on the part of the robot. Finally, we apply our algorithm to several canonical social situations and review the resulting indications of whether or not the situation demands trust.**

## I. Introduction

TRUST. The term itself conjures vague notions of loving relationships and lifelong familial bonds. But is trust really so indefinable? The phenomenon of trust has been seriously explored by numerous researchers for decades. Moreover, the notion of trust is not limited to interpersonal interaction. Rather, trust underlies the interactions of employers with their employees, banks with their customers, and governments with their citizens. In many ways trust is a precursor to a great deal of normal interpersonal interaction.

For interactions involving humans and robots, 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 occurred because of a robot's actions [1]. A great deal of research is currently focused on bringing robots out of labs and into people's homes and workplaces. These robots will interact with humans—such as children and the elderly—unfamiliar with the limitations of a robot. It is therefore critical that human-robot interaction research explore the topic of trust.

The work presented here builds upon our prior research developing an interdependence framework for social action selection by a robot [2]. Our framework (discussed in greater detail in the next section) uses a matrix representation to represent the robot's social interactions. These outcome matrices contain values relating to the robot's and the human's reward with respect to the selection of particular actions. Interdependence theorists have shown that these outcome matrices can be used to represent any interaction [3]. In contrast to much of the prior work on trust, the research presented here does not begin with a model for trust. Rather, it begins with a very simple hypothesis: *if it is true that outcome matrices serve as a representation for interaction, and that all interactions can be represented as an outcome matrix, then should it not also be true that some outcome matrices include trust while others do not?* If so, then one should be able to delineate conditions segregating those outcome matrices that require trust from those that do not. The task then becomes one of determining what the conditions for trust are.

This article presents an investigation into the theoretical and computational aspects of trust as applied to robots. It begins with an in-depth review of the trust literature in search of a definition for trust suitable for implementation on a robot. Next this definition is applied to our interdependence framework for social action selection and developed into an algorithm for determining if an interaction demands trust on the part of the robot. Finally, the algorithm is applied to several canonical social situations and the results reviewed for indications of whether or not the situation demands trust.

The algorithm we describe has been implemented and tested on embodied robots [4]. Because of space considerations, it was not possible to both thoroughly detail our lengthy investigation of the phenomena of trust and its theoretical underpinnings and to also present our experiments involving robots. The results from these experiments will be presented in a subsequent paper.

## II. Trust: A brief Review

### A. Defining Trust

Early trust research focused on definitions and characterizations of the phenomenon. Morton Deutsch is widely recognized as one of the first researchers to study trust [5]. Deutsch, a psychologist, describes trust as a facet of human personality [6]. He claims that trust is the result of a choice among behaviors in a specific situation. Deutsch's definition of trust focused on the individual's perception of the situation and the cost/benefit analysis that resulted. He also proposes the existence of different types of trust. Other types include trust as despair, innocence, impulsiveness, virtue, masochism, faith, risk-taking, and confidence [7]; see [5] for an overview).

Niklas Luhmann, another early trust researcher, provides a sociological perspective [8]. Luhmann defines trust as a means for reducing the social complexity and risk of daily life. He argues that the complexity of the natural world is far too great for an individual to manage the many decisions it must make in order to survive. Because a trusting society has greater capacity for managing complexity, it can afford to be more flexible in terms of actions and experience.

Manuscript received March 3, 2011.

Alan R Wagner is a research scientist with Georgia Institute of Technology Research Institute, Atlanta, GA 30332 USA (corresponding author phone: 404-407-6522; e-mail: alan.wagner@ gtri.gatech.edu).

Ronald C. Arkin is a Regents' Professor and Associate Dean in College of Computing at Georgia Institute of Technology, Atlanta, GA 30332 USA. (phone: 404-894-8209; e-mail: arkin@gatech.edu).

Bernard Barber, another sociologist, defines trust as an expectation or mental attitude an agent maintains regarding its social environment [9]; see [5] for an overview). He claims that trust results from learning in a social system and is used by an individual to manage its expectations regarding its relationships and social environment. Hence, trust is an aspect of all social relationships and is used as a means of prediction for the individual.

Gambetta describes trust as a probability [10]. 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" [10] p. 216). Gambetta defines trust as a probabilistic assessment of another agent's intent to perform an action on which the agent will rely.

Rousseau et al. have examined the definitional differences of trust from a variety of sources [11] and concluded that trust researchers generally agree on the conditions necessary for trust, namely risk and interdependence.

Lee and See consider trust from the perspective of machine automation, providing an extremely insightful and thorough review of the trust literature [12]. They review many definitions of trust and propose a definition that is a compilation of the many previous definitions. Namely, trust is *the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability*. This research uses Lee and See's definition of trust to generate a more conceptually precise and operational description of trust. We define trust in terms of two individuals—a trustor and a trustee. The trustor is the individual doing the trusting. The trustee represents the individual in which trust is placed. Based on the definitions and descriptions above, trust is defined 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.***

Like the work of Luhmann and Gambetta, the presence of risk is addressed. As with Deutsch, the proposed definition also makes note of the trustor's choice of action. The term belief is used, instead of attitude, because agent belief has been operationalized as the probability of a truth statement in related literature and because we feel that difference is inconsequential to the definition [10].

### *B. Models and Measures of Trust*

Researchers have generated many different computational models of trust. This research approaches the problem of modeling trust by showing that our interdependence framework for social action selection implicitly contains mechanisms for determining if trust is necessary for a given social situation. Before detailing our framework several different methods for modeling and measuring trust are described.

King-Casas et al. used a situation in which two human players iteratively interact for ten rounds exchanging money as an investor and as a trustee [13]. In each round the investor selects some proportion of money to invest ($I$) with the trustee. The money appreciates ($3I = R$). Finally the trustee repays a self determined proportion of the total amount ($R$) back to the investor. King-Casas et al. found previous reciprocity to be the best predictor of changes in trust for both the investor and trustee ($\rho = 0.56; \rho = 0.31$ respectively where $\rho$ is the correlation coefficient) [13]. Hence, by measuring quantities of reciprocity, trust is operationalized as monetary exchange in a way that allows for online analysis of the relationship from its inception. The work by King-Casas et al. puts the trustor at monetary risk. Yet, if previous reciprocity indicates the presence of trust then the trustor expects that the trustee will act in a manner that mitigates his or her risk, and both parties benefit from mutual trust.

Marsh defines trust in terms of utility for a rational agent [5]. Further, Marsh recognizes the importance of the situation and includes this factor in his formulation of trust. He estimates trust as, $T_X(y,\alpha) = U_x(\alpha) \cdot I_x(\alpha) \cdot \hat{T}_x(y)$ where $T_X(y,\alpha)$ is $x$'s trust in $y$ for situation $\alpha$, $U_x(\alpha)$ is the utility of $\alpha$ for $x$, $I_x(\alpha)$ is the importance of $\alpha$ for $x$, and $\hat{T}_x(y)$ is the general trust of $x$ in $y$. Marsh notes many weaknesses, flaws, and inconsistencies in this formulation. For example, he states the value range he has chosen for trust, $[-1, +1)$, presents problems when trust is zero.

Recently a trend in trust research has been to focus on the use of probability theory to measure and model trust. Josang and Lo Presti use probabilities to represent an agent's assessment of risk [14]. They describe an agent's decision surface with respect to risk as $F_C(p, G_S) = p^{\frac{\lambda}{G_S}}$ where $C$ is the agent's total social capital[2], $F_C \in [0,1]$ is the fraction of the agent's capital it is willing to invest in a single transaction with another agent, $p$ is the probability that the transaction will end favorably, $G_S$ is gain resulting from the transaction and $\lambda \in [0, \infty]$ is a factor used to moderate the gain $G_S$. Josang and Lo Presti define reliability trust as the value of $p$ and decision trust as $T = \begin{cases} \frac{p-p_D}{p_D} & p < p_D \\ 0 & \text{for } p = p_D \\ \frac{p-p_D}{1-p_D} & p > p_D \end{cases}$ where $p_D$ is a cut-off probability. Josang and Pope later use this model of trust to propagate trust and reputation information for the purpose of developing a secure network cluster [14-16]. Beth et al. also use probability for the purpose of developing trust in network security claiming that the equation $v_Z(p) = 1 - \alpha^p$, where $p$ is the number of positive experiences and $\alpha$ is chosen to be a value high enough to produce confident estimations should be used to measure trust [17].

Castelfranchi and Falcone have been strong critics of defining trust in terms of probability because they feel this description of trust is too simplistic [18]. Rather, they describe a cognitive model of trust that rests on an agent's mental state. This mental state is in turn controlled by an

[2] Social capital is concept from economics used to describe the value of the connections within a social network.

agent's beliefs with respect to the other agent and an agent's own goals [19, 20].

Researchers have also explored the role of trust in machine automation. Trust in automation researchers are primarily concerned with creating automation that will allow users to develop the proper level of trust in the system. Lee and See note that one fundamental difference between trust in automation research and intrapersonal trust research is that automation lacks intentionality [12]. Another fundamental difference is that human-automation relationships tend to be asymmetric with the human deciding how much to trust the automation but not vice versa.

Researchers have explored many different methods for measuring and modeling trust. Trust measures have been derived from information withholding (deceit) [21], agent reliability [22, 23], agent opinion based on deceitful actions [16], compliance with virtual social norms [24], and compliance with an a priori set of trusted behaviors from a case study [25]. Models of trust range from beta probability distributions over agent reliability [16], to knowledge-based formulas for trust [25], to perception-specific process models for trust [24].

Often these measures and models of trust are tailored to the researcher's particular domain of investigation. Luna-Reyes et al., for example, derive their model from a longitudinal case study of an interorganizational information technology project in New York State [25]. This model is then tested to ensure that it behaves in a manner that intuitively reflects the phenomena of trust. A review of computational trust and reputation models by Sabater and Sierra states, "… 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 [26]."

The alternative methods for evaluating trust discussed in this section highlight a diversity of approaches and domains the topic of trust touches on. As will be shown, rather than creating another computational model of trust, the definition of trust above can be used in conjunction with our interdependence framework to determine if a social situation demands trust. In the section that follows our framework for social action selection is briefly described.

## III. Representing Interaction

Social psychologists define *social interaction* as influence—verbal, physical, or emotional—by one individual on another [27]. Representations for interaction have a long history in social psychology and game theory [28, 29]. Interdependence theory, a type of social exchange theory, is a psychological theory developed as a means for understanding and analyzing interpersonal situations and interaction [29]. The term interdependence specifies the extent to which one individual of a dyad influences the other. Interdependence theory is based on the claim that people adjust their interactive behavior in response to their perception of a social situation's pattern of rewards and costs. Thus, each choice of interactive behavior by an individual offers the possibility of specific rewards and costs—also known as outcomes—after the interaction. Interdependence theory represents interaction and social situations computationally as an outcome matrix (Figure 1). An outcome matrix represents an interaction by expressing the outcomes afforded to each interacting individual with respect each pair of potential behaviors chosen by the individuals.

**Example Outcome Matrices**

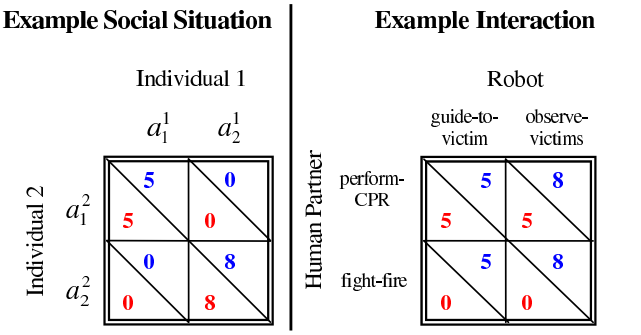


Figure 1 Example outcome matrices are depicted above. The right hand side depicts an outcome matrix representing a interaction between a robot and a human partner in a search and rescue paradigm. The left hand side depicts a social situation. Social situations abstractly represent interactions.

Game theory also explores interaction [28]. As a branch of applied mathematics, game theory thus focuses on the formal consideration of strategic interactions, such as the existence of equilibriums and economic applications [28]. Game theory and interdependence theory both use the outcome matrix to represent interaction [28, 29]. Game theory, however, is limited by several assumptions, namely: both individuals are assumed to be outcome maximizing; to have complete knowledge of the game including the numbers and types of individuals and each individual's payoffs; and each individual's payoffs are assumed to be fixed throughout the game. Because it assumes that individuals are outcome maximizing, game theory can be used to determine which actions are optimal and will result in an equilibrium of outcome. Interdependence theory does not make these assumptions and does not lend itself to analysis by equilibrium of outcomes.

The outcome matrix is a standard computational representation for interaction [29]. The term interaction describes a discrete event in which two or more individuals select interactive behaviors as part of a social situation or social environment. The term individual is used to indicate either a human, a social robot, or an agent. This research focuses on interaction involving two individuals—dyadic interaction.

Because outcome matrices are computational representations, it is possible to describe them formally. The notation presented here draws heavily from game theory [28]. A representation of interaction consists 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 obtained by each individual for each combination of actions that could have been selected [29]. Let $a_j^i \in A^i$ be an arbitrary action $j$ from

individual *i*'s set of actions. Let $(a_j^1, \ldots, a_k^N)$ denote a combination of actions, one for each individual, and let $u^i$ denote individual *i*'s utility function: $u^i(a_j^1, \ldots, a_k^N) \rightarrow \mathcal{R}$ is the utility received by individual *i* if the individuals choose the actions $(a_j^1, \ldots, a_k^N)$. The term $O$ is used to denote an outcome 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.

## IV. Recognizing Situations that Require Trust

The task of delineating the conditions for trust begins with our working definition of trust. Recall that, trust was defined 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 trustee has put its outcomes at risk.***

Recall that social situations abstractly represent a class of interactions. This section develops conditions for classifying a situation in terms of trust. Classification of a situation in terms of trust is a binary task, i.e. a true/false statement concerning whether or not the selection of an action in a situation would require trust.

Figure 2 An example of the trust fall. The trust fall is a trust and team-building exercise in which one individual, the trustor, leans back prepared to fall to the ground. Another individual, the trustee, catches the first individual. The exercise builds trust because the trustor puts himself at risk expecting that the trustee will break her fall.

Consider, for example, the trust fall. The trust fall is a game played in an attempt to build trust between two or more people. One person simply leans backward and falls into the awaiting arms of another person (Figure 2). The trust fall will be used as a running example to explain our conditions for trust.

Assume that the trust fall involves two people. The person leaning back acts as the trustor, whereas the person doing the catching represents the trustee. The trustor decides between two potential actions in the interaction: lean back and do not lean back. The trustee also decides between two potential actions: catch the falling person and do not catch the falling person. Hence the interaction can be represented as a 2x2 outcome matrix (Figure 3). In this interaction the trustor holds the belief that the trustee will break their fall before they hit the ground. Moreover, the action of leaning back puts the trustor at risk of possible injury. The actual result of the interaction depends on the actions of the trustee. Their choice of action can result in injury or no injury to the trustor. As described, the situation implies a specific pattern of outcome values.

The definition for trust listed above focuses on the actions of the trustor and trustee. These individuals can be arbitrarily listed as the interacting individuals in an outcome matrix (Figure 3). Without loss of generality, our discussion of the decision problem will be limited to two actions ($a_1^i$ and $a_2^i$ for the trustor, $a_1^{-i}$ and $a_2^{-i}$ for the trustee). Let $a_1^i$ arbitrarily labeled as the trusting action and $a_2^i$ as the untrusting action for the trustor. Similarly, for the trustee the action $a_1^{-i}$ arbitrarily denotes the action which maintains trust and the action $a_2^i$ the action which does not maintain trust. 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 temporal order is described with the condition in outcome matrix notation as $i \Longrightarrow -i$ indicating that individual *i* acts before individual *-i*.

The definition for trust notes that risk is an important consideration for the trustor. Risk refers to a potential loss of outcome. The occurrence of risk implies that the outcome values received by the trustor depend on the action of the trustee. Our second condition notes this dependence relation by stating that *2) the outcome received by the trustor depends on the actions of the trustee if and only if the trustor selects the trusting action.* Recall that $o^i$ denotes the outcome received by the trustor. If the trustor selects the trusting action then the outcomes ${}_{11}o^i$ and ${}_{21}o^i$ from Figure 3 are being compared. The statement indicates that there will be a difference, ${}_{11}o^i - {}_{21}o^i > \varepsilon_1$, where $\varepsilon_1$ is a constant representing the minimal amount of outcome for dependence.

The trustor may also select the untrusting action, however. The existence of the untrusting action implies that this action does require risk on the part of the trustor. In other words, the outcome $({}_{x2}o^i)$ received by the trustor when selecting the untrusting action does not depend on the actions of the trustee. This leads to a third condition, *3) the outcome received when selecting the untrusting action does not depend of the actions of the trustee.* In other words, the outcomes for action $a_2^i$ do not depend on the action selected by the trustee. Stated formally, $\left|{}_{12}o^i - {}_{22}o^i\right| < \varepsilon_2$, where $\varepsilon_2$ is a constant representing the maximal amount of outcome for independence.

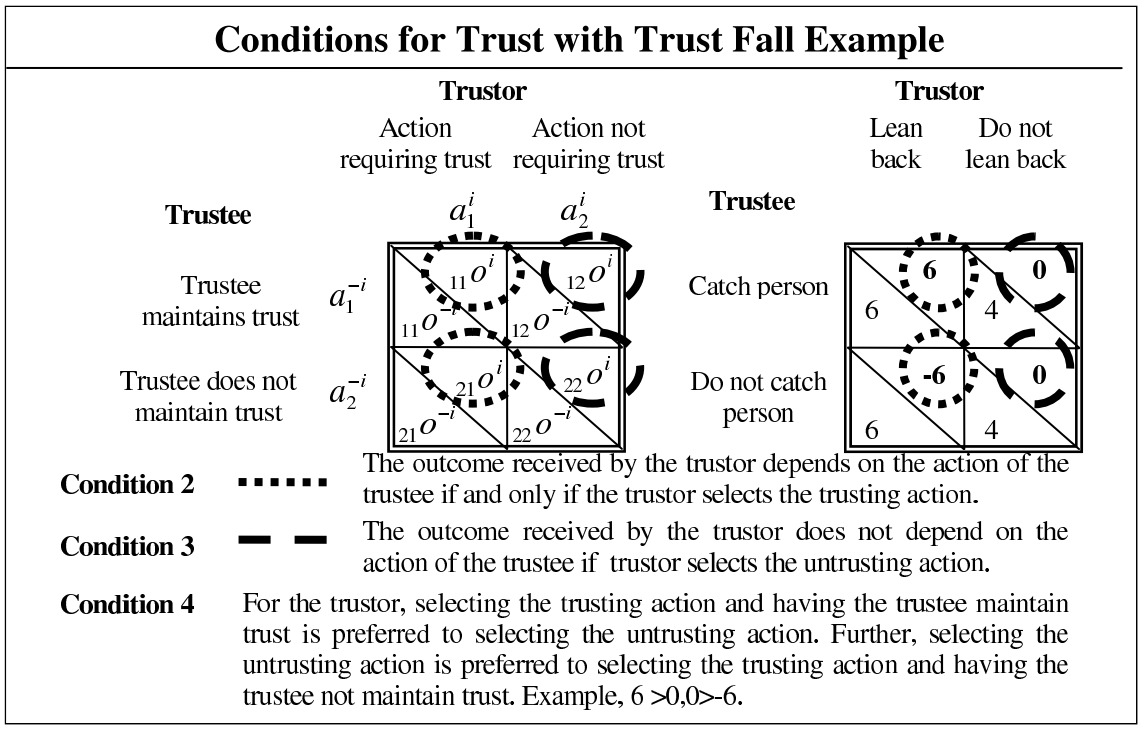


Figure 3 The figure visually depicts the reasoning behind the development of the conditions for trust. The matrix on the left depicts the conditions for trust on an abstract social situation. The matrix to the right presents the conditions for the trust fall example. Conditions two, three and four are depicted in the example. Conditions one and five are not depicted.

The definition for trust implies a specific pattern of outcome values. It indicates that 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 mitigate the trustor's risk 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 (the trustee acts in manner which mitigates the risk) is greater than the value of not trusting at all, is greater than the value of having one's trust broken.* Again described formally, the outcomes are valued ${}_{11}o^i > {}_{x2}o^i > {}_{21}o^i$.

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.

The preceding conditions are necessary for a trusting interaction. Sufficiency occurs when these conditions are met and the trustor selects action $a_1^i$. The first four conditions describe the **situational conditions** necessary for trust. By testing a situation for these conditions one can determine whether or not an interactive situation requires trust. Figure 4 presents an algorithm for determining if a putative situation requires trust.

The trust fall is used as an example in Figure 3 to delineate the conditions for trust. The first condition will be assumed to be true for this example. The second condition results in values (from the matrix) $6 - (-6) > \varepsilon_1$. Thus, action $a_1^i$ does depend on the actions of the partner for constant $\varepsilon_1 < 12$. The values assigned to the constants $\varepsilon_1, \varepsilon_2, k$ are likely to be domain specific. The constant $\varepsilon_1$ represents a threshold for the amount of risk associated with the trusting action. The constant $\varepsilon_2$, on the other hand, represents a threshold for the lack of risk associated with the untrusting action. The third condition results in values, $|0 - 0| < \varepsilon_2$. Here, the action $a_2^i$ does not depend on the actions of the partner for constant $\varepsilon_2 \geq 0$. The final condition results in values $6 > \{0,0\} > -6$. Hence, for individual one, the selection of action $a_1^i$ involves risk that can be mitigated by the actions of the partner and the selection of action $a_2^i$ does not involve risk that is mitigated by the actions of the partner.

## Testing for Situational Trust

**Input**: Outcome matrix *O*
**Assumptions**: Individual *i* is trustor, individual *-i* is trustee, is the trusting action, is not a trusting action.
**Output**: Boolean stating if *O* requires trust on the part of individual *i*.

1. **If** $i \Rightarrow -i$ is false //the trustee does not act before
   **return false** //the trustor
2. **If** ${}_{11}o^i - {}_{21}o^i < \varepsilon_1$ //the trustor's outcome must
   **return false** //depend on the action of trustee
   // when selecting the trusting action
3. **If** $|{}_{12}o^i - {}_{22}o^i| > \varepsilon_2$ //the trustor's outcome must not the
   **return false** //depend on the action of the trustee
   //when selecting the untrusting action
4. **If** ${}_{11}o^i > {}_{x2}o^i > {}_{21}o^i$ is false //the value of fulfilled trust
   **return false** //is greater than the value of not
   **Else** //trusting at all, is greater than the value
   **return true** //of having one's trust broken

Figure 4 The algorithm above depicts a method for determining whether a social situation requires trust. The algorithm assumes that the first individual is the trustor, the second individual is the trustee, the action $a_1^i$ is the trusting action, and the action $a_2^i$ is not a trusting action.

## V. COMMON SITUATIONS AND THE CONDITIONS FOR TRUST

In this section those situations which meet the conditions for trust are qualitatively compared to those which do not. The goal is to demonstrate that the situations selected by the algorithm as demanding trust intuitively match those situations in which humans use trust. Additionally, we strive to show that situations which are typically not considered to demand trust are also deemed to not require trust by the algorithm. It is suspected that additional experiments involving human subjects will be necessary to provide more conclusive evidence that our algorithm does indeed classify situations similarly to humans with respect to trust.

Kelley et al. recently published an atlas of social situations based interdependence theory's interdependence space [3]. The atlas describes several canonical social situations as well as each situation's characteristics. Five situations listed in Kelly et al.'s atlas of social situations were selected as input to our algorithm. Table 1 lists these five social situations. The situations were selected because they represent different areas within the interdependence space. Each situation was used as input to the algorithm in Figure 4. The values for constants were arbitrarily set at $\varepsilon_1 = 6$ and $\varepsilon_2 = 6$. The independent variable is the situations selected for testing. The dependent variable then is the determination of whether or not the situation demands trust.

The results are listed in the rightmost column of Table 1. This column states whether or not the algorithm indicates that the situation demands trust on the part of the trustor. The trustor is assumed to be the individual depicted on the top of the matrix. The trusting action is assumed to be located in the first column of each matrix. These assumptions are simply for clarity of exposition and do not limit the algorithm.

For example consider the Cooperative Situation, the first row from Table 1. The outcome matrix for the situation is used as input to the algorithm. The first line in the algorithm is assumed to be true. The second line of the algorithm calculates ${}_{11}o^i - {}_{21}o^i > \varepsilon_1$ as $13 - 6 > 6$. Hence the second condition for situational trust is true. The third line of the algorithm calculates $\left|{}_{12}o^i - {}_{22}o^i\right| < \varepsilon_2$ as $|6 - 6| < 6$. This third condition for situational trust is also found to be true. Finally, the forth line of the algorithm computes ${}_{11}o^i > {}_{x2}o^i > {}_{21}o^i$ to be $13 > \{6,6\} > 6$ which is false. The final output of the algorithm for this situation is false.

Details for each of the situations examined follows:

1. The Cooperative situation describes a social situation in which both individuals interact cooperatively in order to receive maximal outcomes. Given the algorithm's parameters, the trustor faces a situation in which the trusting action is dependent on the trustee. The untrusting action, in contrast, is not dependent on the trustee. Nevertheless, the trustor stands to lose nothing if the trustee does not maintain trust (6 versus 6). Hence, selection of the trusting action does not involve risk as the trustor stands to minimally gain as much by selecting this action as by selecting the untrusting action. The algorithm therefore returns false because the situation does not meet all of the conditions for trust.

Table 1 Several situations from Kelley et al.'s atlas of social situations are depicted below. The table includes a description of the situation and the situation's outcome matrix. The first condition the algorithm in Figure 4 is assumed to hold for all situations. Columns 3-5 present the results for the remaining conditions. The right most column presents the algorithm's final output, stating whether or not the situation demands trust.

**Social Situations for Qualitative Comparison**

| Situation | Matrix | Cond. 2 | Cond. 3 | Cond. 4 | Sit. Trust? |
|---|---|---|---|---|---|
| Cooperative Situation | 13,12 \| 6,6<br>6,6 \| 6,0 | True | True | False | No |
| Competitive Situation | 6,6 \| 12,0<br>0,12 \| 6,6 | False | False | False | No |
| Trust Situation | 12,12 \| 8,0<br>0,8 \| 4,4 | True | True | True | Yes |
| Prisoner's Dilemma | 8,8 \| 12,0<br>0,12 \| 4,4 | True | False | False | No |
| Chicken Situation | 12,4 \| 8,8<br>0,0 \| 4,12 | True | True | True | Yes |

2. The Competitive situation also does not demand trust, but for different reasons. In this situation the trusting and untrusting actions afford equal risk. Thus the trustor does not face a decision problem in which it can select an action that will mitigate its risk. Rather, the trustor's decision problem is simply of a matter of selecting the action with the largest guaranteed outcome. Trust is unnecessary because the trustor's decision problem can be solved without any consideration of the trustee's beliefs and actions.

3. The Trust Situation describes a situation in which mutual cooperation is in the best interests of both individuals. As the name would portend, this situation demands trust. The trustor's outcomes are dependent on the action of the trustee if it selects the trusting action. Further, nominal outcomes are risked when selecting untrusting action. Finally, the trustor stands to gain the most if it selects the trusting action and the trustee maintains the trust. The trustor's second best option is not to trust the trustee. Finally, the trustor's worst option is to select the trusting action and to have the trustee violate that trust.

4. The Prisoner's Dilemma is perhaps the most extensively studied of all social situations [30]. In this situation, both individuals depend upon one another and are also in conflict. Moreover, selection of the trusting action by the trustor does place outcomes at risk dependent on the action of the trustee. Given the parameters selected, however, the untrusting action is also critically dependent on the action of the trustee. Hence, both actions require some degree of risk on the part of the trustor. Our conditions for situational trust demand that the decision problem faced by

the trustor offer the potential for selecting a less risky action. As instantiated in Table 1, this version of the prisoner's dilemma does not offer a less risky option. Note, however, that by changing one of the trustor outcomes, say 8 to 9, and the algorithm's constants to $\varepsilon_1 = 8$ and $\varepsilon_2 = 9$ the situation does then demand situational trust. Overall, the prisoner's dilemma is a borderline case in which the specific values of the outcomes determine whether or not the situation demands trust. This social situation raises important issues that will be discussed in the conclusion.

5. The Chicken situation is a prototypical social situation encountered by people. In this situation each interacting individual chooses between safe actions with intermediate outcomes or more risky actions with more middling outcomes. An example might be the negotiation of a contract for a home or some other purchase. This situation, like the Trust Situation, demands trust because it follows the same pattern of risks as the Trust Situation.

Overall, this analysis provides some intuitive evidence that our algorithm does correctly classify situations with respect to trust. A more comprehensive test of the algorithm's accuracy will require the use of a human subject pool, placing them in imaginary or virtual social situations, asking them if they believe that the situation or the selection of an action in a situation demands trust and then applying the algorithm in the same situation. We are currently setting up the infrastructure to perform this experiment.

## VI. Summary and Conclusions

This article has focused on the theoretical and computation underpinnings necessary for developing a robot with the capacity to recognize social situations in which it must trust the actions of a person or a person is trusting it. It began with a thorough interdisciplinary investigation of the trust literature which resulted in an operationalized definition of trust. Next it was hypothesized that the definition could be used to develop a series of conditions which could then segregate those social situations demanding trust from situations that do not demand trust. Finally an algorithm was developed based on these conditions and then tested on several well known social situations.

This work is important for several reasons. First, it addresses an important area of robotics and human-robot interaction research. Second, the approach taken does not demand artificial or ad hoc models of trust developed for a particular problem or paradigm. Our definition and algorithm is completely general and not impacted by the characteristics of the human, the robot, or the problem being explored. This work highlights theoretical and conceptual aspects of trust that have not been explored in previous work—most notably the role of interdependence in social situations involving trust. Finally, the algorithm and concepts expounded here should hold regardless of whether the robot is the trustor or the trustee. Hence, the robot should be able to determine if it is trusting someone else or if someone else is trusting it.

This research has hinted at another possible type of trust—forced trust. Forced trust is defined here as a situation in which the trustor has only trusting actions from which to choose. As mentioned in the previous section, our algorithm indicated that the outcome matrix representing the Prisoner's Dilemma did not demand trust because all actions available to the trustee presented risk. Hence, the decision faced by the trustor is one in which they are forced to trust the trustee. This result hints that, as argued by Deustch, there may be multiple types of trust with differing characteristics [7]. Colloquially the term trust may apply to a single action choice. Hence, in situations in which the trustor's action space is greater than two actions, one often describes trust with respect to the selection of a particular risky action among many potential actions. In force trust situations, then describe situations in which all of the trustor's actions present risk mitigated by the actions of the trustee.

This article has largely been silent as to the decision problem faced by the trustee. Situations in which the trustee has a large incentive to maintain the trust or to violate the trust tend to influence one's measure of trust by mitigating or enhancing the trustor's risk [10, 13, 14]. Nevertheless, these situations still demand trust because of the existence of risk and the trustor's dependence on the trustee. The techniques described in this paper have been used to formulate a measure of trust which takes into consideration the decision problem faced by the trustee [4].

The challenges of creating matrices from the perceptual information available to a robot have not discussed. These challenges have been addressed in our previous research [4].

We believe that applications of the algorithm to human-robot interaction problems constitute an important area of research. Further, numerous aspects of our interdependence framework for social action selection have been verified on embodied robots [4].

Our continuing work in this area centers on demonstrating these techniques on robots in interactive situations with humans. We are investigating both situations in which the robot is the trustee and the trustor. We are also empirically verifying our definition for trust and exploring the conjecture pertaining to forced trust. Finally, we are attempting to identify demonstrable applications in the domain of home healthcare and military robotics. Because trust underlies some much of normal interpersonal interaction, we believe that a principled approach will be critical to the success of creating trusting and trustworthy robots.

## References


[1] Economist, "Trust me, I'm a robot," in *The Economist*, vol. 379, 2006, pp. 18-19.

[2] A. R. Wagner and R. C. Arkin, "Representing and analyzing social situations for human-robot interaction," 2006.

[3] H. H. Kelley, J. G. Holmes, N. L. Kerr, H. T. Reis, C. E. Rusbult, and P. A. M. V. Lange, *An Atlas of Interpersonal Situations*. New York, NY: Cambridge University Press, 2003.

[4] A. R. Wagner, "The Role of Trust and Relationships in Human-Robot Social Interaction," in *College of Computing*, vol. Doctoral. Atlanta, GA: Georgia Institute of Technology, 2009, pp. 254.

[5] S. Marsh, "Formalising Trust as a Computational Concept," in *Computer Science*, vol. PhD: University of Stirling, 1994, pp. 170.
[6] M. Deutsch, "Cooperation and Trust: Some Theoretical Notes," in *Nebraska Symposium on Motivation*, M. R. Jones, Ed. Lincoln, NB: University of Nebraska, 1962, pp. 275-315.
[7] M. Deutsch, *The Resolution of Conflict: Constructive and Destructive Processes*. New Haven, CT: Yale University Press, 1973.
[8] N. Luhmann, *Trust and Power*. Chichester: Wiley Publishers, 1979.
[9] B. Barber, *The Logic and Limits of Trust*. New Brunswick, New Jersey: Rutgers University Press, 1983.
[10] D. Gambetta, "Can We Trust Trust?," in *Trust, Making and Breaking Cooperative Relationships*, D. Gambetta, Ed. Oxford England: Basil Blackwell, 1990, pp. pages 213--237.
[11] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not so Different After All: A Cross-Discipline View of Trust," *Academy of Management Review*, vol. 23, pp. 393-404, 1998.
[12] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors*, vol. 46, pp. 50-80, 2004.
[13] B. King-Casas, D. Tomlin, C. Anen, C. F. Camerer, S. R. Quartz, and P. R. Montague, "Getting to Know You: Reputation and Trust in Two-Person Economic Exchange," *Science*, vol. 308, pp. 78-83, 2005.
[14] A. Josang and S. L. Presti, "Analysing the Relationship between Risk and Trust," presented at Second International Conference on Trust Management, Oxford, 2004.
[15] A. Josang, "A Logic for Uncertain Probabilities " *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9, pp. 279-311, 2002.
[16] A. Josang and S. Pope, "Semantic Constraints for Trust Transitivity," in *Second Asia-Pacific Conference on Conceptual Modeling* Newcastle, Australia, 2005.
[17] T. Beth, M. Borcherding, and B. Klein, "Valuation of Trust in Open Networks," in *European Symposium on Research in Computer Security* Brighton, UK, pp. 3-18, 1994.
[18] C. Castelfranchi and R. Falcone, "Trust is much more than subjective probability: Mental components and sources of trust," in *Hawaii International Conference on System Sciences*. Kauai, Hawaii, 2000.
[19] C. Castelfranchi and R. Falcone, "Social Trust: A Cognitive Approach," in *Trust and Deception in Virtual Societies*, C. Castelfranchi and Y.-H. Tan, Eds., 2001, pp. 55-90.
[20] R. Falcone and C. Castelfranchi, "The socio-cognitive dynamics of trust: Does trust create trust?," *Trust in Cyber-Societies: Integrating the Human and Artificial Perspectives*, pp. pp.55-72., 2001.
[21] M. J. Prietula and K. M. Carley, "Boundedly Rational and Emotional Agents," in *Trust and Deception in Virtual Society*, C. Castelfranchi and Y.-H. Tan, Eds.: Kluwer Academic Publishers 2001, pp. 169-194.
[22] M. Schillo and P. Funk, "Learning from and About Other Agents in Terms of Social Metaphors," in *IJCAI Workshop on Agents Learning About, From and With other Agents*. Stockholm Sweden, 1999.
[23] M. Schillo, P. Funk, and M. Rovatsos, "Using Trust for Detecting Deceitful Agents in Artificial Societies," *Applied Artificial Intelligence Journal, Special Issue on Trust, Deception and Fraud in Agent Societies*, 2000.
[24] Y. C. Hung, A. R. Dennis, and L. Robert, "Trust in Virtual Teams: Towards an Integrative Model of Trust Formation," in *International Conference on System Sciences*. Hawaii, 2004.
[25] L. Luna-Reyes, A. M. Cresswell, and G. P. Richardson, "Knowledge and the Development of Interpersonal Trust: a Dynamic Model," in *International Conference on System Science*. Hawaii, 2004.
[26] J. Sabater and C. Sierra, "Review of Computational Trust and Reputation Models," *Artificial Intelligence Review*, vol. 24, pp. 33-60, 2005.
[27] D. O. Sears, L. A. Peplau, and S. E. Taylor, *Social Psychology*. Englewood Cliffs, New Jersey: Prentice Hall, 1991.
[28] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. Cambridge, MA: MIT Press., 1994.
[29] H. H. Kelley and J. W. Thibaut, *Interpersonal Relations: A Theory of Interdependence*. New York, NY: John Wiley & Sons, 1978.
[30] R. Axelrod, *The Evolution of Cooperation*. New York: Basic Books, 1984.