Developing Robots that Recognize when they are being Trusted

Alan R Wagner
Georgia Tech Research Institute
250 14th Street NW, Atlanta GA 30332-0822

Abstract

In previous work we presented a computational framework that allows a robot or agent to reason about whether it should trust an interactive partner or whether the interactive partner trusts the robot (Wagner & Arkin, 2011). This article examines the use of this framework in a well-known situation for examining trust--the Investor-Trustee game (King-Casas, Tomlin, Anen, Camerer, Quartz, & Montague, 2005). Our experiment pits the robot against a person in this game and explores the impact of recognizing and responding to trust signals. Our results demonstrate that the recognition that a person has intentionally placed themselves at risk allows the robot to reciprocate and, by doing so, improve both individuals play in the game. This work has implications for home healthcare, search and rescue, and military applications.

Introduction

Trust underlies a great deal of interpersonal interactions. It allows employers to leave the shop knowing that their employees will act responsibly. It allows depositors to place their entire fortune in the vaults of a bank believing that their assets will be safe. Trust permits a trustor to act in manner that puts them at considerable risk, believing that the actions of their counterpart will mitigate that risk.

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 (Economist, 2006). 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.

Developing computational methods that allow a robot to recognize and react appropriately to indications of human trust has important implications for home healthcare, search and rescue, and military applications. The term trust commonly refers to one's belief that the individual being trusted will act in a manner that reduces the trustor's risk (Lee & See, 2004; Wagner & Arkin, 2011). Hence, a robot that fails to recognize that a person is placing their trust in the robot may fail to consider the needs of the person, and by doing so, place the person at risk.

In previous work we introduced a computational framework and algorithm (discussed in greater detail below) that allows a robot to reason about whether it should trust an interactive partner or whether the interactive partner trusts the robot (Wagner & Arkin, 2011). The work presented here examines our computational framework's ability to recognize when a person attempts to signal their trust in the robot. We employ a well-known economic game which has been shown to involve trust to compare a robot's performance when recognizing trust signals to its performance when it cannot recognize such signals.

This article begins with a brief review of the trust literature. Next, an overview of our interdependence framework for social action selection is provided including our methods for recognizing situations which demand trust. Finally, we present a preliminary experiment performed on a robot which explores how the robot's behavior can be made to change when it detects a person's signal trust. We conclude by discussing the ramifications of this work and directions for future research.

Related Work

Early trust research focused on definitions and characterizations of the phenomenon (Deutsch, 1973; Luhmann, 1979; Barber, 1983). 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. 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 Lee and See's description of trust we define 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*** (Wagner & Arkin, 2011)**.**

Researchers have explored many different methods for measuring and modeling trust. Trust measures have been derived from 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). 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).

A Framework for Social Action Selection

Social psychologists define interaction as influence—verbal, physical or emotional—by one individual on another (Sears, Peplau, & Taylor, 1991). The outcome matrix is a standard computational representation for interaction (Kelly & Thibaut, 1978). It is composed of information about the individuals interacting, including their identity, the interactive actions they are deliberating over, and scalar outcome values representing the reward minus the cost, or the outcomes, for each individual. Figure 1 depicts an interaction involving two individuals. In this article the term individual is used to indicate either a human or a social robot or agent. The rows and columns of the matrix consist of a list of actions available to each individual during the interaction. Finally, a scalar outcome is associated with each action pair for each individual. Outcomes represent unitless changes in the robot, agent, or human's utility.

Because outcome matrices are computational representations, it is possible to describe them formally. The notation presented here draws heavily from game theory (Osborne & Rubinstein, 1994). An outcome matrix consists of 1) a finite set N of 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. 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. The term o^i denotes the outcome received by individual i when a pair of actions has been selected.

Investor-Trustee Game Outcome Matrix

Trustee \ Investor	invest 0	invest 1	invest 2	invest 3	invest 4
return 0	0 4	3 3	6 2	9 1	12 0
return 1		2 4	5 3	8 2	11 1
return 2		1 5	4 4	7 3	10 2
return 3		0 6	3 5	6 4	9 3
return 4			2 6	5 5	8 4
return 5			1 7	4 6	7 5
return 6			0 8	3 7	6 6
return 7				2 8	5 7
return 8				1 9	4 8
return 9				0 10	3 9
return 10					2 10
return 11					1 11
return 12					0 12

Figure 1 The outcome matrix above depicts the rewards received for different patterns of investment and return by an investor and a trustee in the Investor-Trustee game.

Action Selection Strategies

Outcome matrices offer several simple action selection strategies. The most obvious method for selecting an action from an outcome matrix is to choose the action that maximizes the robot's outcome. This strategy is termed *max_own*. An individual's use of the *max_own* strategy results in egoistic interactive behavior. Alternatively, the robot may select the action that maximizes its partner's outcome, a strategy termed *max_other*. An individual's use of the *max_other* strategy results in altruistic behavior.

We term a tendency for an individual to use a specific type of action selection strategy their **disposition**. Hence, if an individual tends to often use the *max_own* strategy, typically selecting the action most beneficial to themself without consideration of their partner then their disposition would be that of an egoist. Similarly, an individual that tends to select actions that maximize their partner's reward would be characterized as altruistic. Social events, such as recognizing that someone trusts you, can change one's disposition. In the experiments detailed below, the robot changes its disposition with respect to the person if it recognizes that the person is trust it.

Recognizing Situations that Require Trust

The definition for trust described above focuses on the actions of the trustor and trustee. The investor takes the role of trustor in the outcome matrix depicted in Figure 1. The definition for trust requires risk on the part of the trustor, hence, 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*.

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.* The statement indicates that there will be a difference, ${}_{11}o^i - {}_{21}o^i > \varepsilon_1$, where ε_1 is a constant representing the minimal amount of outcome for dependence. In Figure 1 the difference to the investor if action pair (invest 4, return 8) is selected over the action pair (invest 1, return 1) is 8-4=4.

The trustor may also select the untrusting action, however. This implies that there is an action available to the trustor that does not require risk on the part of the trustor. This leads to a third condition, *3) the outcome received when selecting the untrusting action does not depend of the actions of the trustee.* Stated formally, $\left|{}_{12}o^i - {}_{22}o^i\right| < \varepsilon_2$, where ε_2 is a constant representing the maximal amount of outcome for independence. In Figure 1 the untrusting action for the investor would be to select 0, no investment.

The definition for trust implies a specific pattern of outcome values. Notably, *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 ${}_{11}o^i > {}_{x2}o^i > {}_{21}o^i$.

These provisions 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 2 presents our algorithm for determining if a putative situation requires trust. This algorithm and these conditions are described in greater detail in our related work (Wagner & Arkin, 2011).

Stereotyped Partner Models

A partner model is an individual's evolving model of their interactive partner. The partner models used in this research contain three types of information: 1) a set of partner features $\left(f_1^{-i}, \ldots, f_n^{-i}\right)$; 2) an action model, A^{-i}; and 3) a utility function u^{-i}. Partner features are perceptual features used for partner recognition. The action model contains a list of actions available to that individual. The utility function includes information about the outcomes obtained by that individual when the robot and the human select a pair of actions.

With respect to this framework, a stereotype is a type of generalized partner model used to represent a collection or category of individual partner models. We have developed algorithms for creating stereotypes from a collection of partner models and for matching of a new interactive partner's perceptual features to an existing stereotype (Wagner A. R., 2012). Stereotype creation is a two phase process. First, partner models are clustered with the centroids of the clusters becoming the partner model stereotype. Next, using the cluster centroids as data, a mapping from partner features to the stereotypes is learned. Stereotype recognition occurs when the robot perceives an new person and uses the mapping to obtain a stereotype.

Testing for Situational Trust
(Wagner & Arkin, 2011)

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** $i \Longrightarrow -i$ 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** $\left|{}_{12}o^i - {}_{22}o^i\right| > \varepsilon_2$ //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 2 The algorithm above depicts a method from (Wagner & Arkin, 2011) for determining whether a social situation requires trust. The algorithm assumes that the first individual is the trustor, the second individual is the trustee.

Empirical Evaluation

Empirically evaluating situations that involve trust is challenging. Economic games, such as the Investor-Trustee game (Figure 1), are a common tool used by researchers for exploring trust (King-Casas, Tomlin, Anen, Camerer, Quartz, & Montague, 2005; Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004). The Investor-Trustee game is a social situation in which an investor acts as the trustor selecting some amount of money to invest with a trustee. In each round the investor selects some amount 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 ρ is the correlation coefficient) (King-Casas, Tomlin, Anen, Camerer, Quartz, & Montague, 2005). Investor reciprocity for round j is quantified as $\Delta I_j - \Delta R_{j-1}$ where ΔI_j is the fractional change in investment during round j and ΔR_{j-1} is the fractional change in repayment from the previous round. Similarly, trustee reciprocity was quantified as $\Delta R_{j-1} - \Delta I_{j-1}$. They found that these measures of reciprocity correlated to trust better than either previous investment/repayment (I and R respectively) or change in investment/return (ΔI_j or ΔR_{j-1}).

The Investor-Trustee game is a valuable tool for trust research for several reasons. First, because it imposes a financial risk on the investor and the trustee the game meets a key condition for the definition of trust. Nevertheless, the game does not place anyone at risk of physical harm. Hence, studies that allow humans subjects to participate can ethically be conducted. Further, because money is used, quantitative evaluation of the situation in terms of outcome is straightforward. In other words, the amount of money gained or lost in an interaction can be interpreted as the subject's change in outcome. Finally, the game has a lengthy literature associated with it that spans from neuroscience to behavioral economics (King-Casas, Tomlin, Anen, Camerer, Quartz, & Montague, 2005; Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004).

Experimental Procedure

We used an Investor-Trustee style game to evaluate our hypothesis that if the person selects actions signifying that he or she trust the robot, then the robot can use our method to recognize this signal (Figure 2) and alter its disposition toward the person, resulting in more reward for both individuals. A single round of the game involved the selection of an investment by the robot and the selection of a return by the person. In our version of the game, the robot could invest up to 4 chips representing $5 each. Investments were made by verbally stating the amounts. Returns by the human trustee were similarly communicated verbally to the robot. Speech recognition was used by the robot to determine the amount returned.

The Nao robot by Aldebaran was used for this experiment. The Nao is a humanoid robot with 25 degrees for freedom including actuated hands. The Nao also has two HD 1280x960 cameras and integrated speech recognition.

The robot played ten rounds of the game with ten notionally different human partners. The humans were notionally different in that the same person (the experimenter) used different costumes and accessories to give the appearance to the robot that it was interacting with individuals that had different perceptual features. These features were used by the robot to create a stereotype that included information about which actions the person tended to select. These stereotypes were used by the robot to then predict the person's investment during each upcoming round.

The experiment consisted of both a control condition and an experimental condition. In both conditions the robot interacted with the same notional human partners displaying the same perceptual features in the same order (Table 1). Further, in both conditions partners P0-P4 resembled doctors and partners P5-P9 resembled fire fighters. Thus, perceptually, two different categories were presented to the robot.

The human followed a fixed pattern when deciding how much investment to return. Individuals from the doctor category returned 4 chips regardless of the robot's investment. This category of partner was meant to simulate a person that did not trust the robot.

Table 1 Partner Features and Values

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

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

The robot's decision on how much to invest reflected its experience playing the game as well as its disposition. The robot was programmed to begin playing the game in a manner that maximized its own profit (a *max_own* disposition).

During the control condition the robot did not use our algorithm to test for situational trust (Figure 2) and hence failed to recognize the human's increased risk taking.

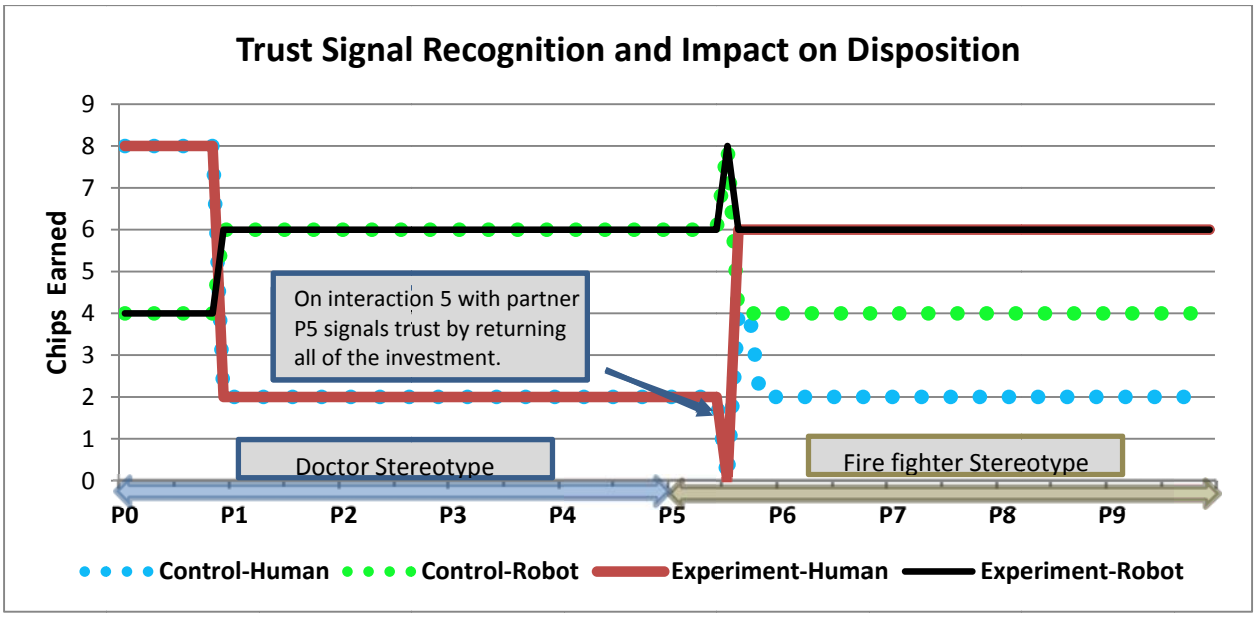

Figure 3. The graph above depicts the number of chips earned during each round of the Investor-Trustee game for both the robot and the human in each condition. The dotted lines indicate the results in the control condition. The solid lines indicate the results in the experimental condition. During the 6th round of play with partner P5 the human trustee returns all of the robot's investment in an attempt to signal the person's trust in the robot. In the experimental condition the robot recognizes this signal and changes its disposition to be more altruistic towards the person, resulting greater outcome for both individuals. In the control condition, the robot does not respond to the person's trust signal and the human trustee retaliates by reducing the return to the robot.

During the experimental condition, however, the robot used our algorithm to recognize that the trustee trusted the robot. The robot then modifies its disposition to be 50% *max_own* and 50% *max_other*. This results in increased investment on the part of the robot.

At the start of game (round 0) the robot began by observing the partner's perceptual features (Table 1). Next the robot selected and stated an amount to invest. The round concluded when the robot recognized the human's verbal statement indicating the return. We recorded both the actions selected and the amounts received by both partners. Because of time, we were only able to run the experiment with each notional partner once. Hence statistically significant results were not possible.

Robot Experiment Results

Figure 3 depicts the results from the experiment. The amount of chips earned in each interaction is displayed along the y-axis. The different partners that the robot interacted with are displayed along the x-axis as P0-P9. The first five human partners were from the doctor category and the later five were from the fire fighter category.

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

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

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

The experimental condition is identical to the control condition until round 6 with partner P5. In this round the human also signals trust by returning the entire investment and appreciation to the robot (8 chips). Here, however, the robot recognizes that the return is not what it predicted. It then uses the method from Figure 2 to determine if the situation demands trust. Once the robot has verified that the situation demands trust and that the person has selected the trusting action, it changes its disposition from 100%

max_own to 50% *max_own* and 50% *max_other*. This change in disposition causes the robot to place greater importance on the outcome received by the partner, which, in turn, causes the robot to increase its investment to 4 chips. The human trustee responds by returning 6 chips. Hence, in this condition the human and the robot receive 6 chips each for the remainder of the experiment.

Summary and Conclusions

This article used our algorithm for situational trust to recognize if a situation demands trust and our framework for social action selection to examine the potential impact these techniques might have on a robot playing the Investor-Trustee game (Wagner & Arkin, 2011). We hypothesized that methods that allow a robot to recognize when a person trusts it would improve the outcomes for both the robot and the human. Our results, albeit limit, support this hypothesis.

Still, the work presented here represents only an initial step in the investigation of this area. Several assumptions and limitations currently exist. First and perhaps most importantly, because the human trustee followed a rather rigid behavioral pattern and even though the pattern was based on observations of game play by people (King-Casas, Tomlin, Anen, Camerer, Quartz, & Montague, 2005), the results may not accurately reflect the play of real humans. Further experiments with real human subjects will thus be needed to confirm these results. Also, the behavioral strategy of the robot was somewhat simplistic in that it the robot simply picked the investment that it believed would maximize its profit based on its model of the human. Although we could develop a more sophisticated approach to playing the game for the robot, the purpose of this work was not to optimize play, but rather to investigate our method to recognizing if a person trusts the robot. Finally, in everyday interpersonal interactions, a human's trust in another person is signaled by a myriad of subtle perceptual cues. This work largely bypasses these cues by focusing on a task in which the trust signal is overt and obvious. Hence, this work does not abate the need for recognizing these subtle perceptual cues. On the contrary, we feel that this work augments our understanding of trust cues by investigating and attempting to formally conceptualize the situational characteristics that lead to these cues.

There are many potential avenues for future research. Perhaps the most pressing will be to test the methods used here on true human subjects. It would also be valuable to expand the research to slightly less structured situations, perhaps involving negotiation and bargaining. Because trust underlies so many different interpersonal interactions finding situations in which it is important for a robot to recognize when they are being trusted should not be difficult.

References

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

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.

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.

Kelly, H. H., & Thibaut, J. W. (1978). *Interpersonal Relations: A Theory of Interdependence.* New York, NY: John Wiley & Sons.

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.

Osborne, M. J., & Rubinstein, A. (1994). *A Course in Game Theory.* Cambridge, MA: MIT Press.

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. pp. 169-194). Kluwer Academic Publishers.

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *NeuroImage, 22*, 1694-1703.

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.

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.* Villamura, Portugal.

Wagner, A., & Arkin, R. (2011). Recognizing Situations that Demand Trust. *20th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2011).* Atlanta, GA.