TRUST DISCOUNTING IN THE MULTI-ARM TRUST GAME

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science

By
MICHAEL COLLINS
B.S. Wright State University 2015

2020
Wright State University

WRIGHT STATE UNIVERSITY
GRADUATE SCHOOL

July 13th 2020

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Michael Collins ENTITLED Trust Discounting in the Multi-Arm Trust Game BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

Ion Juvina, PhD
Thesis Director

Debra Steele-Johnson, PhD
Chair, Department of
Psychology

Committee on Final Examination:

Committee Member: Kevin A. Gluck, PhD

Committee Member: Joeseph Houpt, PhD

Committee Member: Valerie L. Shalin, PhD

Barry Milligan, Ph.D.
Interim Dean of the Graduate School

ABSTRACT

Collins, Michael. M.S. Department of Psychology, Wright State University, 2020. Trust Discounting in the Multi-Arm Trust Game

Social interactions are complex and constantly changing decision making environments. Prior research (Mayer, Davis, & Schoorman, 1995) has found that people use their trust in others as a criterion for decision making during social interactions. Trust is not only relevant for human-human interaction, but has also been found to be important for human-machine interaction as well, which is becoming a growing feature in many work domains (De Visser et al., 2016). Prior research on trust has attempted to identify the behavioral characteristics an individual (trustor) uses to assess the trustworthiness of another (trustee) to determine the trustor's level of trust. Experimental findings have been used to develop into various models of trust (Mayer et al., 1995; Juvina, Collins, Larue, Kennedy & de Mello, 2019) to explain how a trustor comes to trust a trustee. An aspect of trust that has not been investigated is how or if trust changes when a trustor attempts to interact with a trustee, but cannot interact with the trustee. Under such situations Juvina et al.'s (2019) trust model makes the novel prediction that trust will decrease.

To assess the prediction of Juvina et al. (2019) model, a new experimental design (the multi-arm trust game) was developed to evaluate how trust is affected under conditions where an individual variably interacts with multiple trustees. Additionally, the identity the trustee (human and machine) was manipulated to examine differences between human-human and human-machine trust. Before data were collected, the model made *ex-ante* predictions of the participants' behavior. The accuracy of these predictions was then evaluated after the data were collected. The results from our experiment found that our model was able to predict general characteristics of the data confirming the necessity of the model's discounting mechanism, while also highlighting model limitations that are areas for future research.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE PAGE

I. INTRODUCTION

Individuals often make decisions based on the trust they have in others during social interactions (Berg, Dickhaut & McCabe, 1995), in romantic (Gottman, 2011), and work relationships (Colquitt, Scott, & LePine, 2007). For this reason, understanding trust and its influence on decision making is of interest to psychology, sociology, and economics (Rousseau, Sitkin, Burt, & Camerer, 1998). The literature on trust has focused on three main areas of interest. First, trust research has attempted to develop an encompassing definition of trust, in order to separate trust from other related concepts, such as cooperation and risk (Mayer, Davis, & Schoorman, 1995). Second, trust research has attempted to identify different components of trust, such as trait trust, (i.e., one's generalized trust in others) and state trust (i.e., one's trust in a specific person based on his/her behavior in a specific situation; Collins, Juvina, & Gluck, 2016). Finally, research has attempted to develop computational models of trust, specifying how different components of trust interact and influence one's decisions in different environments based on the behavior of others (Dang & Ignat, 2016; Juvina, Lebiere, & Gonzalez, 2015). However, little research has focused on investigating how trust changes over time when individuals do not continuously interact with one another (Lewicki, Tomlinson, & Gillespie, 2006) and how different types of a trustee affect a trustor's trust. The lack of research in these findings limits the scope of which trust theories are applicable. Juvina, Collins, Larue, Kennedy and de Mello (2019) have developed a computational model of trust that contains a discounting mechanism, predicting that trust will decrease under specific circumstances. This discounting is moderated by certain aspects of a trustee's behavior. However, this prediction has yet to be empirically explored. Thus, the purpose of my study is to investigate how one's (trustor) trust in another (trustee) with certain behavioral tendencies changes when the trustor does not continually interact with the trustee. Additionally, this study assesses the predictions of Juvina et al. (2019) trust model, particularly its predictions of trust discounting.

TRUST

Many social situations are characterized by interdependence (Murnighan & Wang, 2016). This means that the outcomes of a given situation depend on the choices made by all involved in a situation. Under these types of situations trust can be used to help simplify these complex and often changing social interactions, such as whom to learn new information from, how to engage in a new situation, and what types of systems to use in a different task. Indeed, research has found trust to influence an individual's decisions in a variety of different domains. Harris and Corriveau (2011) found that young children use their trust in adults as a criterion for deciding from whom to learn new information. Juvina et al. (2013) found that trust mediates transfer of learning across games of strategic interaction. Hancock, Billings, Schaefer, and Chen (2011) noted that trust influences the decisions humans make with inanimate objects, such as machines and robots. These findings support the notion that trust is a means of risk and complexity reduction in situations that are characterized by uncertainty and interdependence (Mayer et al. 1995; McLaain & Hackman, 1999).

Researchers have proposed many different definitions of trust. In this thesis, I use Mayer et al. (1995, p. 712) definition: “the willingness of a party [trustor] to be vulnerable to the actions of another party [trustee] based on the expectation that the trustee will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party.” I chose Mayer et al (1995) definition because it separates trust from other concepts, such as risk and cooperation. Concepts such as risk and cooperation were originally associated with trust, but researchers have since argued that these concepts are separate from trust (Deustch, 1958; Hardin, 1993; Lewicki et al. 2006).

COMPONENTS OF TRUST

Trust is composed of several components that influence the decisions of a trustor in different contexts. The first component is trait trust, also referred to as trust propensity or dispositional trust (Mayer et al., 1995). Trait trust refers to one’s general willingness to trust other people absent any information from a trustee. Trait trust influences a trustor’s decisions under circumstances in which they have no information about a trustee’s trustworthiness (Berg et al., 1995; Engle-Warnick & Sloim, 2004). Those higher in trait trust are more willing to initially trust another person without knowing any information about their trustworthiness whereas those lower in trait trust are less willing to initially trust others. Rotter (1967) originally characterized trait trust as a highly stable predisposition. More recent research has suggested that trait trust changes as a function of experience over a lifetime. Collins, Juvina, and Gluck (2016) found a small but statistically significant change in an individual's trait trust that depended on the trustworthiness of a trustee with whom they recently interacted.

The second component is state trust, which is the trustor's trust in a specific individual in a given situation. One's state trust develops from interacting with a specific trustee in a given situation, either increasing or decreasing based on the trustee's trustworthiness (i.e., willingness to "perform an action important to the trustor", Mayer et al. 1995, p.712). State trust is affected by a variety of different factors such as the trustee's ability, benevolence, integrity, and familiarity (Alarcon, Lyons, & Christensen 2016; Collins et al. 2016, Juvina et al. 2013; Mayer et al. 1995). In addition, state trust is affected by a trustor's perceived trust necessity (Juvina et al., 2015; Collins et al., 2016) and ability to detect and decode trustworthiness signals (Juvina et al., 2019).

MODLES OF TRUST

Models of trust have been developed in order to specify how different components of trust interact and in what context trait and state trust influence an individual's decisions. Many models of trust have been proposed (Deutsch 1958; Hardin, 1993). However, for this paper I review two models of trust Mayer et al. (1995) and Juvina et al. (2015).

MAYER ET AL.'S (1995) VERBAL MODEL OF TRUST

Mayer et al. (1995) model explains how a trustor comes to develop trust in a specific trustee. According to Mayer et al. (1995) model, a trustor's trust in a trustee is based on two components. The first component is a trustor's trait trust[1]. The second component is a trustor's perception of the trustee's trustworthiness. The model assumes that the more trustworthy a trustor perceives a trustee to be, the greater trust the trustor will have for a trustee. Three antecedents moderate the trustor's perception of trustworthiness. The first antecedent is ability (i.e, the trustee's competency on a particular task). The second antecedent is benevolence (i.e., the trustee's desire to do good for the trustor). The third antecedent is integrity (i.e., the moral character of the trustee).

Finally, Mayer et al. (1995) proposed that a trustor's state trust is a combination of both trait trust and perceived trustworthiness, weighted differently based on the time point of the interaction. Under circumstances in which a trustor does not have any experience with a trustee, a trustor's state trust is only determined by the trustor's trait trust. After interacting with a trustee, the trustor's state trust is weighted towards their perception of the trustee's trustworthiness. Mayer et al. (1995) model proposes that a trustor's state trust changes over time, increasing or decreasing based on the trustee's behavior. Continuously updating one's perception of a trustee's trustworthiness allows the trustor's trust to reflect the trustee's recent behavior.

Mayer et al (1995) interpersonal trust model is an improvement over other previously proposed trust models (Deutsch 1958; Hardin, 1993) for three reasons. First, Mayer et al.

[1] Mayer et al. (1995) uses the term trust propensity, but for consistency in this document I use the term trait trust.

(1995) takes into account both trait and state trust when attempting to account for the trustor's trust in a trustee. Second, Mayer's et al (1995) model explicitly states specific antecedents that affect a trustor's perception of a trustee's trustworthiness. Third, Mayer's et al (1995) model incorporates a feedback loop proposing how state trust changes over time. Despite the strengths of Mayer et al.'s (1995) model three limitations of the model do exist. First, Mayer et al.'s (1995) model does not make any claims about the degree that trust increases or decreases over time. For example, Mayer et al. (1995) model does not state if trust increases or decreases according to a linear or a non-linear function. The assumption about how trust changes over time makes a difference in the predicted behavior that is to occur between a trustee and trustor (Juvina et al. 2019). Additionally, the Mayer et al. (1995) model does not state if or how trust changes during instances when the trustor cannot interact with the trustee. Finally, Mayer et al.'s (1995) model does not specify how a trustor combines the judgments of a trustee's behaviors (i.e., ability, integrity, and benevolence) into judgement of trust. Though Mayer et al. (1995) specifies behavioral actions that a trustor will be sensitive to (i.e., benevolence, ability, and integrity), Mayer et al.'s (1995) theory does not specify how certain actions by a trustee are weighted by trustor who will attribute a certain degree of trust to a trustee. Humans have been shown to initially place more trust in automated systems compared to humans, but quickly lose trust in an automated system once it displays an untrustworthy behavior (De Visser, 2016; Lee & See 2004. However, this quick loss of trust in an automated system has shown to be moderated by a system's anthropomorphic features or the ability to explain its actions (De Visser et al. 216, De Mello et al. 2010, Lee & See 2004). This suggests that the type of system (i.e., human or machine) and its capabilities affect the trustworthiness that a trustor is willing to attribute to the system. However, Mayer et al.'s (1995) model currently has no way of accounting for these effects. Understanding these three limitations of Mayer et al. (1995) model is important for understanding the role of trust in decision making over longer periods of time. Juvina et al. (2015) proposes a computational model of trust that addresses these three limitations of Mayer's et al. (1995) trust model.

JUVINA ET AL.'S (2019) COMPUTATIONAL MODEL OF TRUST

Juvina et al. (2019) computational model of trust accounts for how a trustor learns to trust a trustee. As in Mayer et al. (1995) model, in the absence of experience with a trustee, a trustor's trust is influenced by trait trust (T_0). After the trustor interacts with a particular trustee, the trustor's trust in the trustee is influenced by the trustee's perceived evidence of trustworthiness (*PET*). PET is a positive or negative value associated with an action taken by a trustee. A trustor's state trust at time *t* accumulates according to a power law (Equation 1).

The trustor's previous trust in a particular trustee (ST_{t-1}) is raised to a power (that we refer to as the discounting parameter *a*), which is a positive value less than 1, and then is added to PET_t. Raising the trustor's previous value of trust (ST_{t-1}) to *a*, allows trust to increase or decrease according to a power law. Allowing state trust to increase according to a power law gives trust three particular properties. First, state trust increases at a faster rate when trust is low compared to when it is high. Second, a trustor's state trust in a trustee will eventually plateau. Third, trust can quickly be lost when exposed to negative evidence of trustworthiness, allowing the trustor's state trust to reflect the trustee's most recent behavior.

$$ST_t = ST_{t-1}^{a} + PET_t \qquad (1)$$

Juvina et al. (2015) model of trust has been implemented within the Adaptive Control of Thought – Rational (ACT-R) architecture (Anderson, 2007) and has been found to account for human learning across multiple domains (Collins et al., 2016; De Mello et al., 2011; Lount et al., 2008; Juvina et al. 2013; De Visser et al. 2016). Additionally, the Juvina et al. (2019) trust model makes novel predictions about how trust will change under conditions in which the trustor attempts to but cannot interact with a trustee. Specifically, Juvina et al. (2019) model predicts that, under conditions in which the trustor expects to interact with the trustee and the trustee does not interact with the trustor, the trustor's trust in the trustee will be discounted (i.e., decrease). Trust is predicted to be discounted under these circumstances because in Juvina et al. (2019) model the trustor's previous assessment of trust (ST_{t-1}) is decreased by raising the trustor's previous trust assessment by a discounting parameter (*a*). Due to the fact that the trustee did not interact with the trustor, PET for the interaction would be zero, not allowing to offset the discounting of the trustor's previous trust assessment. Given enough failed attempts to interact with a trustee, the trust discounting equation (Eq. 1) predicts that a trustor's trust in that trustee will change from trust to distrust.

TRUST AND TRUSTEE IDENTITY

Beyond the regularities that are observed in trust between humans (i.e., trust is based on the behavior of the trustee, trustworthy behavior leads to more trust, untrustworthy actions lead to distrust, etc.) aspects other than the trustee's overt behavior, such as a trustee's identity, have also been found to affect the trustor's trust in the trustee. Research on trust in automated systems has revealed differences in a trustor's trust in a trustee, based on whether the trustee is a human or an automated system (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Lee & See 2004; de Visser et al. 2016). In line with these findings is research showing that humans exhibit a tendency to interpret particular types of systems as being animate and having goal oriented behavior, even if they are non-human entities (Hieder & Simmel, 1944 ; Csibra, Gergely, Biro, Koos & Brockbank, 1999). Finding that humans have a tendency to assume that at least some objects give goals, leads to particular inferences about their current and future behavior, such as trust. Two commonly observed differences in the trust between a human and automated system are that, (1) humans have been found to have higher initial trust in an automated system compared to humans and (2) humans have been shown to be more apt to lose their trust in an automated system compared to a human following an untrustworthy action (Nass; Steur & Tauber, 1994; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; de Visser et al. 2016). These two regularities suggest that humans' trust is moderated by the identity of the trustee (i.e., animate vs non-animate). Automated systems are thought to be designed to be effective at a particular task, leading humans to have high initial trust in an automated system. Furthermore, automated systems are often developed to complete a particular task and not have robust abilities to complete a range of tasks. If a system makes an error, then a user might assume that the error made by a system is suggestive of its overall ability, leading the user to decrease its trust in a system. Additional research has focused on how these differences in human and automated trust can be overcome, allowing a trustor's trust in an automated system to be more dynamic and robust like that of a human. Slight modifications to an automated system, such as making a system anthropomorphic (de Visser et al. 2016), giving a system facial expressions (De Melo et al. 2010), or allowing the system to communicate (de Visser et al. 2016; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003) have all been found to make a user's trust in an automated system more similar to a human agent, increasing the user's tendency to infer goal directed behavior. The literature on trust in automated systems shows that a trustee's identity plays an important role in how a trustor trusts a trustee (i.e., human or automated system) affecting its initial beliefs and how it updates information about the trustee. A fully developed model of trust should be able to account for the differences in interacting with multiple types of trustees.

LIMITATIONS WITHIN THE LITERATURE

The literature on trust is vast and continuously growing. Currently, three limitations exist within the research on trust. One, many researchers have examined the role of trust during contextually limited games of strategic interaction (Berg et al., 1995;

Collins et al., 2016; Yamagishi et al., 2005). During games of strategic interaction, participants play simple games representing an abstract situation with other participants or confederate agents. In these situations, participants often have complete information about the game, cannot verbally communicate interact with one another, and have limited choices. Examining the role of trust during abstract games of strategic interaction makes it difficult to understand how other contextual factors, such as trustee identity, facial expressions and body language might interact and affect trust as well (De Melo et al. 2011; Lee et al. 2013). Second, many models of trust are verbal theories. Verbal theories offer general descriptions of what factors affect trust and how trust develops between individuals (Deutsch, 1958; Mayer et al. 1995; Lewicki, Mcallister & Bies, 1998). A limitation of verbal theories is that quantitative predictions are difficult to make compared to mathematical or computational models (Hoffrage & Marewski, 2015).

Finally, many studies examining the role of trust are conducted under conditions where participants interact only once (Berg et al. 1995) or consistently (Collins et al., 2016; Engle-Warnick & Slonim, 2004; Juvina et al., 2013). Investigating the effects of trust development between people continuously interacting during short periods of time does not allow for the investigation of how a trustee's variable interaction schedule with a trustor might affects trust development. Due to the schedule of interaction between participants used in different studies, it is currently unknown if trust decreases, increases, or remains the same during periods of time when a trustor attempts to interact with the trustee, but cannot. In this thesis, I address the effect of the trustee's interaction schedule and trustee identity on trust.

TRUST DISCOUNTING

As previously mentioned many studies on trust have the trustor continually interact with a trustee. Under these circumstances, how (or if) trust changes during periods in which a trustor and trustee do not interact cannot be assessed. Understanding how trust changes during periods when a trustee does not interact with the trustor is necessary for understanding more complex real-world environments. For example, in many organizations employees have a set list of tasks to complete. Employees can either attempt to complete all of their assigned tasks on their own or attempt to delegate task(s) to other employees, who may or may not accept to complete the task. In this situation an employee can be viewed as a trustor and all other employees are trustees. A trustee, if delegated a task by a trustor may choose to complete the task, doing either high or low quality work, or decline to help the trustor. Under this situation an employee (i.e., trustor) when deliberating on whom to attempt to delegate a task to may choose another employee based on their trustworthiness. Trustworthiness of a trustee in this context might be based off of two factors (1) quality of their previous work (i.e., ability) and (2) the frequency at which the trustee has interacted with the trustor (i.e., interaction schedule).

Failing to understand how trust changes over extended interactions across players limits the generalizability of current models of trust. It is unlikely that a trustor interacts continually with all of the trustees they have trust in. In turn, it is also unlikely that trustees are always available to interact with all the trustors who might place trust in them. Instead, it is more likely that there are instances where the trustor and the trustee do not interact for periods of time. We focus here on the case in which a trustor attempts to interact with a trustee, but the trustee cannot interact with the trustor. Mayer et al. (1995) and Juvina et al. (2019) trust models make different predictions about how a trustor's trust for a trustee changes during periods when the trustor attempts to but cannot interact with a trustee.

The Mayer et al. (1995) model predicts that a trustor's trust will remain the same, under conditions when the trustor and the trustee cannot interact. Trust is predicted to remain at the same level, due to the fact that under these circumstances, there is no evidence of trustworthiness or untrustworthiness and therefore the trustor should not change their perception of the trustee's trustworthiness. However, Juvina et al. (2019) trust model predicts that the trustor's trust in a trustee will decrease or be discounted. Juvina et al. (2019) trust model predicts trust discounting, based on the assumption that the uncertainty about the trustee's trustworthiness increases as the time passes. The increased uncertainty about the trustee's trustworthiness is due to the fact that a trustee's trustworthiness is not static but can change. A trustee's trustworthiness could change for a variety of reasons unknown to the trustor, such as change of internal motives or external incentives. Furthermore, an automated system's trustworthiness might shift due to computer malfunction or change in the environment leading to its algorithm no longer being effective. Discounting of old information in favor of more recent information is common in reinforcement learning and belief learning models used to model standard decision making tasks (Anderson, 2007; Camerer, Ho, & Chong, 2002; Zaki, Kallman, Wimmer, Oschsner, & Shohamy, 2016). By discounting previously assessed trust in a particular trustee, Juvina et al. (2019) trust model places more emphasis on the trustee's most recent behavior.

Indeed, de Visser et al. (2016) found that a trustor's trust in a trustee is not held constant during periods where the trustor cannot assess the behavior of a trustee. De Visser et al. (2016), found evidence that, when interacting with different anthropomorphized agents, trust was quicker to decrease under situations in which the trustor could not continuously monitor the behavior of the agent. De Visser's et al. (2016) finding suggests that a trustor's representation of a trustee's trustworthiness is not static during periods of time when a trustor cannot directly monitor the behavior of the trustee. Instead, De Visser et al.'s (2016) finding suggest that trust is discounted during the period of time when the participant could not observe the behavior of the anthropomorphized agent.

During situations where a trustor attempts to interact with a trustee and there is no observable evidence of trustworthiness, Juvina et al.'s (2019) trust updating equation (Eq. 1) predicts that trust will be discounted. This counterintuitive prediction stems from two aspects of the Juvina et al.'s (2019) trust update equation. First, the trustor's previous trust assessment of the trustee is raised to a trust discounting parameter, decreasing the previous assessment of trust. Second, due to the fact that the trustor did not observe the trustee's behavior, there is no perceived evidence of trustworthiness to offset the decrease in the trustor's previous trust assessment. The overall effect of these two factors is that the trustor's trust in the trustee decreases when the trustor cannot directly examine the trustworthiness of a trustee.

INVESTIGATING TRUST DISCOUNTING

One reason that the effects of trustee's interaction schedule on trust have not been investigated is due to the types of games and their implementations that are often used in trust research. For example, one commonly used game is the trust game, also referred to as the investment game (Berg et al. 1995). The trust game has been used extensively in both psychology and economic research to examine the role of trust in decision making (Johnson & Mislin, 2011). The trust game is a simple game of strategic interaction played with two players. Each player is randomly assigned to a specific role of either trustor (Player 1) or trustee (Player 2). At the start of the trust game, Player 1 is given a particular endowment of points[2]. Player 1 is then given the opportunity to allocate any amount of their endowment to Player 2 and keep the remainder. Any points allocated from Player 1 to Player 2 are first tripled by the experimenter before being given to Player 2. For example, if Player 1 sent Player 2 10 points, then Player 2 would receive 30 points. The multiplication of the Player 1's endowment creates the social dilemma that is

[2] The trust game can also be played with money, but in this thesis I use the term points for consistency throughout the document.

relevant for a trust scenario, because Player 1 took a risk allocating their endowment to Player 2 (Rieskamp & Gigerenzer, 2005) After Player 2 receives the tripled number of points allocated from Player 1, Player 2 is given the opportunity to send any number of their received points back to Player 1 keeping the remaining amount. Once Player 2 makes a decision about how many points to send back to Player 1 the interaction ends and players earn their respective payoffs. Player 1's payoff is determined by the number of points of their endowment they kept for themselves and the number of points sent back to them by Player 2. Player 2's payoff is based on the number of points of tripled endowment Player 2 kept for themselves.

The trust game is often used in trust research due to the game's simplicity and ability to operationalize behavioral trust. The trust game allows for an easy quantitative measure of behavioral trust and trustworthiness (Camerer, 2003, Berg et al. 1995, Murnighan & Wang, 2016). During the trust game, the amount of points sent by Player 1 to Player 2 is thought to be a measure of Player 1's trust in Player 2 and the number of points Player 2 sends back to Player 1 is thought to be a measure of trustworthiness, for two reasons. One, Player 1 is under no obligation to allocate any of their endowment to Player 2. Two, Player 1 understands that Player 2 is not obligated to send any of their received number of points back to Player 1. These two factors satisfy the constraints of Mayer et al.'s (1995) definition of trust. The "willingness to be vulnerable to the actions of another" (Mayer et al. 1995 p . 712) (i.e., sending any part of their endowment to Player 2) with the "expectations that that the trustee (Player 2) will behave in a beneficial way towards the trustor (Player 1)” (Mayer et al. 1995 p . 712) (i.e., sending back part of the tripled allocation). The number of points that Player 2 sends back to Player 1, are thought to be a behavioral measure of Player 2’s trustworthiness, due to the fact that Player 2 is under no obligation to send any points or money back to Player 1. By sending points back to Player 1, Player 2 is behaving both benevolently and with integrity, both of which are characteristics of trustworthiness (Mayer et al., 1995).

The trust game has been implemented in experiments in primarily three different ways. Each of the three common implementations of the trust game is inadequate to investigate trust discounting. The first common implementation of the trust game is the one shot game (Berg et al., 2015). During a one shot implementation of the trust game, participants are randomly assigned to roles (Player 1 or Player 2) and to a specific experimental pair. Once assigned to an experimental pair, participants play one round of the trust game after which the experiment ends. During the one shot trust games, the trust measured by a one shot interaction is Player 1’s trait trust. Due to the fact that Player 1 and Player 2 have no experience with one another, Player 1 must rely on their general trust in others (trait trust) to influence his decision about how to interact with Player 2. Though one shot games are relevant for understanding how trait trust influences initial behavior between two people, one shot games do not allow for trust development between a trustor and a trustee to be examined.

The second type of implementation of the trust game is repeated anonymous interactions within a group (Ignat, Dang, Shalin & 2019 – Control condition). During this implementation of the trust game, participants are randomly assigned to play the trust game repeatedly with other participants within a particular group. At the start of the experiment participants are assigned to interact with others within a particular group. After being assigned to a group all members are randomly assigned to play one of the two roles (i.e., Player 1 or Player 2). Then participants are partnered with another group member to play a single instance (i.e., round) of the trust game. After each round, all members of the groups are again randomly assigned to a new role (i.e., trustor or trustee) and a new partner to play with. Over the course of repeated rounds of the trust game played with different group members, the identity of all players remains anonymous. The effect of player anonymity is that participants cannot develop specific trust relationships with certain players during the game. The lack of specific trust relationships means each group member has to place trust in the group as a whole.

Finally, the third type of implementation is the repeated sequential interaction (Dubois, Willinger & Blayac, 2012). During this implementation of the trust game, participants are assigned to a particular role (i.e., Player 1 or Player 2) and are paired with another participant. The two participants then play the trust game repeatedly with the same partner for a set number of iterated rounds. During this implementation of the trust game, specific trust relationship between Player 1 and Player 2 can develop. Specific trust relationships can develop due to the fact that both players gain experience with one another, which allows Player 1’s trust behavior to become specific to the behavior of Player 2. Repeated sequential interactions of the trust game resemble the most common development of a trust relationship in a real world situation. Both players begin the game with no experience with each other, but over time through repeated interactions, Player 1 comes to develop some level of trust in Player 2 based on their trustworthiness.

Although repeated sequential interactions of the trust game are more similar to real world interactions (as compared to one shot or repeated anonymous interactions within a group), they still miss crucial features of how a trustor uses trust to inform their decisions in real world environments. First, in most situations a trustor is not forced to interact with a single trustee. Instead, a trustor likely has the opportunity to stop interacting with a trustee if they are found to be untrustworthy. Second, a trustor in many situations likely has the opportunity to interact with multiple trustees and can choose to interact with as many or as few of these trustees as the trustor wants. Finally, it is unlikely that all trustees interact with a trustor continually. A more realistic assumption is that trustees will interact with the trustor on a variable schedule.

One similarity between these three factors (i.e., unencumbered interaction, multiple trustees, and variable schedule) is that these features all capture aspects of the exploration-exploitation dilemma. The exploration-exploitation dilemma is found in many different domains, like choosing a restaurant, network development, and business (Cohen, McClure, Yu, 2007). It occurs when a decision maker has the opportunity to choose from multiple decision options with each option offering a different potential payoff that can only be revealed through experience. The dilemma facing the decision maker is to develop a strategy that allows the decision maker to explore enough of the decision options to infer the possible reward that can be obtained from each decision option and exploit the decision options that yield the highest rewards by continually choosing them.

This exploration-exploitation dilemma shares several similarities with a trustor having to choose a trustee to interact with. Each trustee likely has its own level of trustworthiness and a variable schedule in which to interact with a trustor, which affects the benefit the trustor can obtain by attempting to place trust in the trustee. The trustor then must decide which of the multiple trustees to interact with. It is this characteristic that is not fully represented in the standard implementations of the trust game. In order to add this decision dilemma into the trust game, I combine the standard trust game with the multi-arm bandit game. The multi-arm bandit game is an experimental paradigm explicitly used to study the exploration-exploitation dilemma. By pairing the multi-arm bandit game with the trust game, I examined the effect of a trustee's variable interaction schedule on a trustor's behavior.

THE MULTI-ARM BANDIT GAME

The multi-arm bandit (MAB) game is a decision making experiment first proposed by Robbins (1952), to highlight the dilemma of exploration and exploitation in sequential experimental designs. The MAB game has since been of interest to many different fields, such as economics, psychology, and computer science (Cohen, McClure, & Yu, 2007). During the MAB an individual is given the opportunity to select from *N* choice options (i.e., arms). Each arm when selected randomly chooses a reward from an unknown reward distribution. The individual is then given multiple opportunities to repeatedly select from any of the *N* arms. It is assumed that participants would attempt to maximize the number of points that they earn. In order to accomplish this goal, the participant must select from multiple different arms in order to determine which arms offer the highest reward (exploration) and then exploit the arm(s) that the subject perceives as giving the best reward (exploitation).

THE MULTI-ARM TRUST GAME

The exploration-exploitation dilemma is similar to the situation of a trustor choosing among multiple trustees to interact with. In principle, every other person is a trustee and could potentially be trustworthy. However, individuals learn over time that people vary in the extent to how trustworthy they are. As previously mentioned, this is analogous to the multi-arm bandit game. Each arm has the potential of giving out a reward, but the decision maker knows that some arms may provide a higher reward than others. To solve this dilemma, individuals sample from multiple different arms seeing the rewards that are offered from each arm (i.e., interact with multiple trustees assessing their trustworthiness). From the sampled arms individuals attempt to choose predominately from a single arm (i.e., choosing to interact with the most trustworthy trustee(s)).

In order to invoke the exploration and exploitation dilemma into the standard trust game, I created a new experimental design making three modifications to the standard trust game, based on the multi-arm bandit game[3]. First, participants were told that they were playing the game simultaneously with three other participants who have a particular identity (human or automated agent). In reality, participants played with 3 confederate agents whose behavior was predetermined. All participants were told that they played with a particular type of agent to trigger their initial beliefs about the identity of the confederate agent. The use of confederate agents is common in research using games of strategic interaction, in order to control for different experimental variables (Alexrod, 1984; Nowak & Sigmund, 1993; Craig, Asher, Oros, Brewer, & Krichmar, 2013; Collins et al., 2016). In this experiment I used confederate agents in this study to specify the three trustees' trustworthiness and frequency of interaction with the participant. Second, instead of only choosing a single arm during each round, as in the standard multi-arm bandit game, participants had the opportunity to interact with as many or as few of the trustees as they wish during a round. Third, instead of making a discrete choice (e.g., choose an arm or not to choose an arm) participants were given a per round endowment of points during each round and then had the choice to freely allocate those points to themselves or to any of the 3 other confederate agents.

All participants were told that the number of points they sent to another confederate agent would be multiplied by 4[4] and then the agent will have the ability to

[3] The modifications presented here are based on the pilot study presented in Appendix D.
[4] The multiplier for the points allocated from a trustor to a trustee was increased from 3 as in the standard trust game to 4 to encourage participants to allocate their endowment to the confederate agents. In our previous pilot study (Appendix D), we found that some

freely choose how much to return to the participant. This decision is analogous to the decision made in the standard trust game (Berg et al., 1995), with the trustor deciding how many points to allocate to a particular trustee. Under these conditions the amount sent by the participant is a behavioral measure of trust and the amount sent back from the confederate agent is the measure of trustworthiness.

Finally, to investigate the effect of not interacting with a trustee, the schedule of when each confederate agent is able to interact with the participant is manipulated. Participants were told that during each round the confederate agents have the opportunity to choose between two tasks. The confederate agent can choose to either interact with the participant and accept the number of points sent by the participant or choose not to interact with the participant. If the confederate agent decides not to interact with the participant, they will have the opportunity to receive a reward randomly selected from an unknown distribution. If the participant decides to allocate any of their endowment to a confederate agent who has chosen not to interact with the participant during a particular round, the participant will be notified that they could not send their allocation to that counterpart during that round. By manipulating the schedule of the confederate agents the effect of trust discounting can be examined.

participants were reluctant to interact with the confederate agents. For this reason we increased the incentive to interact with confederate agents, increasing incentives to participants were found to modify participants behavior during a game (Akçay, & Roughgarden, 2011 & Rapport, 1967).

Each of the 3 confederate agents were placed on a unique interaction schedule with the participant. The first interaction schedule is the *high* interaction schedule, where they had the opportunity to interact with the confederate agent during each round. The second interaction schedule is the *medium* interaction schedule. On the medium interaction schedule, the confederate agent had the opportunity to interact with the participant during every 3 rounds. The third interaction schedule is the *low* interaction schedule. On the low interaction schedule, the confederate agent had the opportunity to interact with the confederate agent during every 6 rounds.

The behavior of all three confederate agents was held constant, with added stochasticity in the models behavior, over the course of the experiment. During the game each confederate agent utilized two different strategies during different parts of the game. During the first part of the game (rounds 1- 70) each confederate agent returned back 75% of the multiplied number of points sent by the participant (i.e., three times the allocated amount), during rounds when it can interact with the participant. The purpose of the initial strategy used by the confederate agent is to allow the participant to develop varying degrees of trust in the 3 confederate agents based on the interaction schedule of each confederate agent. During the second part of the multi-arm trust game (i.e., rounds 70-120) the confederate agent changed its strategy. The second strategy used by the confederate agent was to send back on average the same number of points sent by the participant to the confederate agent (i.e., 25% of the multiplied amount). Participants on average do not gain or loose any points while the confederate agents use their second strategy. The purpose of confederate agents second strategy is to observe the effect the interaction schedule on trust discounting under conditions were potential payoff that can be earned across all confederates was equal.

Although, the behavior of each of the four confederate agents was held constant over the course of the study, manipulating the schedule of each confederate agents creates a difference in potential payoff across the four confederate agents. For example, the total payoff that can be earned from each confederate differs depending on how frequently a participant can interact with a particular confederate agent. Confederate agents with a higher interaction schedule, on average, yielded a higher payoff than a confederate agent with a lower interaction schedule. The difference in potential payoff between confederate agents is confounded with interaction schedule only in the high interaction schedule. However, this confound is mitigated by the addition of the confederate agents' trustworthiness neutral strategy.

A second experimental condition was added, where human participants were told that they would interact with an automated agent (non-animacy condition) who has been developed to play this game as a human. This creates a situation where participants are exposed to the same payoff as in the animacy condition, but are told that they are not interacting with humans and instead a computer agent. This type of design has been used to highlight the differences between human-human and human-machine trust. A comparison between the two experimental conditions allows the difference in behavior between the two conditions where participants are exposed to the same level of payoff to be examined.

HYPOTHESES

In summary, the goal of our experiment is to examine the effects of a trustee's trustworthiness, identity, and interaction schedule on a trustor's trust development and behavior within the multi-arm trust game. By investigating the effects of a trustee's interaction schedule on the participants' behavior, I can assess if a trustor's trust in a trustee is discounted during periods when a trustor cannot interact with a trustee, according to the predictions of Juvina et al.'s (2019) trust model.

Model *ex-ante* predictions of the participant's behavior in the two experimental conditions (i.e., animacy, and non-animacy) were generated prior to collecting data for the study. All of the *ex-ante* model predictions (including specifics of both the model implementation and fit) are presented in Appendix D and E. They were generated from fitting the model to pilot data (Appendix C) collected in a pilot study with a slightly different experimental design and were preregistered and made available online prior to data collection[5]. From the set of model predictions 10 experimental hypotheses were developed.

Across the experiment, I predicted that participants' trust would be sensitive to the behavior of the confederate agent's strategy (high or neutral trustworthiness) and

[5] https://osf.io/3e25a

interaction schedule (i.e., high, medium, and low) of the confederate agent. According to Juvina et al.'s (2019) trust model each of these factors are predicted to affect the participant's trustworthiness over the course of the experiment. The confederate agent's high trustworthiness strategy should lead to more trust being developed, increasing the participants allocations overall the confederate agents (H1, H1a). Whereas the neutral trustworthiness strategy should lead to a decrease in trust leading to a decrease in the participants allocation (H2, H2a). Furthermore, our trust model made specific predictions about the effect the confederate agent's trustworthiness would have on the participant's trust over the course of the experiment, with trust being discounted in the medium and low interaction schedules (H3). These predictions are consistent with the trust literature's consensus that trust is dependent on the trustworthiness of the trustee.

Hypothesis 1*: We predict participants will allocate a positive amount of their endowment to the 3 confederate agents over the course of the multi-arm trust game while the confederate agents use the high trustworthiness strategy (Figure 1-A).*
H1a*: We predict that the participants' rate of their per round allocation will have a positive relationship with the confederate agents interaction schedule while the confederate agents use the high trustworthiness strategy (Figure 1-B).*

Hypothesis 2: *We predict that participants will decrease their overall rate of allocation to the 3 confederate agents during the trustworthiness neutral portion of the experiment (Figure 1-A).*
H2a: *We predict that participants will decrease their rate of allocation to the confederate agent on the high interaction schedule, while the confederate agent uses the trustworthiness neutral strategy (Figure 1-B).*

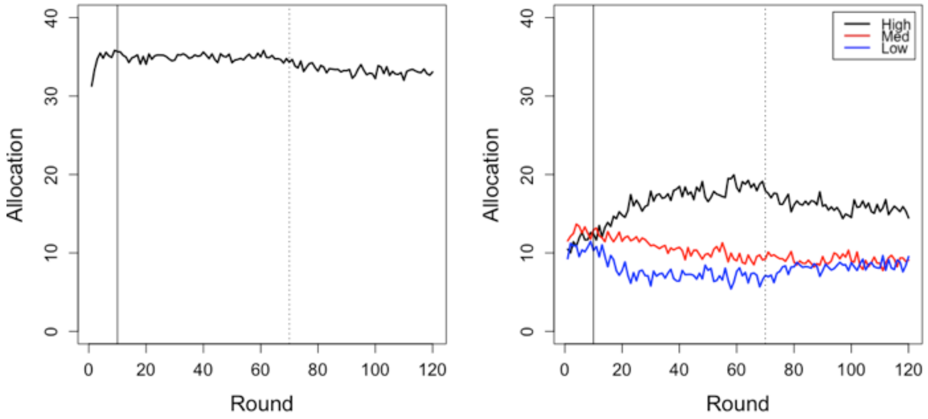


Figure 1. The average round by round allocation of the trust models' predictions averaged across all three of the confederate agents (1-A, left panel) and average round by round trust model predictions to the confederate agents (1-B, right panel) on the high (black line), medium (red line), and low (blue line) interaction schedule averaged across both the animacy and non-animacy conditions.

Hypothesis 3: *We predict that participants' overall average allocation across the animacy and non-animacy conditions to the confederate agents will have a positive relationship with the confederate agents' interaction schedule (Figure 2).*

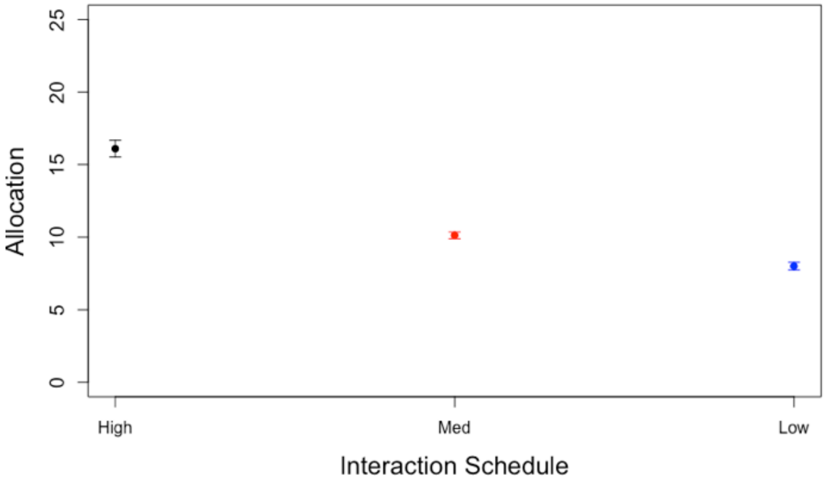


Figure 2. The average allocation and 95% confidence intervals of the trust model's predictions of human behavior relative to the confederate agents on the high (black dot), medium (red dot), and low (blue dot) interaction schedules, averaged across both the animacy and non-animacy conditions.

Along with the behavioral characteristics of the confederate agent (i.e., strategy and interaction schedule), I also predicted that the participant's trust would depend on the confederate agent's identity (i.e., animacy or non-animacy). Previous research has shown that humans interpret the behavior of humans and automated systems differently, having more resilient trust for humans compared to automated system (De Visser 2016, Nass; Steur & Tauber, 1994). Based on the prior research, our model predicts that overall allocations would be lower in the non-animacy compared to the animacy condition (H4) and would interact with various aspects of the confederate agents behavior (H5, H5a, H6).

Hypothesis 4*: We predict that, on average, participants will allocate a greater portion of their endowment to the confederate agents in the animacy condition compared to the non-animacy condition (Figure 3).*

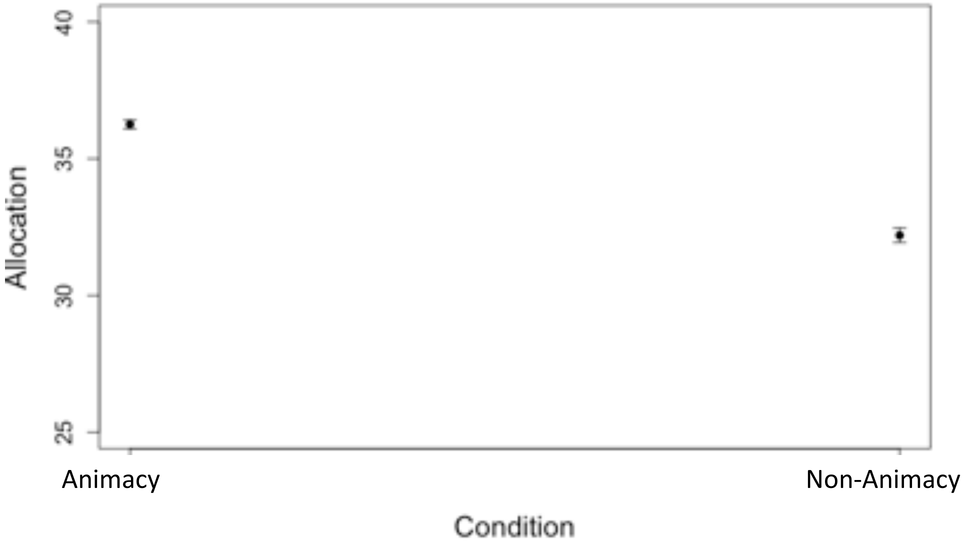


Figure 3. The average allocation and 95% CI of the trust model's predictions of the participants' behavior in animacy and non-animacy conditions.

Hypothesis 5: *We predict that there will be an two way interaction between the confederate agents' strategy (high and neutral trustworthiness) and identity.*

H5a. We predict a greater positive difference between the participants' allocation in the animacy and non-animacy conditions during the confederate agents' use the trustworthiness neutral strategy compared to the high trustworthiness strategy.

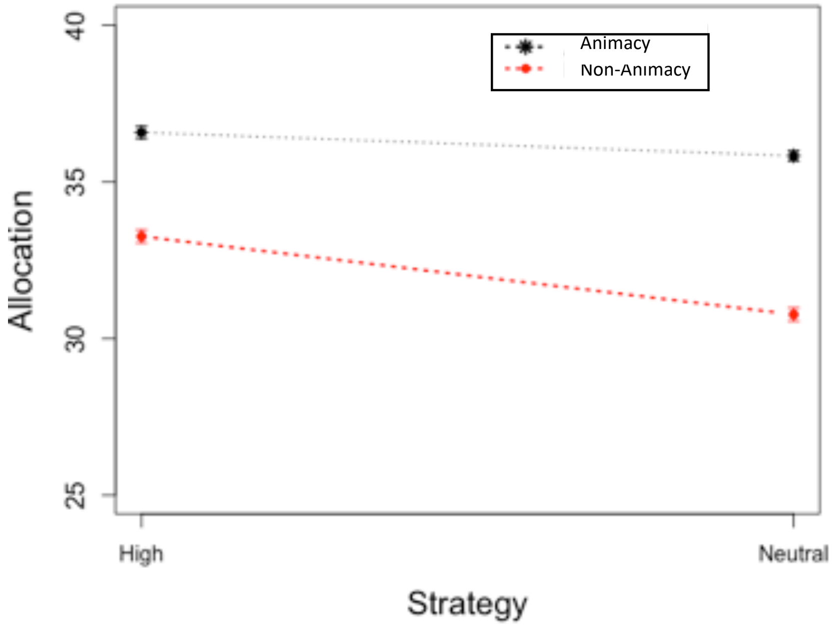


Figure 4. The average allocation and 95% CI of the trust model's predictions of the participants' behavior in animacy and non-animacy condition, while the confederate agents use the high and neutral trustworthiness strategy.

Hypothesis 6: *We predict that there will be a significant three-way interaction between the identity of the confederate agent (i.e., animacy and non-animacy), the confederate agents' interaction schedule (high, med, and low), and confederate agents' strategy (high vs neutral trustworthiness (Figure 5).*

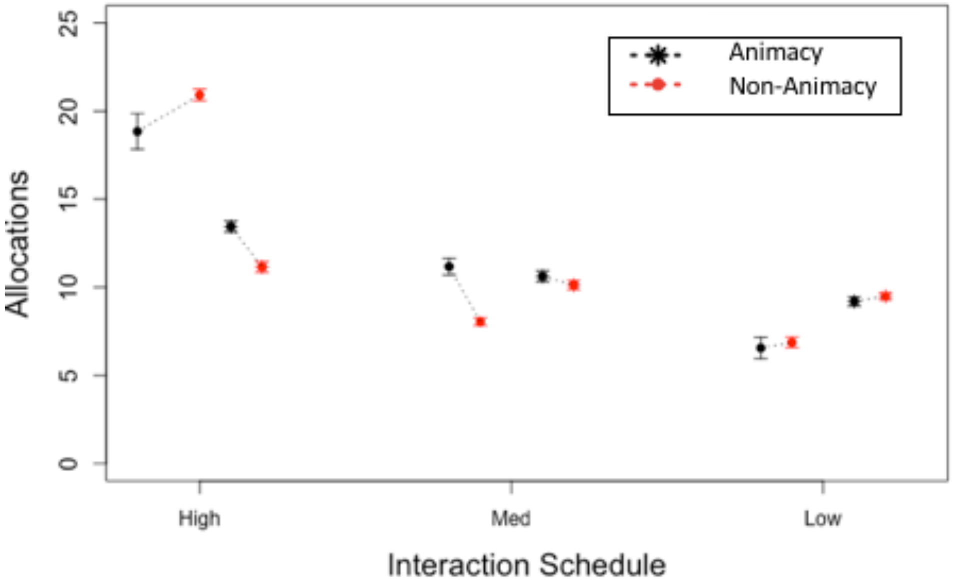


Figure 5. The trust model's predictions of the average and 95% CI allocation during the animacy (solid dot) and non-animacy (star) while the confederate agents on the high, medium, low interaction schedule use the high (black) and neutral trustworthiness strategy (red).

In addition to these 6 hypotheses about the participants' behavior, four additional predictions, based on previous literature, regarding the participants' response to the state and trait trust survey measures were made. I predicted that participants would use two constructs to govern their behavior (i.e., trait trust and state trust), but that these constructs would be moderated by the confederate agents' identity, interaction schedule, and trustworthiness. Trait trust is one's general predisposition to trust others and has been shown to inform a trustor's initial choices with a trustee. Individuals have shown to initially place a greater trust in an automated agent compared to another human. State trust is one's trust in a specific individual in a specific situation and develops over time based on the trustworthiness of a trustee. Individuals have been shown to have more resilient trust in humans compared to automated agents (de Visser et al., 2016). Based on the differences between when trait and state trust are thought to influence behavior, I predict that participants' trait and state trust survey results would correlate with the participants' behavior at different points during the game.

Hypothesis 7: *The participants' trait trust in the animacy condition will positively correlate with the participants' overall allocation of points sent to the four confederate agents during the first round.*

Hypothesis 8: *We predict that the participants' state trust in each of the confederate agents will positively correlate with the participants' average allocation.*

Hypothesis 9: *We predict that the participants state trust in the animacy condition in each of the three confederate agents will be moderated by the confederate agents' interaction schedule.*

Hypothesis 10: *We predict that the participants state trust in the confederate agents will be moderated by the confederate agents' identity.*
H10a: *We predict that the participants in the animacy condition will have a higher level of trust in the confederate agent on the high interaction schedule that the non-animacy condition.*

> ***H10b:*** *We predict that the participants in the animacy and non-animacy condition will have the same level of trust in the confederate agents on the medium and low interaction schedule.*

II. METHOD

PARTICIPANTS

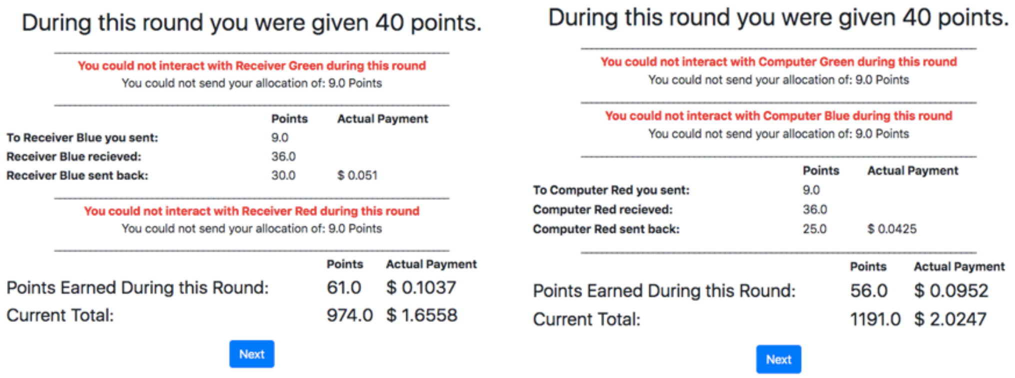
Forty four (Age: M = 38.25, SD = 11.8, Gender: 17% female) participants were recruited from the website Amazon Mechanical Turk (AMT) to take part in this study. All participants were evenly split between the two experimental condition. Participants received a base payment of $10 for taking part in the study and earned up to an additional $10 based on their performance during the game. The average total experimental payment for the experiment was $14.48.

EXPERIMENTAL TASK

The experimental task used in this study was the multi-arm trust game (MATG). The MATG is a game of strategic interaction combining features of two different games, the multi-arm bandit game (Robbins, 1952) and the trust game (Berg, 1995). The MATG is played between 4 players who interact repeatedly. One of the four players is randomly assigned the role of the Sender while the other three players are assigned the role of the Receiver. Over a series of rounds in the MATG, each player makes a set of decisions depending on their role in the game. At the start of each round both the Sender and Receiver each make an initial decision. First, the Sender is given a per-round endowment of 40 points. The Sender is then allowed to freely allocate their 40 point endowment between themselves and the Receivers. The Sender can give as much or as little of the 40 points as they wish to either themselves or to any of the 3 Receivers. As the Sender allocates their per-round endowment, each Receiver must decide to interact or not to interact with the Sender. If a Receiver decides not to interact with the Sender, then the Receiver will earn a random number of points selected from a distribution that is unknown to the Receiver. If a Receiver decides to interact with the Sender, then the Receiver will be given the number of points allocated to them by the Sender multiplied by 4. For example, if a Receiver decides to interact with a Sender and the Sender allocated 4 points to that Receiver, then the Receiver would be given 16 points. Additionally, Receivers who choose to interact with the Sender are allowed to return any number of their received multiplied allocation to the Sender. After all the Receivers have made their respective choices, the Sender is then notified of the choices made by each of the Receivers for that round. If the Sender allocated points to a Receiver who chooses not to interact with the Sender during that round, then the Sender is notified that they could not send their points to the Receiver during this round and the Sender is given back the points allocated to the Receiver. If a Sender allocated points to a Receiver who chooses to interact with the Sender, then the Sender is notified about the number of points allocated to the Receiver, the multiplied number of points that the Receiver was given, and how many points the Receiver returned to the Sender. The Sender is also told the total number of points earned during a given round. After the Sender observes the information about the Receivers the next round begins and the same procedure is repeated.

EXPERIMENTAL MANIPULATIONS

During each experimental condition (i.e., animacy and non-animacy) all participants played the role of the Sender with the same 3 confederate agents playing the role of the Receivers. Additionally, each experimental condition had a unique narrative to be consistent with the identity of the confederate agents. In the animacy condition, participants were told that they are 1 of 4 participants that have been recruited to participate in this experiment. Each participant in the animacy condition were be told that they have been "randomly" selected to play the role of the Sender in the experiment, while the 3 other "participants" were assigned to play the role of the Receiver. Additionally, during the MATG, when one Receiver chose not to interact with the participant, participants were told the Receiver chose not to interact with the participant and they could not be given their allocation during that round (Figure 6). In the non-animacy condition, participants were told that they were interacting with 3 separate computer algorithms that were developed to play this game. As in the animacy condition, each time the confederate agent chose not to interact with the participant, the participant was notified in the same way as in the animacy condition (Figure 6).



During this round you were given 40 points.

You could not interact with Receiver Green during this round
You could not send your allocation of: 9.0 Points

	Points	Actual Payment
To Receiver Blue you sent:	9.0	
Receiver Blue recieved:	36.0	
Receiver Blue sent back:	30.0	$ 0.051

You could not interact with Receiver Red during this round
You could not send your allocation of: 9.0 Points

	Points	Actual Payment
Points Earned During this Round:	61.0	$ 0.1037
Current Total:	974.0	$ 1.6558

During this round you were given 40 points.

You could not interact with Computer Green during this round
You could not send your allocation of: 9.0 Points

You could not interact with Computer Blue during this round
You could not send your allocation of: 9.0 Points

	Points	Actual Payment
To Computer Red you sent:	9.0	
Computer Red recieved:	36.0	
Computer Red sent back:	25.0	$ 0.0425

	Points	Actual Payment
Points Earned During this Round:	56.0	$ 0.0952
Current Total:	1191.0	$ 2.0247

Figure 6. An example of the results page in the multi-arm trust game shown to participants in the animacy (left plot) and the non-animacy condition (right plot).

SURVEY MEASURES

During the study, participants in the animacy and non-animacy conditions answered a set of two survey measures.

Trait trust measure. To measure a participants' trait trust, a 24-item questionnaire using items from two different trait trust surveys from Rotter (1967) and Yamagishi (1986) (Appendix A). All items are rated on a 1 (strongly disagree) – 5 (strongly agree) graphical interface scale. An example of an item used on the trait trust survey is "Most people are basically honest".

STATE TRUST MEASURE. To measure participants' state trust, a 14-item state trust survey using items from Collins et al. (2015) (Appendix B). All items are rated on a 1 (strongly disagree) – 5 (strongly agree) graphical interface scale. An example of an item used on the state trust survey is "Receiver 1 can be trusted".

BEHAVIORAL MEASURES

During the experiment two aspects of the participants' behavior during the MATG was measured: (1) the amount of points that participants allocated to each of the confederate agents and (2) the amount to time (milliseconds) that a participant spent on each page.

PROCEDURE

For this study, all participants were recruited from the website Amazon Mechanical Turk (AMT). Participants signed up for the experimental session and were given a link to go to and complete the experiment. Before beginning the experiment all participants gave their consent to participate. Afterwards participants went on to read the instructions for the study and to take a trait trust survey. Following the initial survey measures, participants were given instructions on how to play the multi-arm trust game. After reading the instructions for the experiment, participants went through and played 120 rounds of the MATG. After completing the game, all participants again took the state and trait trust survey. After completing the experiment participants were debriefed to the true nature of the study and paid for both their time and performance (Figure 7).

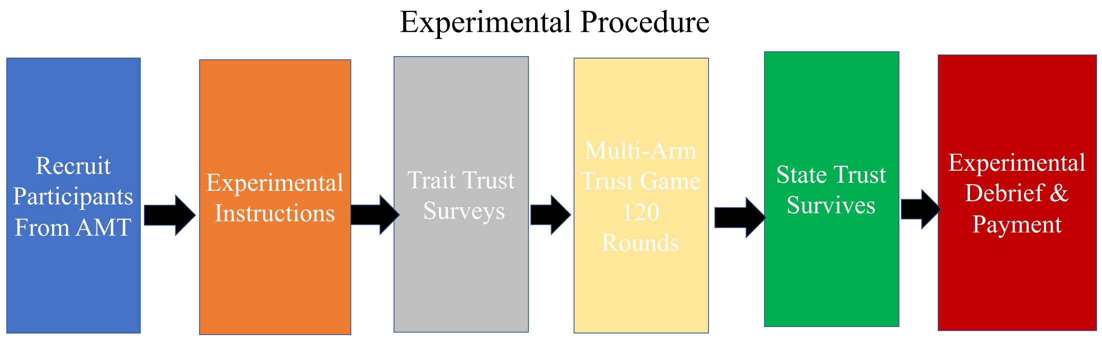

Experimental Procedure
Recruit Participants From AMT
Experimental Instructions
Trait Trust Surveys
Multi-Arm Trust Game 120 Rounds
State Trust Survives
Experimental Debrief & Payment

Figure 7. A diagram of the experimental procedure for the study.

III. RESULTS

The goal of this study was to evaluate the *ex-ante* predictions made by the trust model of Juvina et al. (2019) for the MATG. Prior to data collection, the model made round by round predictions of the participants' average behavior with each of the confederate agents over the course of the experiment. Additionally, these predictions were used to develop a series of hypotheses about how participants' behavior would be affected by the experimental manipulations. In this section, we evaluate both the model's ex-ante predictions as well as each of the experimental hypotheses.

MODEL PREDICTIONS

The trust model's predictions of the participants' average behavior were evaluated by comparing the average round by round allocation of the model and participants to the three confederate agents (Figure 8). The model's predictions were compared to the human data using two common fit statistics, correlation (*r*) and the root mean squared deviation (*RMSD*). Overall, a significant positive correlation,($r(718) = .50$, $p < .01$, *Cohen's d* = *medium effect*, *RMSD* = 9.05) between the average round by round allocation of the participants' behavior and the model's predictions was found. The finding suggests that the model was able to predict particular trends in the participants' behavior. Although the relatively large RMSD between the participants and the model data highlights discrepancies between the model and participants' allocations to the confederate agent (Figure 8), these discrepancies were found to differ between the animacy and non-animacy conditions.

The model's predictions of the animacy condition, (*r*(358) = .69, *p* < .01, *Cohen's d = large effect)*, *RMSD* = 7.62 (Figure 8 left column), were found to best account for human behavior while the confederate agents used the high-trustworthiness strategy (rounds 1-70), even though the model overestimated the extent to which participants would allocate their per round endowment to the confederate agent on the medium interaction schedule. The accuracy of the model's predictions decreased after the confederate changed its strategy from high to neutral trustworthiness. The animacy model predicted that after the confederate agents changed their strategy the participants' change in allocation would be minimal. However, participants quickly decreased their allocation to the confederate agents on the high interaction schedule (Figure 8).

In the non-animacy condition, (*r*(358) = .73, , *Cohen's d = large effect*, *RMSD* = 10.29), the accuracy of the model's predictions while the confederate agents used the high and neutral trustworthiness strategy, were opposite of the animacy condition (Figure 7). During the high trustworthiness condition, the model predicted little differentiation between the participants' average allocation across the three confederate agents. However, participants allocated a majority of their per-round endowment to the confederate agents on the high interaction schedule. Despite the model not capturing human behavior well during the high trustworthiness condition, the model did predict human behavior fairly well after the confederate agents switched to the trustworthiness neutral strategy. After the confederate agents' strategy shift, participants quickly decreased their allocation to a stable amount across all confederate agents, in line which the predictions of the non-animacy model.

Overall, examining the accuracy of the *ex-ante* predictions of Juvina et al. (2019)'s to the human data collected reveals interesting findings. The overall positive correlation between human data and the model predictions suggested that the model predicted various aspects of human behavior during the experiment. This is in line with prior work examining the extent of *ex-ante* model predictions (Collins, 2015). We now evaluate the hypotheses the model made for both the behavioral and survey results.

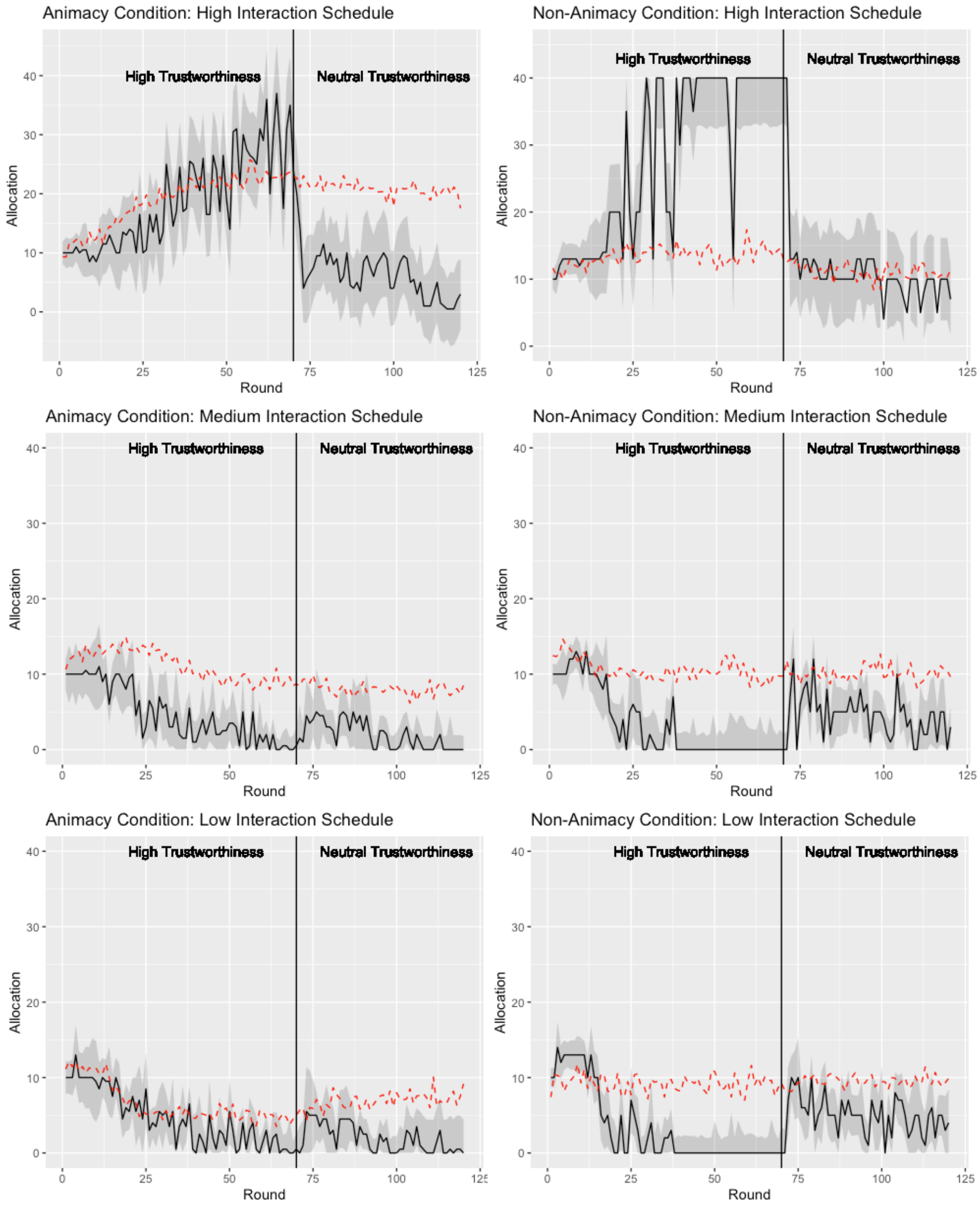


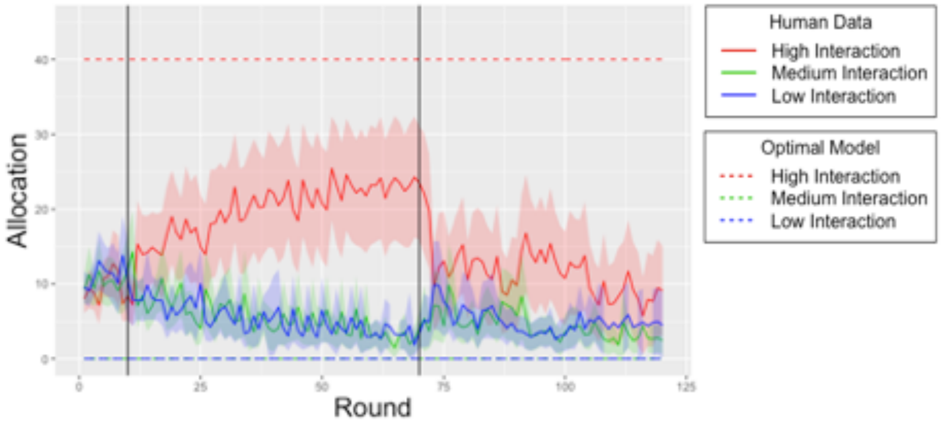
Figure 8. The average round by round (black line) and 95% confidence interval (gray ribbon of the participants' allocation to each of the three confederate agents (High

interaction - top most plot, Medium interaction - middle plot, and Low interaction - bottom most plot) in the animacy condition (left panel) and non-animacy condition (right panel), as well as the average per round ex-ante model predictions (dashed red line).

OPTIMAL MODEL

In addition to the trust model's predictions, the participant's average allocations was compared against an optimal model (Figure 9). The behavior of the optimal model was based on the assumption that the agent was omniscient and had full knowledge of the confederate agents' interaction schedule and strategy. Under this assumption, we assumed that the optimal strategy would be to allocate the full per round endowment to the confederate agent on the high interaction schedule during every round. Allocations to the confederate agents on the medium and low interaction schedule is assumed to be unnecessary due to the fact that their behavior is the same as the high confederate agent. Because when the confederate agents changed strategies the optimal agent does not lose points on average, we predicted that even after the change from high to the neutral trustworthiness that the optimal agent would still allocate to the confederate agent on the high interaction schedule. From this assumption, the optimal model was found to have an overall high correlation with the overall behavior across all of the data, but did have a higher RMSD compared the *ex-ante* predictions from the trust model (*r*(718) = .80, *p* < .01, *Cohen's d = large effect* , *RMSD* = 14.49). This finding suggests that the human data trended in the direction of the optimal strategy, but the high RMSD shows a high deviation from the optimal model's choices. The participants were found to allocate the greatest portion of their endowment to the confederate agent on the high interaction schedule during the first portion of the experiment, but never to the extent predicted by the optimal model in both the animacy and non-animacy condition. The likely reason for this deviation is that the participants had to learn through experience both the reward and interaction schedule of the confederate agents, leading them not to fully allocate their full endowment consistently to the confederate agent on the high interaction schedule.

Animacy Condition



Non-Animacy Condition

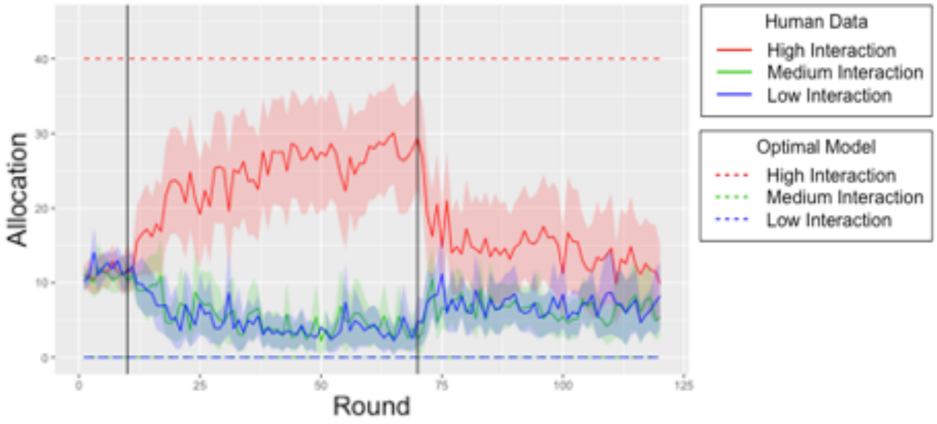


Figure 9. The average round by round and 95% confidence interval of the participants (solid line and ribbon) allocation and the optimal model allocations (dashed line) on the high (red), medium (green) and low (blue) interaction schedule

BEHAVIORAL MEASURES

To assess the hypotheses of our study a linear mixed-effects model was used. To determine the mixed effects that were included in the linear mixed-effect model, three linear mixed effects models were compared against each other using the Akaike Information Criterion (AIC). Each of the three models regressed the participants' per round allocation onto the same four factors (i.e., round, confederate agents' interaction schedule, strategy, and identity), varying only in their mixed effects[6].

The first model was a standard fixed-effects model (i.e., fixed effect model). The second model included a unique intercept for each participant (i.e., random intercept model). The third model included a unique intercept for each participant and a slope for each of the participant's allocations to the three confederate agents (random intercept and slope model)[7].

These three models were compared using *AIC* which considers the variance accounted for by a given model relative to each model's free parameters (Burnham & Anderson, 2003). AIC punishes models that are overly complex favoring simpler models (Burnham & Anderson, 2003). Of the three models compared, the random intercept and

[7] The formulas for the model and the R code are included in an Appendix F.

slope model (LL(43) = -49760.41 , *AIC* = 97215.27) had the lowest AIC relative to the random intercept (LL(38) = -49583.93, *AIC* = 99243.87, p < .01) and fixed effect model (*LL*(37) = -49760.41,*AIC* = 99594.83, p < .01); both of the latter models were found to be significantly different from the random intercept and slope model (*L.Ratio* = 352.9594 <.0001; *L.Ratio* = 2038.5977 <.0001). From this comparison of *AIC* values, I concluded that despite the additional free parameters of the random slope and intercept models, the lower AIC values suggest the additional complexity of the model allows for a more comprehensive explanation of variance in the data[8]. From these results, the random slope and intercept model was used to test for the main effect and interactions of the experimental manipulation to assess the ex-ante hypotheses (Appendix E). Further post-hoc tests were guided by the model's ex-ante predictions.

RATE OF ALLOCATION

Over the course of the experiment, the model predicted that through interacting with the three confederate agents participants would learn to have different degrees of trust based on the confederate agents' different features (i.e., strategy, identity, and interaction schedule). The overall level of trust in the three confederate agents would then affect the total amount of endowment that the participants allocated per round. The model

[8] Due to unbalanced gender representation in our sample, the additional factor of gender was explored as an additional source of variance in our data. However, when a linear mixed effect model with gender added as an additional factor and compared against to the random intercept and slope model, no significant difference was observed between the two models ((*L.Ratio* = 5.295638, *p* = .07).

hypothesized that participants' total per round allocation over the course of the experiment would interact with the confederate agents' strategy.

Consistent with this hypothesis a significant interaction was observed between the confederate agents' strategy and round, *F*(1, 13266) = 6.2523, *p* < .0124, *f* = .02). While the confederate agents used the high trustworthiness strategy, the model predicted that participants would allocate an increasing positive amount across the three confederate agents over the course of the experiment and decrease their allocation while the confederate agents used the trustworthiness neutral condition. A post-hoc test revealed the rate of the participants' allocation to be consistent only with the predicted rate of allocation during the low trustworthiness strategy. During the high trustworthiness condition, participants slightly decreased their average allocation across the confederate agents (*B* = -0.0015), opposite of the predicted effect, while during the low trustworthiness strategy participants decreased their overall allocation to the confederate agents. (*B* = -0.13379) (Figure 10), consistent with the predicted effect.

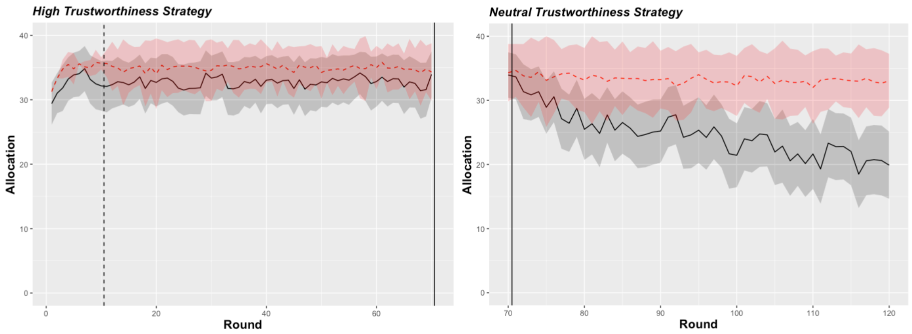


Figure 10. The average and 95% CI for both the participants (solid black line and grey ribbon) and the model predictions (dashed red line and red ribbon) for the high trustworthiness strategy (left panel) and neutral trustworthiness strategy (right panel).

In addition to the model predicting an interaction between round and strategy, the model also predicted that the rate of allocation to the confederate agents would interact with the confederate agents' interaction schedule. Following this hypothesis, a significant interaction between Round, interaction Schedule, and Strategy was observed, F(2, 13266) = 11.5236, $p < .01$ $f =$ .04) , but further analysis indicates a departure from the expected direction. The model predicted that the rate of the participants' allocation to the confederate agents would positively correlate with the confederate agent's interaction schedule. Though participants had the highest rate of allocation to the confederate agent on the high interaction schedule, no significant difference was observed in the rate of allocation between the medium and low interaction schedule (Figure 11). During the second portion of the experiment, where the confederate agent used the trustworthiness neutral strategy the model predicted the opposite effect of allocation to the three confederate agents on the different interaction schedules. However, again the only difference in the rate of allocations was observed between the high interaction condition and the medium and low interaction condition. Based on these findings, we find no evidence to support H1, find partial support for H1a. Furthermore, we accept H2 and find partial support for H2a.

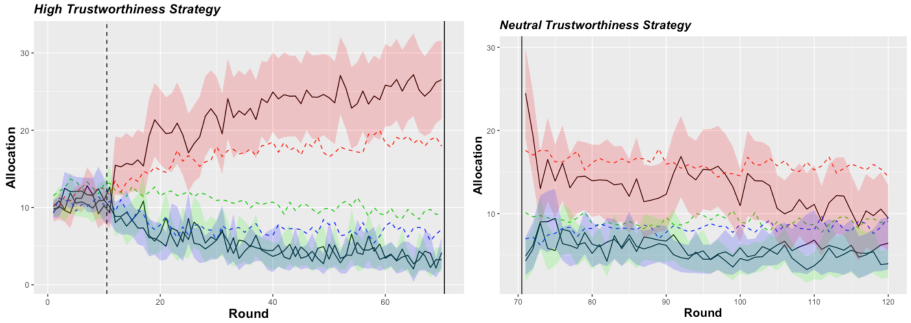


Figure 11. The average and 95% CI round by round allocation for the human (solid line) and model (dashed line) to the confederate agent on the high (red), medium (green), and low (blue) interaction schedule, while the confederate agents used the high (left plot) and neutral (right plot) strategy.

CONFEDERATE AGENT INTERACTION SCHEDULE

As was seen within the interaction between round, confederate agents' strategy, and interaction schedule on the rate of the participants' allocations, participants' overall allocations were sensitive to the confederate agents' interaction schedule. Our trust model predicted that the participants' trust would be sensitive to the confederate agents' interaction schedule during the experiment, because participants' were predicted to discount their trust during instances where they attempted to interact with the confederate agent but the confederate agent did not interact with them. From this assumption, the model predicted a positive relationship between the participants' allocation and the confederate agent's interaction schedule. As predicted a significant main effect of the confederate agents' interaction schedule was found (*F*(2, 13266) = 19.30, *p* < .01 *f* = .05) In line with our pilot study (Appendix C), allocations to confederate agents differed according to the confederate agent's interaction schedules, the largest difference in the participants' total average allocation to the confederate agents was between the confederate agent on the high (*M* = 15.58, *SD* = 14.93) interaction schedule and the other confederate agents (i.e., medium *M* = 5.48, *SD* =7.60 and low *M* = 5.64, *SD* = 7.13 interaction schedule, Figure 12). No difference in the participants' total allocation to the confederate agents on the medium and low interaction schedule as predicted was observed. From these results, I find partial support for *H*3.

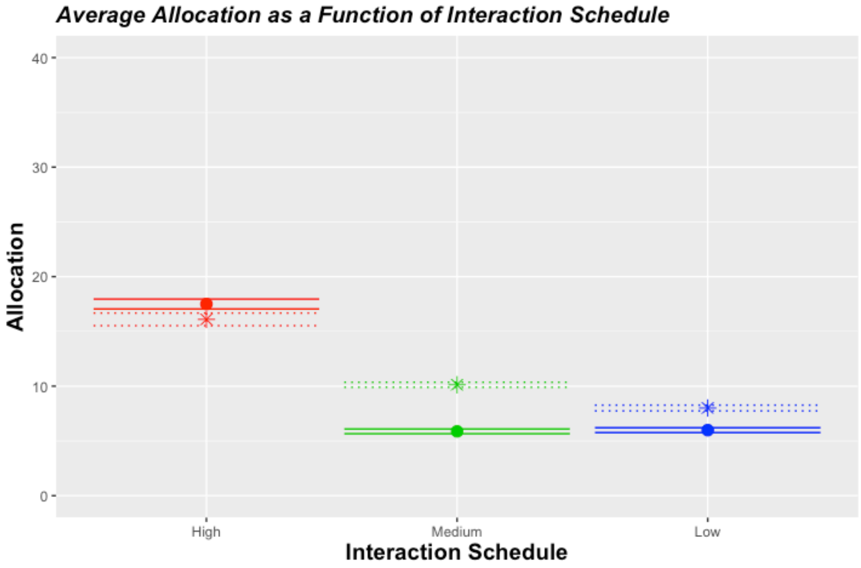


Figure 12. The average and 95% CI allocation across rounds and participants by participants (solid dot) and model (star) to each of the three confederate agents on the different interaction schedules.

CONFEDERATE AGENT IDENTITY

In addition to a main effect of the confederate agent's interaction schedule, the model predicted that the participants' allocations would be sensitive to the confederate agents identity (animacy and non-animacy). The predicted difference between the animacy and non-animacy model stems from the assumption that participants would discount the trust in confederated agents in the non-animacy condition more than the animacy condition. From this assumption, the model predicted that participants would allocate a greater portion of their endowment over the course of the game in the animacy condition compared to the non-animacy condition. A significant effect of the confederate agent's identity (animacy vs non-animacy) was found ($F(1, 13266) = 112.69, p < 05, f = .09$). However, participants were found to allocate less of their endowment to the confederate agent in the animacy ($M = 8.9$, $SD = 11.59$) compared to the non-animacy ($M = 10.62$, $SD = 12.47$) condition (Figure 13). Though there was a significant interaction effect of animacy, the observed effect was in the opposite direction. Thus no evidence was found to support H4.

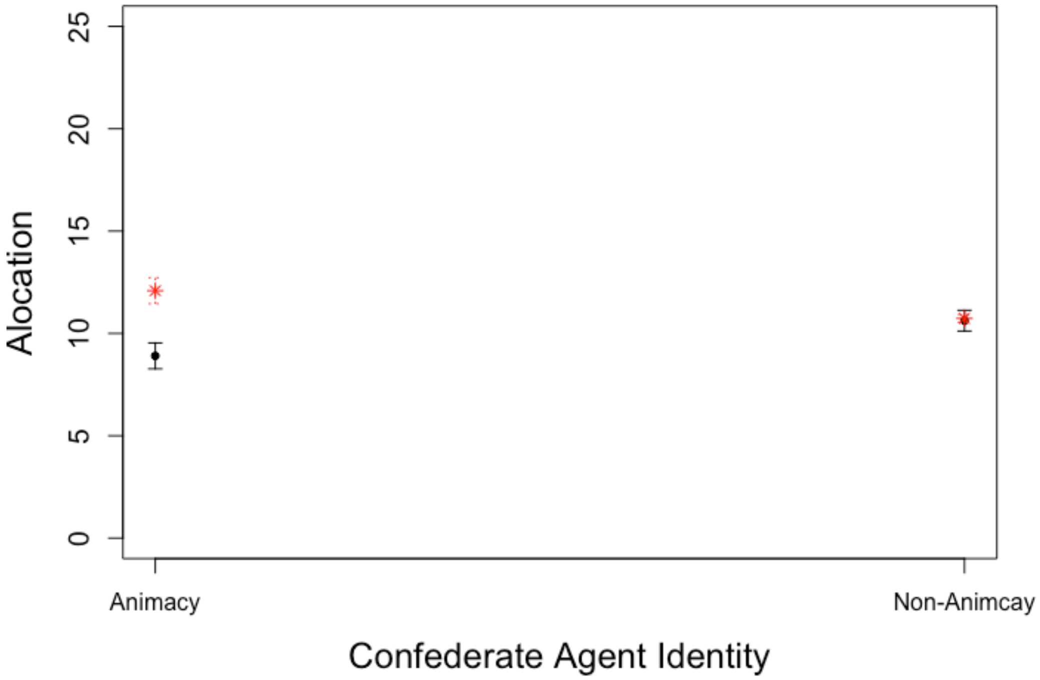


Figure 13. The average and 95% CI allocation by participants (black solid dots) and model (red star) in the animacy and non-animacy condition.

CONFEDERATE AGENT STRATEGY

Next, we examined the extent that the participants' average allocation was affected by the confederate agents' two strategies (i.e., high or neutral trustworthiness) over the course of the experiment. The model predicted a two-way interaction between the confederate agents' identity (animacy and non-animacy) and strategy (high and neutral) but no significant interaction effect was found ($p < .05$). However, a significant effect main effect of strategy was found ($F(1, 13266) = 6.99$, $p < .01$, $f = 0.02$). As expected, participants decreased their allocations to the confederate agents after they switched from the high ($M = 10.90$, $SD = 12.72$) to the neutral ($M = 8.22$, $SD = 10.85$) trustworthiness strategy (Figure 14). Due to the fact that no significant interaction effect was observed between the confederate agent's strategy and identity we find no evidence to support *H5*, *H5a*, or *H5b*.

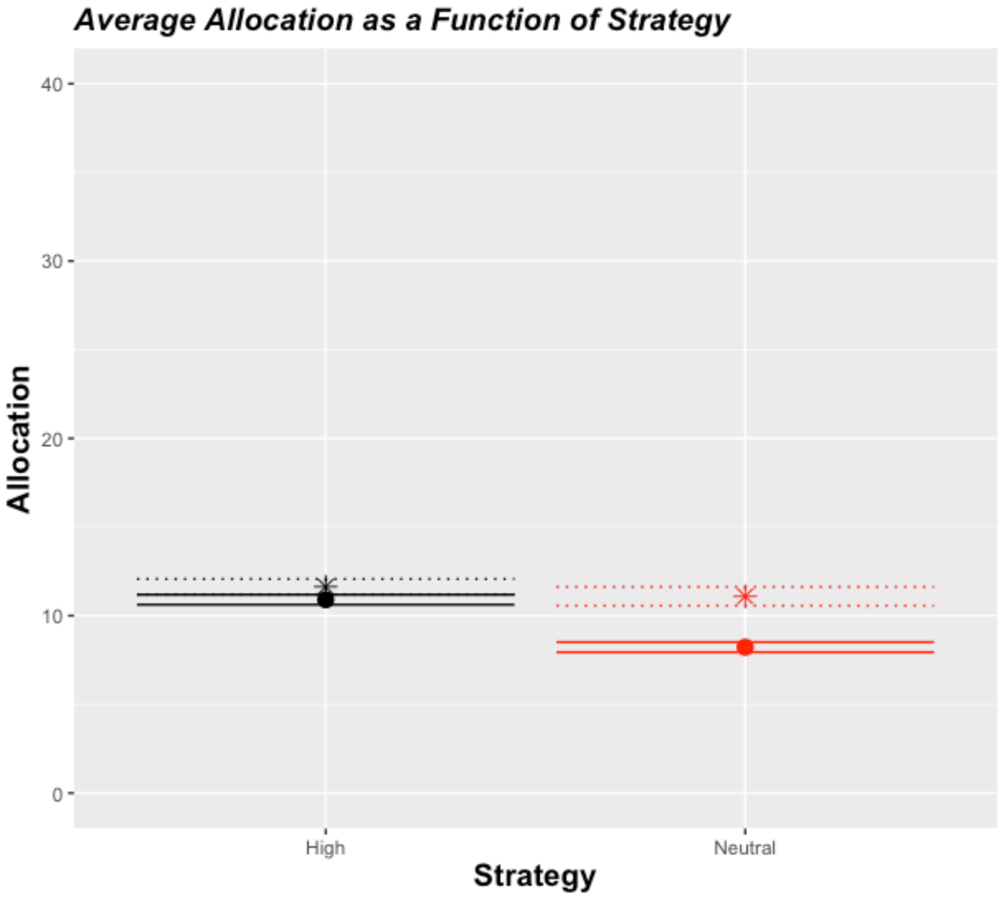


Figure 14. The average allocation of participants (solid dots) and model (star) while the confederate agents used the high and neutral trustworthiness strategy.

.

Finally, the model predicted a significant three-way interaction between each of the confederate agent's characteristics (i.e., identity, strategy, interaction schedule).

As seen in the results reported so far, the effect of the confederate agents' identity was not found to be significant. However, a significant two-way interaction between the confederate agents' interaction schedule and strategy was observed (F(2,13266) = 240.2312 , *p* < .01, *f* = .19). Over the course of the experiment, participants' only changed their average allocation to the confederate agent on the high interaction schedule between the high (*M* = 20.60 , *SD* = 15.38) and neutral (*M* =13.50 , *SD* = 14.22) condition (Figure 15). Given that there was no significant three-way interaction we find no support for *H6.*

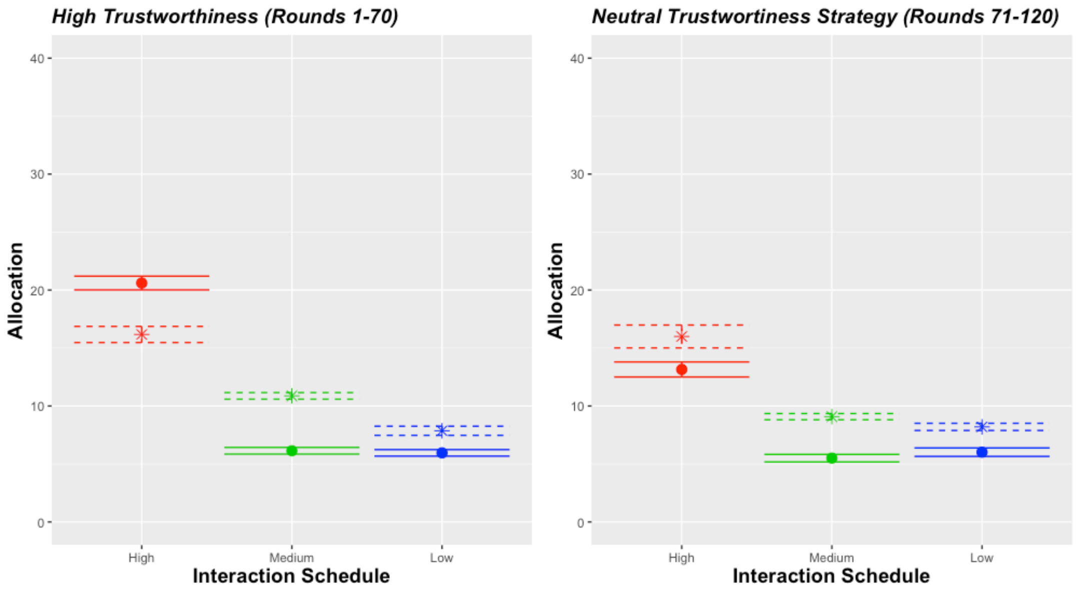


Figure 15. The average and 95% CI allocation by participants (solid dots and solid error bars) and model (star and dashed error bars), during the high trustworthiness strategy (left plot) and neutral trustworthiness strategy (right plot) to the three confederate agents on the three interaction schedules (High – red, Medium – green, Low – blue).

SURVEY MEASURES

Trait Trust

Before participants interact with the confederate agents, it is thought that the participant's initial allocation decision was governed by their trait trust. For this reason, we predicted an overall positive correlation between participants' trait trust and 1st round allocation. Additionally, we predicted a difference between the strength in the correlation between animacy and non-animacy condition. As predicted, a significant positive correlation (*r* (35) = .33, *p* < .04, *Cohen's d* = *medium effect*) was found between the participants' first round allocation and their trait trust measures (Figure 16). However, no significant correlation was found within the animacy (*r*(18) = .44, *p* = .07) or non-animacy (*r*(17) = .13, *p* = .56) conditions. From these results, we accept *H*7 but find no evidence to accept *H7a.* Furthermore no significant correlation was observed between the participant's trait trust and average allocation over the course of the experiment (*r(35)* = -.06, *p* = .70). Taken together these findings suggest that participants' trait trust influenced their initial allocation and we did not find evidence to suggest that trait trust influenced their behavior over the course of the experiment.

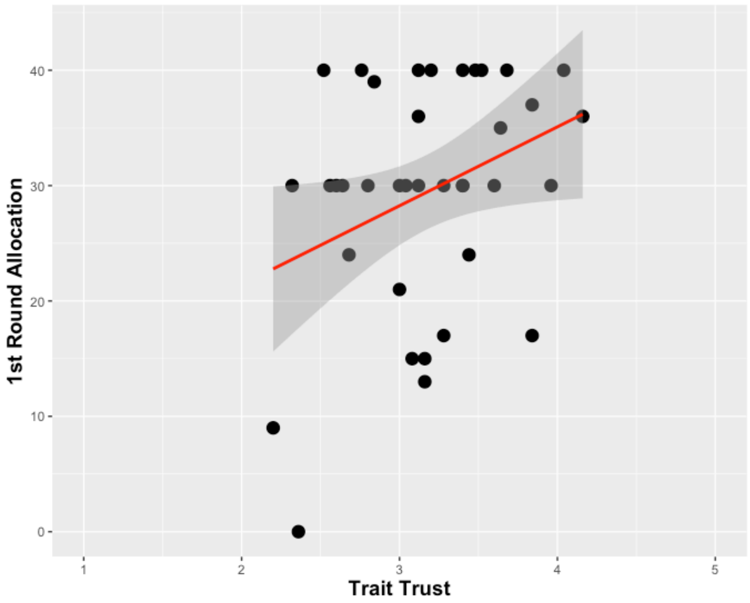


Figure 16. The participants total 1st round allocation as a function of their trait trust (black dots).

State Trust

After participants began to interact with the confederate agents, we predicted that participants would learn to trust the confederate agents based on their behavior over the course of the experiment. From this assumption, we predicted that there would be a positive correlation between the participants' state trust in the confederate agents and their average allocation to each of the three confederate agents. In line with this prediction, a significant positive correlation (*r(109)* = .44, *p* < .01, *Cohen's d = medium effect*) (Figure 17) between participant's average allocation to each of the confederate agents and their state trust for the confederate agent was found (Figure 16). The positive correlation between average allocation and state trust suggests that the participants' behavior was motivated by their state trust in a confederate agent. Additionally, we did not find evidence to suggest that participants' trust influenced their first round allocation (*r* = .08, p = .35).

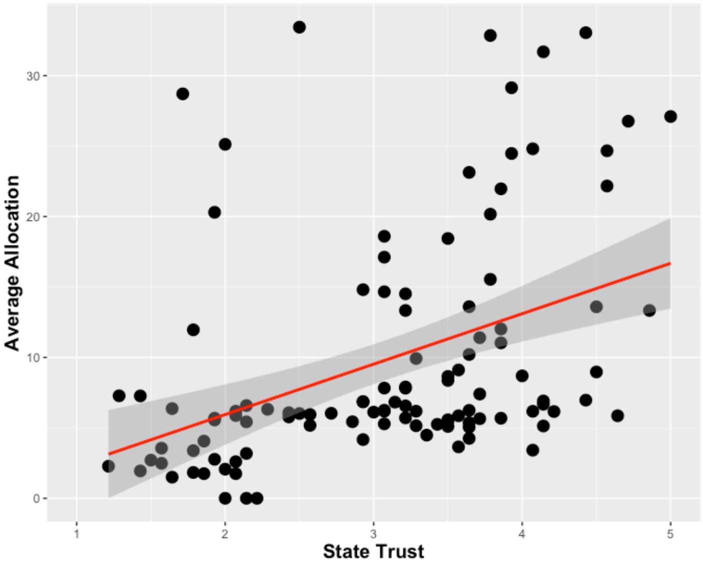


Figure 17. The participants' average allocation to a confederate agent as a function of their state trust.

In addition to state trust being positively correlated with the average allocation, we also predicted that state trust would be influenced by the confederate agent's interaction schedule and identity. To examine the extent that these variables would influence the participants' state trust, an additional linear mixed effect model was run, regressing each participant's state trust onto the confederate agent's interaction schedule (high, medium, and low) and identity (animacy non-animacy), with each participant being given a random intercept. The model found a significant effect of the confederate agents' interaction schedule (F(2,87) = 9.42, $p < .01$, $f = .47$), but not identity ($p = .96$). Mimicking participants allocations to the confederate agents, state trust was highest in the confederate agent on the high ($M = 3.54$, $SD = .86$) interaction schedule, with no significant difference being observed between the state trust of the confederate agents on the medium (*M* = 2.90, *SD* =.86), and low schedule (*M* =2.76 , *SD* = .93) (Figure 18). From these results, we find partial support for H9, no evidence was found to accept H10 or H10a, and we confirm H10b.

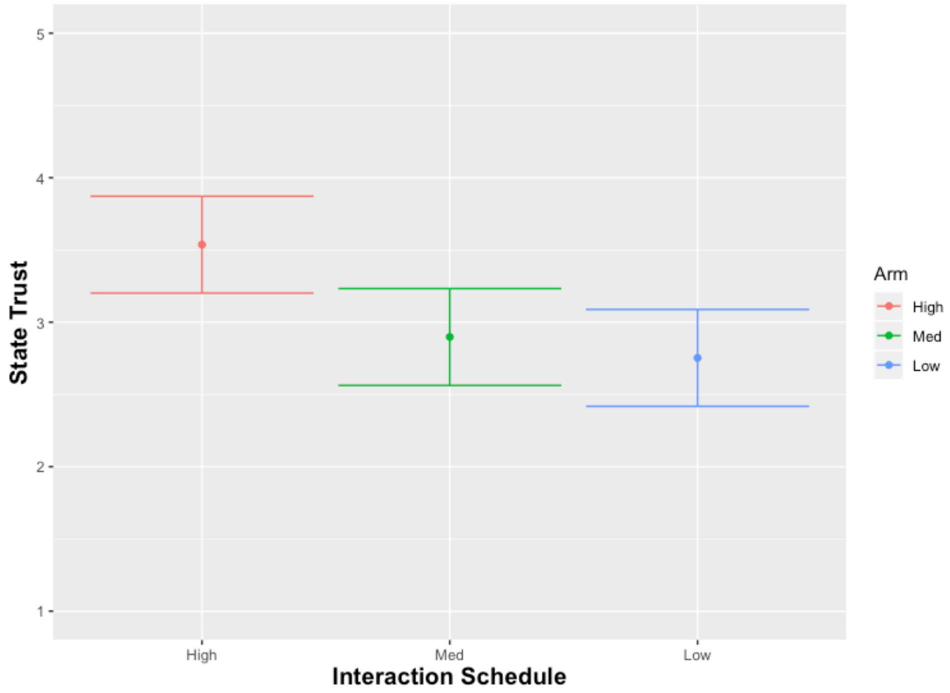


Figure 18. The state trust for each of the confederate agents on three interaction schedules (high- red, medium-green, and low - blue).

IV. DISCUSSION

Trust research has focused primarily on how trust influences decisions during short or one-shot interactions between two people. The typical experimental designs used to assess trust development lack particular qualities of real-world social interactions, such as the capability of interacting with multiple people or interacting with people on a variable schedule. It are these features are common in the workplace, virtual team environments, and the growing use of autonomous systems. If theories on trust are to be used for real-world applications, the interaction with these real-world features on trust needs to be better understood. One common feature of each of these types of scenarios is that they are examples of the exploration-exploitation dilemma, where an individual has multiple decision options each containing a variable potential reward. To better understand how individuals use their trust in others to make decisions under these constraints we developed the multi-arm trust game to assess how trust is influenced by multi-person variable interactions. Besides creating a new experimental design, a previously published trust model fit to pilot data was used to make ex-ante predictions of the participants' behavior in our experiment. In this discussion, I address the theoretical and methodological contributions of this study to both the trust and broader cognitive science literature.

THEORETICAL CONTRIBUTIONS

The theoretical contributions of this research were threefold. First, we examined the effect of the interaction schedule and trustworthiness on trust development. The data collected from our study suggested both interaction schedule and trustworthiness affected the participants' trust behavior. Participants allocated a majority of their allocation to the confederate agent on the high interaction schedule, relative to the medium or low interaction schedule. Second, participants' decreased their allocation to the confederate agent on the high interaction condition after the confederate agent changed from a high to a neutral trustworthiness strategy. Third, our results suggested that the confederate agent's identity did affect participant's allocations, but in the direction opposite of what was predicted. Participants allocated slightly more of their overall endowment in the non-animacy condition compared to the animacy condition, but the confederate agent’s identity was not found to interact with the other features of the experimental design. Finally, the model predicted a three-way interaction with the confederate agents’ strategy, identity, and interaction schedule, but no evidence was found to support this hypothesis. Only the confederate agent’s strategy and interaction schedule interacted with the participants’ allocation. Similar to the previous effects observed in our pilot study (Appendix C), the only significant interaction was found between the confederate agent on the high interaction schedule while using the high and neutral trustworthiness strategy. After the confederate agent’s strategy change from the high the neutral trustworthiness strategy participants decreased their allocation to the confederate agent on the high interaction. While no evidence was found to suggest that the participants' allocation to the confederate agents on the medium and low interaction schedule changed during while using the high and neutral trustworthiness strategy. Besides the behavioral data collected in the experiment, additional hypotheses were made about the participants’ responses to state and trait trust questionnaires and how they would relate to the participants' behavior. Our results found that both the trait and state trust questionnaire correlated with relevant aspects of the participants’ behavior. Trait trust had a positive relationship with the confederate agents’ first round allocation, but no evidence was found to suggest that trait trust was moderated by the confederate agent’s identity. State trust positively correlated with the participant’s average allocation to the confederate agent and was systematically moderated by the confederate agent’s interaction schedule (e.g., high, medium, and low), but not the confederate agent’s identity (e.g., Animacy vs. Non-Animacy). Furthermore, the effect of the interaction schedule was reflected in the participants’ allocations to the confederate agents based on their interaction schedule. The participants’ state trust was the highest for the confederate agent on the high interaction and lowest for the confederate agent on the medium and low interaction conditions, with no other significant interactions observed. Taken together these results support the assumption that trust was a motivating factor for the participants’ decisions throughout the experiment. Furthermore, our results highlight that both interaction schedules and neutral trustworthiness behavior are important to trust development and behavior.

In addition to the results of our study being of interest to trust research, the findings of this study are relevant to several other areas of cognitive science, such as deontological reasoning, theory of mind, causal reasoning, and risk perception. One idea that has been proposed to account for differences between human and computer interaction would be to posit that decisions made when interacting with other humans are the results of specific mechanisms developed to handle the complex human decision-making type task, such as logical reasoning problems, which are solved with a higher degree of accuracy when contextualized in a social situation to detect cheating (Cosmides & Tooby, 1992). From a deontological reasoning perspective, individuals would be thought to show different patterns of behavior in the animacy compared to the non-animacy condition. Although a significant difference in the average allocation between the animacy and non-animacy condition, no evidence was found to suggest a difference between the round by round allocations. This suggests the animacy and non-animacy condition was not found to have a significant effect on learning over the course of the experiment.

Furthermore the trust research as a whole does not appear to support the central claim of deontological reasoning. Although differences exist in the participants' trust behavior when interacting with other humans compared to machines, it is found these differences stem not from a fundamentally different mechanism, but instead from differences in attributions from the trustor to the trustee. Juvina et al (2019), found that differences between human-human and human-machine interaction could be accounted for by simply modifying the trust discounting parameter. These differences might reflect fundamental differences in deontic reasoning, but could also be accounted for by other aspects of behavior, such as theory of mind, causal reasoning, or risk perception.

The second area of adjacent interest is the theory of mind literature, which attempts to understand how individuals attribute mental states to others to explain and predict the behavior of others (Samson, 2013). This line of research is relevant to trust research because the trustworthiness of another individual is thought to be based on several different perceived personality characteristics (i.e, ability, benevolence, and integrity, Mayer et al., 1995). One particular view of theory of mind, which opposes the deontological reasoning literature is the theory-theory (Gopnik & Wellman 1992). Gopnik proposes that one's theory of mind is based on a theory where individuals attribute mental states to others in order to understand their behavior. From this theory, a better explanation of our research might be found. When a trustor interacts with a trustee, the trustor might prescribe particular mental states to the trustee and use this information to make a summary judgment to explain their behavior and infer trustworthiness. From this perspective, no fundamental difference is observed in the underlying mechanism when interacting with a human or machine, but simply a difference in the mental states that are projected to the different agents. This notion is supported by Gray, Gray, and Wegner (2007) that humans tend to project differences in mental abilities to humans, machines, robots, and other non-animated objects.

Besides the attribution of mental states to individuals, the third area of related research is causal reasoning. Causal reasoning is the process by which an individual comes to learn and infer the causal structure of an environment or domain to make judgments about the future. Research on causal reasoning has shown that it depends both on information about covariance and domain-independent knowledge (Cheng, 1997). From a causal learning standpoint, our multi-arm bandit game holds a great deal of relevance. Participants in this experiment must learn many different aspects of the task but must also learn the relationship between their allocation to a confederate agent and the return that they receive from the confederate agent. The quick decrease in the participants' allocation after the confederate agents' strategy shift from high to neutral trustworthiness suggest that participants were going beyond simple covariation information and had constructed some type of theory as to the causal relationship between their allocations to the confederate agent and the returns they received from the confederate agents.

The final adjacent area of interest related to our work is risk perception.

Risk perception deals with how an individual subjective assessment of the ratio of costs and benefits for a particular action (Weber & Milliman, 1997; Stewart, Chater, & Brown, 2006). Risk perception is affected by a variety of different factors such as the domain, information presentation, and personality. These findings are at odds with how trustworthiness was modeled in this study. In the current study, it was assumed that there was no difference in the participant's perception of risk attitudes since the trustworthiness of the confederate agent was solely a reflection of their return to the participant. However, research on risk perception has shown an individual's perceived risk of a particular domain influences their decision-making. The payoff returned by a confederate agent is affected by both the confederate agents' strategy and interaction schedule, both of these factors could have influenced a participant's perceived risk of interacting with a confederate agent, making them more or less likely to interact with the confederate agents on the medium or low interaction schedule. Though trust is often considered separate than risk, trust is thought to help lubricate scenarios with high risk. The participant's perceived risk of a given situation might interact with their current trust with a confederate agent affecting their allocation behavior.

METHODOLOGICAL CONTRIBUTIONS

In addition to the theoretical contributions of our work, this research also makes two major methodological contributions. The first methodological contribution dealt with the trust model and the level of granularity that the model was able to account for and predict. In this paper, we provided evidence that trust can be accounted for using a simple accumulator model. This finding is in line with previously published work (Juvina et al. 2019), which showed that the trust model could take into account a wide range of trust phenomena. This supports the notion that trust is developed over time and increases and decreases according to interaction schedule and the trustworthiness of other people. The benefit of this modeling approach is that the model can both account for and make predictions at the trial level of aggregation, instead of only predicting average performance across a range of conditions or simple differences. Attempting to account for human performance at these low levels of aggregation is more difficult, but also provides a stronger test of our model. In the data collected in this study, the model predicted the general trends observed in the data and but also failed to certain aspect of the participant's behavior, such as the less allocation to the confederate agents on the medium and low interaction condition and the participants sudden change in behavior after the confederate agents' change from the high to the neutral strategy. Though the model's predictions were found to have a lower correlation compared to our optimal model, the trust model provides a more probable account for human behavior because the optimal model's assumption of full knowledge is cognitively implausible. Our trust model shows how an individual can learn from experience and trend towards optimality. It is these subtle nuances in behavior that would have been lost had the focus of model predictions been at the level of overall behavior across conditions, compared to the average trial by trial level data.

The use of this modeling approach allows for a more rich analysis of the data and allows for a more rigorous test, that cannot be accomplished by verbal models of trust such as Mayer et al. (1995) or Sperber, Clément, Heintz, Mascaro, Mercier, Origgi, and Wilson (2010).

The second methodological contribution of our study deals with the experimental task, extending the paradigm of the trust game. The initial motivation of the trust game was to highlight the limits of standard normative decision making analysis, that simply looking at payoffs would not be able to account for human decision making (Berg et al. 1995). This initial research showed that aspects other than payoff incentives affect human decision making. Since this time, the trust game has been a useful tool for trust and decision making research, but limitations of that design already discussed at the beginning of this paper exist. Our modification and addition to the standard trust game paradigm combining it with the multi-arm trust game to allow it to include multiple different individuals and investigate non-continuous interactions. With these modifications, we were able to expand the trust game into domains that before it was unable to account for. This kind of contribution is a hallmark of progress in the literature of behavioral game theory, where additional features of simple games representing particular aspects of strategic interaction are adapted to account for more real-world scenarios (Camerer 2003).

Each of these methodological contributions furthered the goals of the trust literature as a whole and allows for a new paradigm that can be utilized by other areas. Accounting for trust using a simple accumulator model allows for the model to account for data at a higher level of fidelity that can be accounted for by simple verbal models of trust. The extension of the normal experimental design of the trust game and combining it with the multi-arm bandit game allows for more real-world scenarios on trust to be examined as well as offering a paradigm for other areas of research such as deontological reasoning, theory of mind, and risk perception.

PRACTICAL IMPLICATIONS

The results presented in this paper suggest that the rate of interaction between a trustor and a trustee influences trust development over time. All of the confederate agents in our study used the same strategy during the experiment, differing only in their rate of interaction between them and the participant. From this manipulation, differences in both participants' behavior and trust assessments with the confederate agents were observed. Furthermore, participants did not allocate their per round endowment proportionally according to the confederate agents' interaction schedule, instead, participants optimized their allocations placing a majority of their trust into the confederate agent on the high interaction schedule. These two findings taken together have implications for applied applications, such as teams or groups or automated technology. In group interactions such as virtual/in-person team or when a user is working with multiple automated systems, emphasis should be added to make such the lines of communication between all parties human or otherwise remain open. Failure for group members to respond to requests for interaction can lead to a decrease in trust and failure to interact. Additionally, if an individual optimizes their behavior, requests for assistance might be focused on the most responsive individual. In our experiment, the trustworthiness of the trustees was held constant, but in real-world interactions where trustworthiness might be variable between participants an individual might end up incorrectly misplacing their trust in a trustee who is the most responsive. Our research suggests that possible misattribution of trust might be mitigated by encouraging open lines of communication among individuals or giving individuals reasons for why a particular person cannot interact with a certain individual at a given time. Simple messages of explanations between parties have shown to be effective at decreasing misattributions of trust due to mistakes (de Visser et al., 2016).

LIMITATIONS AND FUTURE RESEARCH

The results of this study found confirming evidence for the assumption of the discounting mechanism proposed by Juvina et al. (2019)’s trust mechanism and showed that interaction frequency played a role in trust development. However, our experiment was also found to have several limitations that warrant future research. First, we focused on assessing the average behavior of participants throughout the experiment. An understanding of average behavior and basic behavioral trends in human behavior is often useful in many domains, but the extent of the conclusions that can be made from the analysis are limited. Model fits and predictions from average data ignore individual differences within a dataset and favor models that cannot account for individual decisions (Estes & Maddox, 2005). Our model assumed that all participants developed trust and used the information to make allocations in the same way. However, when faced with complex decisions multiple strategies are often seen across individuals and even within individuals (Lee, Gluck, and Walsh, 2019). Future research should explore individual differences in both trust development and allocation strategies used by participants. Second, in our effort to model the data we assumed that the confederate agents' behavior was a pure measure of trustworthiness. This assumption was a practical simplification for our modeling effort. However, trustworthiness is perceived and contextual factors do affect a trustor’s trustworthiness assessment of a trustee's behavior (De Melo et al. 2011). This contextual sensitivity in the trustors was an unexamined factor that could also account for the behavioral effects found in the dataset. Moreover, we did not take into account subjective risk perception as influencing the trustor's behavior. Individuals have shown to have various levels of risk-taking attitudes, which could affect their willingness to allocate their per-round endowment or expectations of the confederate agents. Future research should investigate the trustor’s perceptions of the confederate agents' behavior and risk-taking attitudes as a source of behavioral variation in future studies using the multi-arm bandit game.

Third, though statistically significant trends were observed in the data, all of the effect size measures obtained in this study were small. This small effect size calls for more data to be collected to determine the robustness and stability of these effects in this experiment and explore possible ways that the experimental design could be made more stable. For example, an end of study or mid experiment attention check could be delivered to participants to ensure that participants are aware of the various manipulations in the study. Employing these types of methods would allow more a better way to decrease the noise in the data caused by the subject not attending to the experiment.

Finally, future research could examine the extent that trust discounting could be moderated by additional manipulations. In the current study, the assumption was made that trust was discounted only when participants attempted to interact with the confederate agent but the confederate agent could not interact with them during the experiment. It might be the case that the effects of discounting can be mitigated if participants are given a reason as to why they could not interact with the confederate agent during a particular round. For example, when people go on vacation they often set up an automatic email reply to let others know they are not available to read their email. Removing the ambiguity of interaction might help maintain properly calibrated trust relationships over time in variable environments.

CONCLUSION

The interdependence between individuals and others (humans or machines) is a fundamental component of social interactions. This interdependence allows for a division of expertise between multiple people allowing individuals to accomplish tasks they would not have been able to do themselves. However, for these benefits to arise individuals need to be able to trust and rely on others, which depends on the environment and the characteristics of the trustee.

Theories of trust need to be formalized so that they can be used to understand particular human interactions or predict how certain changes to the environment will modify the trust between individuals. The lack of formalization in other theories of trust (Mayer et al. 1995; Sperber et al. 2010) has led to a great deal of ambiguity and confusion in the literature limiting their generalizability and their ability to make sense of more complex interactions. Formal theories of trust, such as the one presented in this paper, allows for greater precision in explanation and prediction. In this study, we examined the predictions of a formalized trust model for a novel multi-arm trust game. An evaluation of our model found that overall the general patterns in the data were accounted for. As predicted, the interaction schedule of a trustee was found to affect both the participants' behavioral allocation and state trust evaluations, supporting the necessity of the Juvina et al. (2019) trust discounting mechanism. However, there were particular aspects of the behavior in the data the model did not predict, such as a discrepancy between the model's predictions of the animacy and non-animacy condition and the optimization of the participant's allocations. It is always possible post-hoc to develop a model that can account for a set of data after it is collected, though too often in these cases the complexity of these models is unnecessarily increased and this was not the goal of this paper. Instead, we examined the limitations of a simple model to account for a particular dataset, which allows for a more theoretically informative model evaluation (Wikens, 1998). The results of this paper show support for a simple model of trust and future avenues of researc

V. REFERENCES

1. Alarcon, G. M., Lyons, J. B., & Christensen, J. C. (2016). The effect of propensity to trust and familiarity on perceptions of trustworthiness over time. *Personality and Individual Differences*, *94*, 309-315.

2. Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

3. Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological science*, *2*(6), 396-408.

4. Arrow, K., The Limits of Organization, W. W. Norton & Company, New York, 1974.

5. Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, *10*(1), 122-142.

6. Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), 199-231.

7. Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

8. Burns, K. J., & Demaree, H. A. (2009). A chance to learn: On matching probabilities to optimize utilities. Information Sciences, 179(11), 1599-1607.

9. Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481), 933-942.

10. Collins, M., Juvina, I., Douglas, G., & Gluck, A. K. (2015). Predicting trust dynamics and transfer of learning in games of strategic interaction as a function of a player's strategy and level of trustworthiness. In *Proceedings in the 24th Annual Conference on Behavioral Representation in Modeling and Simulation (BRIMS) (Washington, DC)*.

11. Collins, M. G., Juvina, I., & Gluck, K. A. (2016). Cognitive model of trust dynamics predicts human behavior within and between Two Games of Strategic

Interaction with Computerized Confederate Agents. *Frontiers in psychology, section cognitive science*, *7*.

12. Collins, M. G., Juvina, I., & Gluck, K. A. (2016). Game-Specific and player-specific knowledge combine to drive transfer of learning between games of strategic interaction. ” In *Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, 2015, pp. 186–195.

13. Couch, L. L., & Jones, W. H. (1997). Measuring levels of trust. *Journal of research in personality*, *31*(3), 319-336.

14. Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *Journal of applied psychology*, *92*(4), 909.

15. Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. *The adapted mind: Evolutionary psychology and the generation of culture*, 163, 163-228.

16. Csibra, G., Gergely, G., Bıró, S., Koos, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of ‘pure reason’in infancy. Cognition, 72(3), 237-267.

17. de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331.

18. De Melo, C. M., Carnevale, P., & Gratch, J. (2011). The impact of emotion displays in embodied agents on emergence of cooperation with people. *Presence: teleoperators and virtual environments*, *20*(5), 449-465.

19. Deutsch, M. (1958). Trust and suspicion. *Journal of conflict resolution*, *2*(4), 265-279.

20. El Salamouny, E., Krukow, K. T., & Sassone, V. (2009). An analysis of the exponential decay principle in probabilistic trust models. *Theoretical computer science*, *410*(41), 4067-4084.

21. Engle-Warnick, J., & Slonim, R. L. (2004). The evolution of strategies in a repeated trust game. *Journal of Economic Behavior & Organization*, *55*(4), 553-573.

22. Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619-619.

23. Hardin, R. (1993). The street-level epistemology of trust. *Politics & Society*, *21*(4), 505-529.

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

25. Harris, P. L., & Corriveau, K. H. (2011). Young children's selective trust in informants. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *366*(1567), 1179-1187.

26. Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. The American journal of psychology, 57(2), 243-259.

27. Hoffrage, U., & Marewski, J. N. (2015). Unveiling the Lady in Black: Modeling and aiding intuition. *Journal of Applied Research in Memory and Cognition*, *4*(3), 145-163

28. Falvello, V., Vinson, M., Ferrari, C., & Todorov, A. (2015). The robustness of learning about the trustworthiness of other people. *Social Cognition*, *33*(5), 368-386

29. Griffiths, N. (2005, July). Task delegation using experience-based multi-dimensional trust. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, 489-496.

30. Ignat, C. L., Dang, Q. V., & Shalin, V. L. (2019). The influence of trust score on cooperative behavior. *ACM Transactions on Internet Technology (TOIT),* 19(4), 1-22.

31. Juvina, I., Collins, M.G., Larue, O., Kennedy, W., de Visser, E., & de Melo, C. (2019). Toward a unified theory of learned trust in interpersonal and human-

machine interactions. *ACM Transactions in Interactive Intelligent Systems, 9*(4), 1-33.

32. Juvina, I., Saleem, M., Martin, J. M., Gonzalez, C., & Lebiere, C. (2013). Reciprocal trust mediates deep transfer of learning between games of strategic interaction. *Organizational Behavior and Human Decision Processes*, *120*(2), 206-215.

33. Juvina, I., Lebiere, C., & Gonzalez, C. (2015). Modeling trust dynamics in strategic interaction. *Journal of applied research in memory and cognition*. 4(3): 197-211.

34. Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *Academy of management Review*, *23*(3), 438-458.

35. Lewicki, R. J., & Wiethoff, C. (2006). Trust, trust development, and trust repair. *The handbook of conflict resolution: Theory and practice*, 92-119.

36. Lee, J. J., Knox, W. B., Wormwood, J. B., Breazeal, C., & DeSteno, D. (2013). Computationally modeling interpersonal trust. *Frontiers in psychology*, *4*.

37. Lount Jr, R. B., Zhong, C. B., Sivanathan, N., & Murnighan, J. K. (2008). Getting off on the wrong foot: The timing of a breach and the restoration of trust. *Personality and Social Psychology Bulletin*, *34*(12), 1601-1612.

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

39. McLain, D. L., & Hackman, K. (1999). Trust, risk, and decision-making in organizational change. *Public Administration Quarterly*, 152-176.

40. Moisan, F., ten Brincke, R., Murphy, R. O., & Gonzalez, C. (2018). Not all Prisoner's Dilemma games are equal: Incentives, social preferences, and cooperation. *Decision,* 5(4), 306.

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

42. Nairne, J. S. (2016). Adaptive Memory: Fitness-Relevant "Tunings" Help Drive Learning and Remembering. *Evolutionary Perspectives on Child Development and Education*, 251-269.

43. Nairne, J. S., Thompson, S. R., & Pandeirada, J. N. (2007). Adaptive memory: survival processing enhances retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 263.

44. Nairne, J. S., VanArsdall, J. E., Pandeirada, J. N., Cogdill, M., & LeBreton, J. M. (2013). Adaptive Memory The Mnemonic Value of Animacy. *Psychological Science*, *24*(10), 2099-2105.

45. Rapoport, A. (1967). A note on the "index of coop-eration" for Prisoner's Dilemma. *The Journal of Conflict Resolution*, 11, 100–103

46. Rieskamp, J., & Gigerenzer, G. (2005). Heuristics for social interactions: How to generate trust and fairness. *Manuscript submitted for publication*.

47. Robbins (1952). Some aspects of the sequential design of experiments. In: *Bulletin of the American Mathematical Society*, 55, 527–535.

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

49. Samson, D (2013). Theory of Mind." Edited by Daniel Reisberg. Oxford Handbook of Cognitive Psychology, doi:10.1093/oxfordhb/9780195376746.013.0059.

50. Schoorman, F. D., Mayer, R. C., & Davis, J. H. (1996, April). Empowerment in veterinary clinics: The role of trust in delegation. In: *The 11th annual meeting of the Society for Industrial and Organizational Psychology*, San Diego, CA.

51. Sakaki, M., Niki, K., & Mather, M. (2012). Beyond arousal and valence: The importance of the biological versus social relevance of emotional stimuli. *Cognitive, Affective, & Behavioral Neuroscience*, *12*(1), 115-139.

52. Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive psychology*, 53(1), 1-26.

53. Sturgis, P., Read, S., & Allum, N. (2010). Does intelligence foster generalized trust? An empirical test using the UK birth cohort studies. *Intelligence*, *38*(1), 45

54. Wang, Y., & Vassileva, J. (2003, September). Trust and reputation model in peer-to-peer networks. In *Peer-to-Peer Computing, 2003.(P2P 2003). Proceedings. Third International Conference on* (pp. 150-157). IEEE.

55. Weber, E. U., & Milliman, R. A. (1997). Perceived risk attitudes: relating risk perception to risky choice. *Management science*, 43(2), 123-144.

56. Wickens, T. D. (1998). On the form of the retention function: Comment on Rubin and Wenzel (1996): A quantitative description of retention. *Psychological Review*, 105, 379–386.

57. Wang and J. Vassileva, Trust and reputation model in peer-to-peer networks, In: Conf. on Peer-to-Peer Computing, Zurich, Aug. 2003, pp.150 – 157.

58. Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100-1122.

59. Yamagishi, T. 2001: Trust as a form of social intelligence. In Cook, K. (ed) Trust in Society. New York: Russell Sage Foundation.

60. Yamagishi, T., Kanazawa, S., Mashima, R., & Terai, S. (2005). Separating trust from cooperation in a dynamic relationship prisoner’s dilemma with variable dependence. *Rationality and society*, *17*(3), 275-308.

VI. APPENDIX A

TRAIT TRUST QUESTIONNAIRE

[*]Item was reverse coded
[a]Items came from Rotter (1967)
[b]Items came from Yamagishi (1986)

1. I generally have faith in humanity. [a]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree)
(4: Agree slightly) (5: Agree very much)

2. I feel that people are generally reliable.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

3. I generally trust other people unless they give me a reason not to.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

4. Most people are basically honest. [b]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

5. Most people are trustworthy.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

6. Most people are basically good and kind.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

7. Most people are trustful of others.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

8. I am trustful.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

9. Most people will respond in kind when they are trusted by others.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

10. Hypocrisy is on the increase in our society.[a]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

11. One is better off being cautious when dealing with strangers until they have provided evidence that they are trustworthy.[a]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

12. Those devoted to unselfish causes are often exploited by others.[a*]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

13. Fear and social disgrace or punishment rather than conscience prevents most people from breaking the law.[a]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

14. Most experts can be relied upon to tell the truth about the limits of their knowledge.[a]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

15. Most people tell a lie when they can benefit by doing so. [b]*

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

16. The judiciary is a place where we can all get unbiased treatment.[a]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

17. Most people answer public opinion polls honestly.[a]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

18. Most repairmen will not overcharge, even if they think you are ignorant of their specialty.[a]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

19. Most people are primarily interested in their own welfare.[b]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

20. Most students in school would not cheat even if they were sure they could get away with it.[a]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

21. Most people can be counted on to do what they say they will do.[a]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

22. Most salesmen are honest in describing their products.[a]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

23. Most elected officials are really sincere in their campaign promises.[a]

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

24. In these competitive times one has to be alert or someone is likely to take advantage of you. [a]*

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

VII. APPENDIX B

* Item was reverse coded

State Trust Questionnaire

1. I feel safe to take risks in this game knowing that the other player would not take advantage of me.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

2. The other player would not willingly undermine my earnings in this game.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

3. The other player behaves consistently.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

4. I believe that the other player wants to help me to make a good amount of payoff in this game.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

5. The other player can be trusted.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

6. The other player is trying to take advantage of me. *

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

7. I feel that the other player is competent.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

8. The other player tries to make me lose in this game. *

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

9. I believe that the other player is fair.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

10. I would not let the other player have any influence over my payoff. *

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

11. I would be willing to let the other player have complete control over the outcomes of this game.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

12. I understand the reasoning behind the other players moves.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

13. I know in advance what moves the other player will make.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

14. I like playing with the other player in this game.

(1: Disagree very much) (2: Disagree slightly) (3: Neither agree nor disagree) (4: Agree slightly) (5: Agree very much)

VIII. APPENDIX C

PILOT STUDY

PILOT STUDY: METHOD

PILOT STUDY PARTICIPANTS

All participants (N = 40; Age = 22, SD = 2; Male: 46% Female: 56%)were recruited from Wright State University for this experiment through on campus advertisement. Participants were paid $10 dollars for their participation and then received up to an additional $10 dollars based on their performance during the experiment. In total, participants had the opportunity to earn up to a total of $20 during the experiment.

DESIGN

The experimental design for the pilot study was a 2 (animacy, non-animacy) x 4 (high, med-high, med-low, and low interaction schedule) x 2 (high and neutral trustworthiness strategy) between and within subject experimental design. During the pilot study participants played 120 iterated rounds of the multi-arm trust game, in one of two different conditions (i.e., animacy vs. non-animacy). The multi-arm trust game is a game of strategic interaction that is played repeatedly with a group of five players. All participants were told that they will interact with either 4 other participants (animacy condition) or 4 computers (non-animacy condition) over the course of the experiment. In reality all participants interacted with the same 4 confederate agents over the course of the experiment.

PILOT STUDY EXPERIMENTAL TASK

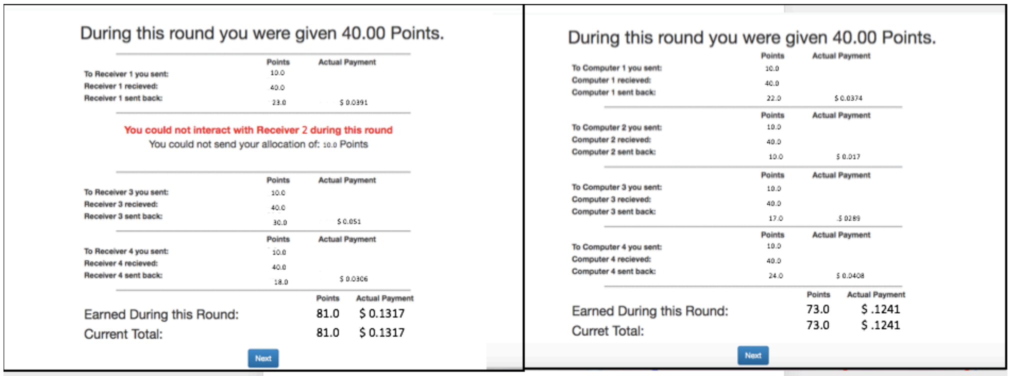
The experimental task used in the pilot study was the multi-arm trust game (MATG). The MATG is a game of strategic interaction combining aspects of two different games, the multi-arm bandit game (Robbins, 1952) and the trust game (Berg, 1995). The MATG is played between 5 players who interact repeatedly. One of the five players is randomly assigned the role of the Sender. While the other four players are assigned the role of the Receiver. Over the course of the MATG each player makes different decisions depending on their role in the game. At the start of a round both the Sender and Receiver each make an initial decision. First, the Sender is given a per round endowment of 40 points. The Sender is then given the opportunity to freely allocate their 40 point endowment between them self and the Receivers. The Sender can give as much or as little of the 40 points as they wish to either them self to any of the 4 Receivers. As the Sender allocates their per round endowment, each Receiver must decide to interact or not to interact with the Sender. If a Receiver decides not to interact with the Sender, then the Receiver earns a random number of points selected from a distribution that is unknown to the Receiver. If a Receiver decides to interact with the Sender, then the Receiver is given the number of points allocated to them by the Sender multiplied by 4. For example, if a Receiver decides to interact with a Sender and the Sender allocated a Receiver 4 points, then the Receiver would be given 16 points. Additionally, Receivers who choose to interact with the Sender are given the opportunity to return any number of their received multiplied allocation back to the Sender. After all the Receivers have made their respective choices, the Sender is then notified of the choices made by each of the

Receivers for that round. If the Sender allocated points to a Receiver who choose not to interact with the Sender during that round, then the Sender is notified that they could not send their points to the Receiver during this round and the Sender is given back the points allocated to the Receiver. If a Sender allocated points to a Receiver who choose to interact with the Sender, then the Sender is notified about the number of points allocated to the Receiver, the multiplied number of points that the Receiver was given, and how many points the Receiver returned to the Sender. The Sender is also told the total number of points earned during the a given round. After the Sender observes the information about the Receivers the next round begins and the same procedure is repeated.

PILOT STUDY EXPERIMENTAL MANIPULATIONS

During each experimental condition (i.e., animacy and non-animacy) all participants played the role of the Sender with 4 confederate agents playing the role of the Receivers. Additionally, each experimental condition had its own scenario to be consistent with the identity of the confederate agents. In the animacy condition, participants were told that they are 1 of 5 participants that have been recruited to participate in this experiment. Each participant in the animacy condition was be told that they have been randomly selected to play the role of the Sender in the experiment, while the 4 other "participants" are assigned to play the role of the Receiver. Additionally, during the MATG, when one Receiver chooses not to interact with the confederate agent, participants were told the Receiver chose not to interact with the participants and they could not be given their allocation during that round (Figure 1).

In the non-animacy condition, participants were told that they are interacting with 4 computers. All participants in the non-animacy condition were told that the behavior of each computer is stochastic, but that each has a different preprogrammed bias. The bias of the computer would make it more or less likely to return more or less points when it is allocated points. Unlike the animacy condition, when participants were explicitly notified when a Receiver chose not to interact with them, participants in the non-animacy condition received no explicit feedback. Instead participants were shown that the computer returned the same number of points allocated by the participant (Figure 1).



During this round you were given 40.00 Points.

	Points	Actual Payment
To Receiver 1 you sent:	10.0	
Receiver 1 recieved:	40.0	
Receiver 1 sent back:	23.0	$ 0.0391

You could not interact with Receiver 2 during this round

You could not send your allocation of: 10.0 Points

	Points	Actual Payment
To Receiver 3 you sent:	10.0	
Receiver 3 recieved:	40.0	
Receiver 3 sent back:	30.0	$ 0.051

	Points	Actual Payment
To Receiver 4 you sent:	10.0	
Receiver 4 recieved:	40.0	
Receiver 4 sent back:	18.0	$ 0.0306

	Points	Actual Payment
Earned During this Round:	81.0	$ 0.1317
Current Total:	81.0	$ 0.1317

During this round you were given 40.00 Points.

	Points	Actual Payment
To Computer 1 you sent:	10.0	
Computer 1 recieved:	40.0	
Computer 1 sent back:	22.0	$ 0.0374

	Points	Actual Payment
To Computer 2 you sent:	10.0	
Computer 2 recieved:	40.0	
Computer 2 sent back:	10.0	$ 0.017

	Points	Actual Payment
To Computer 3 you sent:	10.0	
Computer 3 recieved:	40.0	
Computer 3 sent back:	17.0	$ 0289

	Points	Actual Payment
To Computer 4 you sent:	10.0	
Computer 4 recieved:	40.0	
Computer 4 sent back:	24.0	$ 0.0408

	Points	Actual Payment
Earned During this Round:	73.0	$ .1241
Curret Total:	73.0	$ .1241

Figure 19. Shows the feedback that participants in the animacy (left pane) and non-animacy (right pane) condition received over the course the multi-arm trust game.

PILOT STUDY CONFEDERATE AGENT.

During the experiment participants would interact with 4 confederate agents, whose behavior is pre-determined and the same in both the animacy and non-animacy condition. Confederate agents are used in this study in order to control the frequency of when Receivers interact with the Sender and the amount of the multiplied allocation that Receivers return to the Sender. At the start of the experiment each confederate agent is randomly placed on a unique interaction schedule. The first interaction schedule is the high interaction schedule. On the high interaction schedule, the confederate agent would be able to interact with the participant during every round of the MATG. The second interaction schedule is the medium-high (med-high) interaction schedule. On the med-high interaction schedule the confederate agent would be able to interact with confederate agent every 2 rounds. The third interaction schedule is the medium-low (med-low) interaction schedule. On the med-low interaction schedule the confederate agent would able to interact with the participant every 4 rounds. The fourth interaction schedule is the low interaction schedule. On the low interaction schedule, the confederate agent would be able to interact with the confederate agent every 6 rounds. Noise was added to the confederate agents' interaction schedule. Noise is added to the interaction schedule of a confederate agent by adding a 33% percent chance that the confederate agent would interact with the participant one round before or after its set interaction schedule. The addition of noise to the interaction schedule of a confederate agent decreases the likelihood that participants would be able to determine when they can and cannot interact with a particular confederate agent.

The second aspect of the confederate agents' behavior that is manipulated is the percentage of the multiplied allocation that the confederate agent returns to the participant during rounds when the confederate agent can interact with the participant.

Confederate agents decide about how much of the multiplied allocation to return to the participant using one of two pre-determined strategies. Each of the confederate agent's strategy is used at a different point during the experiment. For each strategy a confederate agent decides how much to return to the participant by randomly choosing a number from a particular truncated normal distribution. The number randomly chosen from the truncated normal distribution is then multiplied by the multiplied allocation or the allocated number of points and then returned to the participant. The confederate agents' first strategy (high trustworthiness) is used during the first part of the MATG (rounds 1-70). During the high trustworthiness strategy, a number is randomly chosen from a truncated normal distribution with a mean of .5, standard deviation of .1, a minimum of 0, and a maximum value of 1. The number selected from this predetermined distribution is then multiplied by the full amount of the multiplied allocation given to the confederate agent and returned to the participant. While using the high trustworthiness strategy each confederate agent on average returned half of the multiplied allocation to the participant, during rounds when the confederate agent can interact with the participant. The purpose of the high trustworthiness strategy is to allow the participants to develop varying levels of trust in the four confederate agents based on their unique interaction schedule.

During the second part of the MATG (rounds 71-120) the confederate agents use the second strategy (neutral trustworthiness). While using the neutral trustworthiness strategy, confederate agents select a number from a truncated normal distribution, with a mean of 1, standard deviation of .1, minimum value of 0, and maximum value of 2. The number chosen from this distribution is then multiplied by the participant's allocation and then returned to the participant. Using this neutral trustworthiness strategy, participants on average received the amount they allocated to the confederate agent. The purpose of the second strategy is to observe the participants' reaction to the confederate agents' trustworthiness neutral behavior, being neither untrustworthy (i.e., sending back less the originally allocated amount) nor trustworthy (i.e., sending back more than originally allocated amount). It is under these conditions that I predict trust discounting would occur.

PILOT STUDY TRUST MEASURES

PILOT STUDY SELF-REPORT MEASURES

During the study participants in the animacy and non-animacy condition took a set of survey measures. Participants in the animacy condition took a trait and state trust survey. While participants in the non-animacy condition took only a trait trust survey.

Trait trust measure

To measure a participants' trait trust, a 24-item questionnaire (Appendix A) using items from two different trait trust surveys from Rotter (1967) and Yamagishi (1986). All items are rated on a 1 (strongly disagree) – 5 (strongly agree) graphical interface scale. An example of an item used on the trait trust survey is "Most people are basically honest".

State trust measure

To measure participants' state trust, a 14-item state trust survey (Appendix B) using items from Collins et al. (2015). All items are rated a 1 (strongly disagree) – 5 (strongly agree) graphical interface scale. An example of an item used on the state trust survey is "Receiver 1 can be trusted".

PILOT BEHAVIORAL MEASURES

During the experiment two aspects of the participants' behavior during the MATG was be measured: (1) the frequency that the participants allocate any of their per-round endowment to each of the confederate agents and (2) the proportion of the participant's per-round endowment a participant allocates to a particular confederate agent during each round.

PILOT STUDY PROCEDURE

For this experiment all participants were recruited from Wright State University. Participants first were brought into the laboratory where they signed an informed consent form. Next all participants were brought to an experimental booth, where they sat in front of a computer to complete the experiment. Next, participants received instructions about how to play the MATG. Based on the experimental condition the participants have been assigned to, participants were told that they are interacting with either other participants (animacy condition) or with multiple computers (non-animacy condition). Additionally, participants were told that they were be paid for their performance during the study earning $0.01 for every 6 points earned over the course of the game. After reading the instructions participants in the animacy condition were told they have been randomly selected to play the role of the Sender and go on to play the MABG with the four confederate agents. Next, all participants took the trait trust survey and begin playing the MATG. Participants were not told the total number of rounds in the game to avoid end game effects.

After playing 120 rounds of the MATG, participants in the animacy condition took 4 state trust surveys, one for each of the 4 confederate agents. Once all participants in the animacy condition have completed the final set of surveys, participants were debriefed to the true nature of the experiment. Participants in the non-animacy condition did not take the state trust surveys. Finally, all participants were compensated based on the total number of points they earned over the course of the 120 rounds of the MATG.

PILOT STUDY RESULTS

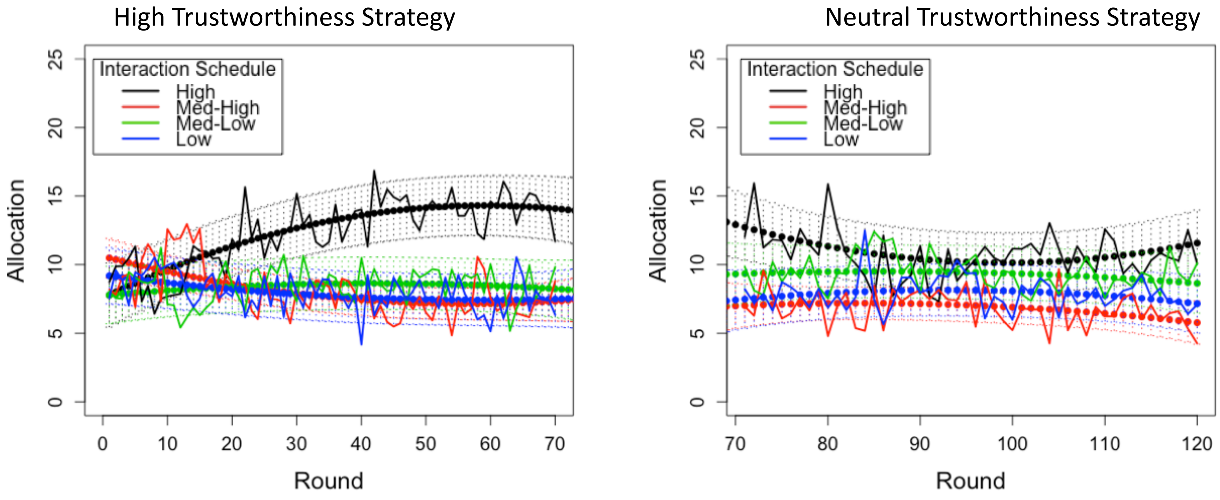
To assess the hypothesis of our pilot study, the participants round by round allocation to the four confederate agents during the multi-arm trust game and the responses to the participant's state and trait questionnaires were examined. To examine the behavior of the participants over the course of the experiment a linear mixed effects model was used. The linear mixed effects model regressed each participants' per round allocation on to a quadratic effect of round, nested within strategy, a confederate agent's interaction schedule (i.e., high, med-high, med-low, and low), strategy (i.e., high and neutral trustworthiness) and identity (i.e., animacy and non-animacy). A linear mixed model was chosen over repeated measure ANOVA, due to the fact the linear mixed model does not aggregate over the behavior of participants (De Visser, 2016).

Before our hypotheses were assessed, three linear mixed effects models each with different mixed effects were compared against each other. Each model regressed the participants per round allocation on to the same four factors (i.e., round, confederate agents' interaction schedule, strategy, and identity) (Appendix D). However, each of the three models differed in their mixed effects. The first model was a fixed effects model with no mixed effects (i.e., fixed effect model). In the second model, each participant was given their own unique intercept (i.e., random intercept model). In the third model, each participant given a unique intercept and each interaction schedule was given a unique slope (random intercept and slope model). To compare the three models the Akaike information criterion (AIC) was used. AIC is a method of model comparison which compares the variance accounted for by a given model relative to its number of free parameters). AIC punishes models that are overly complex favoring simpler models (Burnham & Anderson, 2003). Of the three models compared, the random intercept and slope (*AIC* = 135086.4) had a lowest AIC compared to both the random intercept (*AIC* = 140154.5) and fixed effect model (*AIC* = 140336.2). From the comparison of AIC values, I concluded that despite the additional complexity of the random slope and intercept models, the lower AIC values suggest the additionally complexity of the model allows for a more meaningful explanation of variance in the data. From these results, the random slope and intercept model was used to test the hypotheses for the pilot study (Appendix D).

Rate of Allocation: The linear mixed effect model found a significant quadratic effect of round over the course of the MATG across both experimental conditions (*F*(1,18635) = 13.43, *p* < .01). An investigation into round reveled that there was a slight decrease in the participants' allocations to the four confederate agents over the course the MATG (B = -.00001, SE = -.0001) during both the animacy and non-animacy condition. However, no significant interaction was observed between round and strategy (*p* > .05). Although, a significant interaction between round, strategy, and the confederate agents' interaction schedule was observed (*F*(3,18635) = 8.66, *p* < .01) (Figure 2). The interaction between round, the confederate agents' strategy and interaction schedule suggests participants' allocation to the confederate agents was dependent on the confederate agents' using a particular strategy (i.e., high or low trustworthiness) and interaction schedule (i..e, high, med-high, med-low, low) during both the animacy and non-animacy condition. A Tukey's post hoc test run on the model revealed differences in allocation between the confederate agent strategies and interaction schedules. While the confederate agents used the high trustworthiness strategy it was found there was a positive trend in allocation to the confederate agent on the high interaction condition (*B* = .002, *SE* = .002), which was found to be significantly different than the med-high (*B* = -.001, *SE* = .002), med-low (*B* = -.003, *SE* = .002), and low (*B* = -.006, *SE* = .002) (*t*(16635) = 9.51, *p* < .01; *t*(16635) = 11.38, *p* < .01; *t*(16635) = 6.54, *p* < .01).

Additionally, a significant difference was observed in the trend of the participants' allocation to the confederate agent on the med-high and med-low interaction schedule (t(16635) = -4.84, p < .01). While the confederate agents used the trustworthy neutral strategy, there was a significant change in the confederate agents rate of allocation to the confederate agent on the high (B = -0.01, SE = .002) compared to the low (B = .003, SE = .002) interaction schedule (t(18635) = -3.179, p < .03). No other significant contrasts were observed.

Based on these results examining the rate of the participants' allocation to the confederate agents across both the animacy and non-animacy condition during the MATG, I fail to find evidence for H1 and H2. No significant interaction between the round and the confederate agents' strategy was observed in the participants' allocation. However, weak support was observed for H1a, and H2a. I hypothesized that the rate of the participants' allocation would have a positive relationship with the confederate agents' interaction schedule, while the confederate agents used the high trustworthiness strategy. However, the only significant differences that were observed between the rates of allocation of the confederate agents based on their interaction schedule was between the high interaction and other interaction schedules (med-high, med-low, and low) while the confederate agents used the high trustworthiness strategy. Additionally, while the confederate agents used the neutral trustworthiness strategy there was only a significant difference in the rate of allocation between the low and high interaction schedule. In each of these cases the hypothesized negative relationship between the participants' rate of allocation to the confederate agents on the interaction schedule was not observed. The observed difference that did occur was in the predicted direction. Based on these results, evidence supporting but not confirming H1a and H2a was found.


High Trustworthiness Strategy

Neutral Trustworthiness Strategy

Figure 20 The average per round allocation (solid line) and least mean square estimate and 95% confidence interval (dashed line) to the confederate agent on the high (black), med-high (red), med-low (green), and low (blue) interaction schedule while the confederate agents used the high (left panel) and neutral (right panel) trustworthiness strategy.

Confederate Agent Identity: No main effect of the confederate agents' identity was found over the course of the MATG (p > .05). However, a significant interaction was observed between the confederate agents' identity and interaction schedule (*F*(3,18635) = 8.904, *p* < .001) (Figure 3).

A Tukey's post-hoc test revealed the participants' allocations to the confederate agents' based on their interaction schedule depended on the identity of the confederate agent. During the animacy condition, participants were found to allocate a greater portion of their per round endowment to the confederate agent on the high (*B* = 18.75, *SE* = 1.73) compared to the med-high (*B* = 8.88, *SE* = 1.15), med-low (*B* = 5.22, *SE* = 1.57), and low (*B* = 5.09, *SE* = 1.54) interaction schedule (*t*(18635) = 4.33, *p* < .01; *t*(18635) = 5.22, *p* < .01 ;*t*(18635) = 5.342, *p* < .01). No other significant effects were observed across the confederate agents across the other interaction schedules during the animacy condition (*p* > .05). During the non-animacy condition no significant differences were observed in the allocation across the confederate agents' on the different interaction schedules (*p* > .05). Across the animacy and non-animacy condition the participants' allocations between the confederate agent on the high interaction schedule was found to be significantly different(*t*(37) = -3.364, *p* < .03) . Participants in the animacy condition allocated a greater portion of their endowment to the confederate agent on the high interaction (*B* = 18.75 *SE* = 1.74) schedule compared to the non-animacy condition (*B* = 10.59, *SE* = 1.69).

From these results, I fail to find evidence to support H3, but the significant effects that were observed between allocations between the animacy and non-animacy allocation to the confederate agent on the high interaction schedule but do show partial support for the H3. Although, no main effect of the confederate agent's identity was found, the confederate agents' identity was found to interact with the participants' allocation across the confederate agents. Participants in the animacy condition were found to differentiate their allocation among the confederate agents compared to the participants in the non-animacy condition.

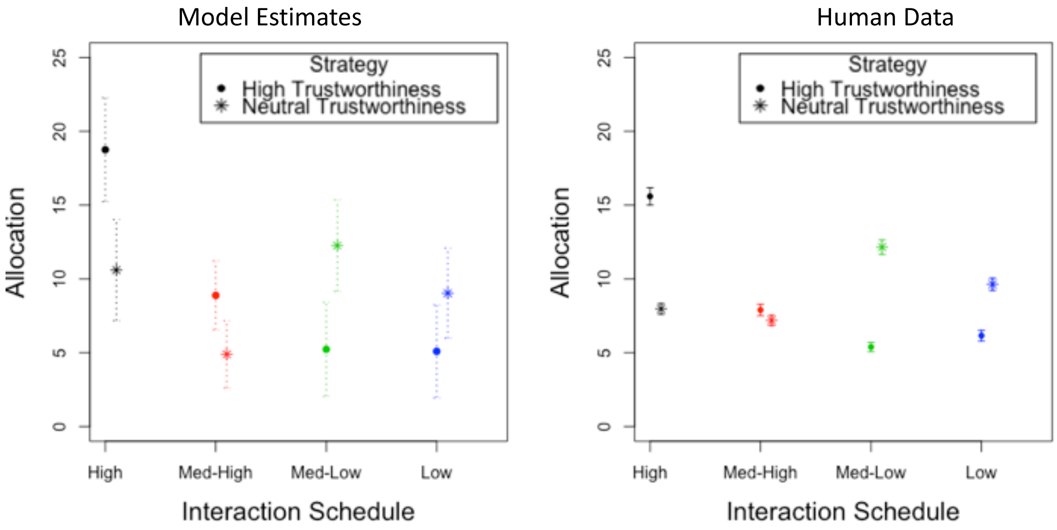


Figure 21. The average allocation +/- 95% CI from the linear mixed effect model (left panel) and human data (right panel) to the confederate agent on the high (black), med-high (red), med-low (green), and low (blue) interaction schedule while the confederate agents used the high trustworthiness strategy (solid dot) and neutral trustworthiness condition (star).

Interaction Schedule: The linear mixed effect model found a significant main effect of the confederate agents' interaction schedule on the participants' allocation to the confederate agents across both the animacy and non-animacy conditions (*F*(3,18635) = 3.277, *p* < .03) (Figure 4).

A Tukey post-hoc test revealed significant differences in the participants' allocation across the confederate agents based on their interaction schedule. Participants allocated the greatest portion of their endowment to the confederate agent on the high (*B* = 14.67, *SE* = 1.21) interaction schedule compared to the med-high (*B* = 6.88, *SE* = .80), med-low (*B* = 8.74, *SE* = 1.09), and low (*B* = 7.06, *SE* = 1.08) interaction schedule (*t*(18635) = 4.167, *p* < .01; *t*(18635) = 4.897, *p* < .01; *t*(18635) = 3.356, *p* < .01). No other significant differences were observed across the confederate agents' interaction schedule (p > .05).

From these results I fail to find evidence to support H5. I hypothesized that there would be a positive relationship between the confederate agents' interaction schedule and the participants' allocation. Little differentiation in the participants' allocation was observed based on the confederate agents' interaction schedule, with participant's allocation differing only between the high interaction schedules and other interaction schedules (med-high, med-low, and low). Though the observed differences between the participants' allocation are not enough to confirm my hypothesis, the observed effects are in the predicted direction.

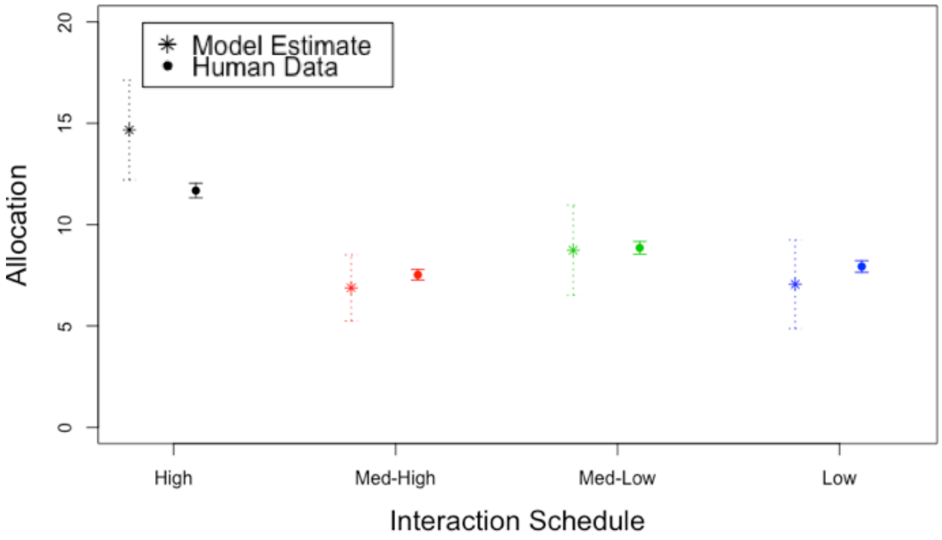


Figure 22. The average allocation and 95% CI from the linear mixed effect model (stars) and human data (solid dots) to the confederate agent on the high (black), med-high (red), med-low (green), and low (blue) interaction schedule.

Strategy The linear mixed effect model revealed no main effect of strategy (p > .05). Based on this results no evidence was found to support H5. The lack of main effect of strategy may be the result of confederate agents' strategy interacting with other aspects of the experimental design (i.e., Confederate Agents Interaction schedule and identity, and round). These multiple interactions masked the main effect of strategy.

PILOT STUDY SURVEY MEASURES

For each participant, the average response of all of the survey measures was taken for both the trait and state trust survey. These trait and state trust measures were then correlated to various measures of behavior over the course of the experiment. I hypothesized that the participants' trait trust measures would correlate with the participants' first round allocation across the 4 four confederate agents (H6). However, no significant correlation was found (p > .05).

Additionally, I hypothesized the participants' state trust in each of the confederate agents would be affected by the participants' allocation (H7) and the confederate agents interaction schedule (H8). To assess these hypotheses two additional linear mixed effects models were run with the participants' average allocation and confederate agents interaction' schedule predicting the participants' state trust in the confederate agent. Both linear mixed effects models found that both allocation (*F*(54) = 6.45, *p* < .01) (Figure 5) and interaction schedule (*F*(54) = 36.73, *p* < .01) (Figure 6) significantly predicted the participants' state trust, in the animacy condition. There was a significant positive relationship between the participants' average allocation and state trust measure (*B* = .05, *SE* = .008). Additionally, a post hoc test revealed that participants had significantly higher trust in the confederate agent on the high (*B* = 2.65, *SE* = .13) interaction schedule compared to the med-high (*B* = 3.15, *SE* = .13), med-low (*B* = 3.15, *SE* = .13), and low (*B* = .287, *SE* = .13) interaction condition (*t*(54) = 4.304, *p* < .03; *t*(54) = 2.80, *p* < .04; *t*(54) = 2.80, *p* < .04). Based on these two results H7 is confirmed and partial support for H8 is found. These results show that both the participants' allocation behavior and the behavior of the trustee (i.e., interaction schedule) both influenced the participants state trust in each of the confederate agents.

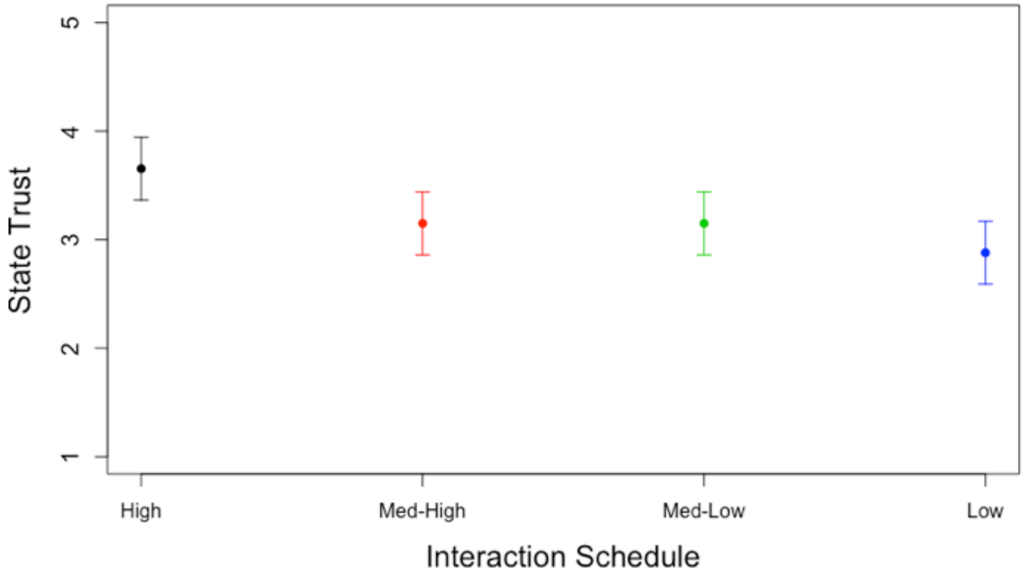


Figure 23. The least square estimate of the participants' state trust in the confederate agents with high (black), med-high (red), med-low (green), and low (blue) interaction schedule.

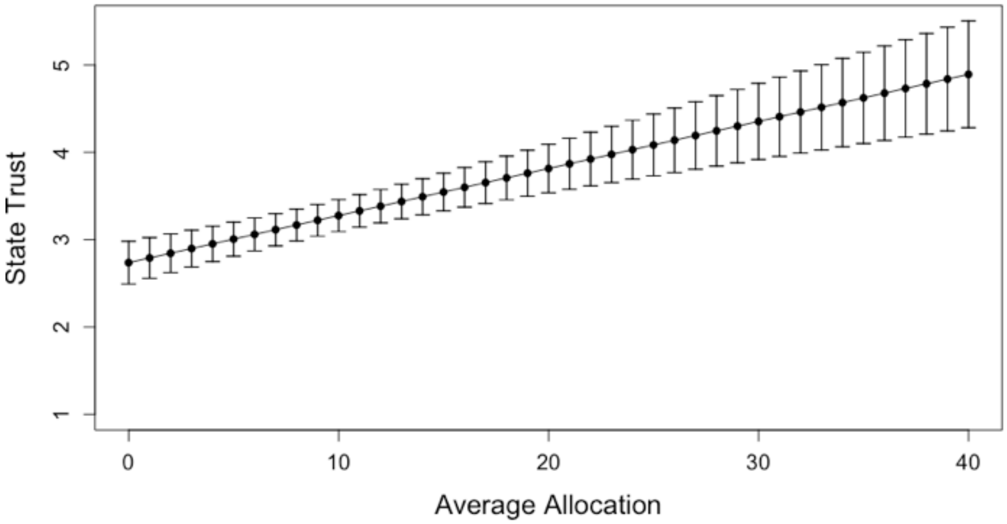


Figure 24. The least square estimates of the participants' state trust in each of the confederate agents as a function of their average allocation over the course of the multi-arm trust game.

IX. APPENDIX D

JUVINA ET AL. (2019) MULTI-ARM TRUST GAME IMPLEMENTATION

TRUST DEVELOPMENT AND ALLOCATION BEHAVIOR

Once the trust model receives the points sent back from each confederate agent, the trust model updates its trust in each of the four confederate agents. The model's trust ($Trust_{i\ t}$) in each confederate agent (i) at a particular time (t) is determined by Juvina et al.'s (2019) trust equation (Equation 1). The trust models' previous (t-1) trust assessment in a confederate agent ($Trust_{i\ t-1}$) is raised to the discounting parameter (a) and then added to that trust model's perceived evidenced of trustworthiness in a particular confederate agent (PET_i). PET_i is determined by the number of points returned by a confederate agent ($Trustee_{\mathrm{i\ t}}$) during a particular trial (Equation 2). The process of updating the trust model's trust in each of the confederate agents is conducted after each time the trust model allocates points to a confederate agent.

After the trust model updates its trust in each of the 4 confederate agents, it then adjusts its willingness to allocate points to each of the four confederate agents during the following round. The trust model adjusts its willingness to allocate points to each confederate agent (Equation 3) based on three factors. The first factor is the number of points the trust model was willing to allocate to a particular confederate agent during the previous round ($Allotment_{i\ t-1}$). $Allotment_{i\ t-1}$ is initially set to the number of points that the trust model allocated to confederate agent during the first round of the simulation. The second factor is the trust model's PET_i of a particular confederate agent during the current round ($PET_{i\,t}$). The third factor is the percent of change in the model's trust for a particular confederate agent. The percent of change in the model's trust is calculated by dividing the trust model's initial trust (T_0)[9] by the model's current trust in a particular confederate agent ($Trust_{i\,t-1}$), subtracted from 1. The trust model then updates its willingness to allocate points to a particular confederate agent by multiplying the confederate agent's $PET_{i\,t}$ by the percent of change in the model's trust for a particular confederate agent and then adding this to its previous willingness to allocate points. This is then repeated for each of the confederate agents that the trust model allocated points to during the previous round.

$$(1).Trust_{i\,t} = Trust^{a}_{i\,t-1} + PET_{i\,t}$$

$$(2)PET_{i\,t} = Trustee_{i\,t}$$

$$(3)\quad Allotment_{i\,t} = Allotment_{i\,t-1} + \left[PET_{i\,t} * \left(1 - \frac{T_0}{Trust_{i\,t}}\right)\right]$$

Finally, after the trust model updates its trust and its willingness to allocate its endowment across the confederate agents, the trust model then chooses how many points it will allocate to each of the confederate agents. The decision of how much the trust model allocates to each confederate agent is determined by the trust model's trust in each confederate agent ($Trust_{i\,t}$) and its willingness to allocate points to a confederate agent

[9] T_0 is initially set to 29, which is an initial trust value determined from previous simulations (Juvina et al., 2019).

($Allotment_t^i$). Before allocating any points to the confederate agents, noise is added to the model's trust assessment of each arm (ε). Noise is added to model's trust assessments to place variability into the order that the trust model's allocates its endowment to the confederate agents. The noise added to the models trust assessment is determined by randomly drawing a single number from a logarithmic distribution with a mean of 0 and variance is determined by the noise parameter (s) (Equation 4). After noise is added to the model's trust assessment each confederate agent is ranked ordinally by their trust assessments. Next the trust model starts to allocate its points from its per round endowment to each confederate agents.

The trust models' current state of trust (i.e., trust or distrust) determines the strategy of the trust model's allocation. If the trust model's trust in a given confederate agent is less than its initial trust value (T_0), then the trust model chooses a number randomly selected from a truncated normal distribution (N(0, .5)), with a minimum value or 0 and maximum value of 1. The number selected from this distribution is then allocated to the confederate agent. If the trust model's trust in a particular confederate agent is greater than its initial trust value (T_0), then a number is randomly drawn from a truncated normal distribution between 0 and 1, with a mean equal to percent of the model's willingness to allocate points to a particular confederate agent (N($\frac{Allotment_t^i}{Endowment}$, .5). The number randomly selected from this distribution is then multiplied by the remaining per round endowment left to distribute among the confederate agents. Due to the fact that on average the trust model allocates its endowment in the order that it trusts each of the confederate agents, the trust model on average will allocate a greater amount of its endowment to confederate agents that it trusts the most. After the trust model makes a decision about how many points to allocate to each confederate agent and the receivers, the trust model updates its trust and willingness to allocate points for each confederate agent. This process is repeated for the remainder of a simulation.

$$(4) \qquad \sigma^2 = \frac{\pi^2}{3} s^2$$

X. APPENDIX E

PILOT STUDY MODEL FIT

The trust model was fit to the average per round allocation to each of the four confederate agents in the animacy and non-animacy condition. Model fitting was obtained by manipulating the models' three parameters (Noise, Trust discounting, and initial 1st round allocation). Both the noise and initial first round behavior were fit to both the animacy and non-animacy condition. The discounting parameter was fit separately to the animacy and non-animacy conditions. The trust model's discounting parameter was manipulated across both conditions, because we hypothesized that the identity of the confederate agents would affect the trust development of the participants and in turn affect their allocations during the MATG.

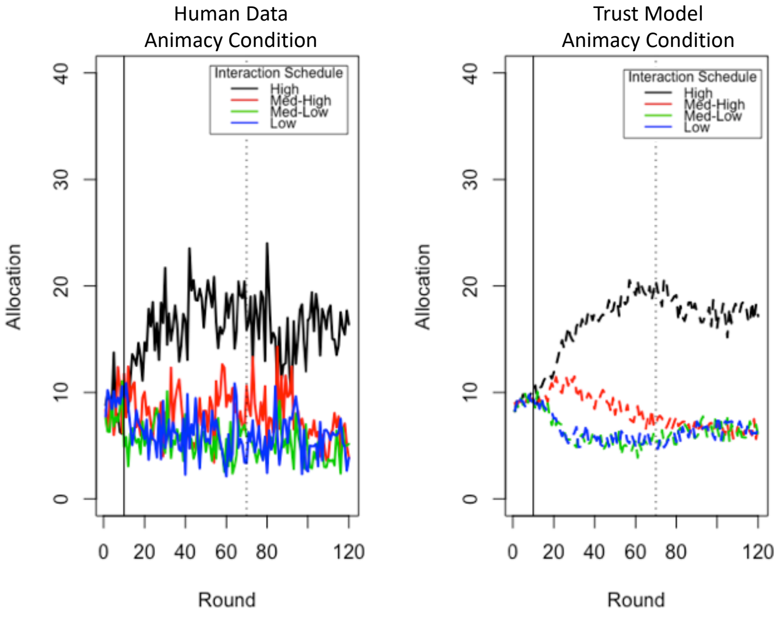


Figure 25. The trust model's fit (right panel) to the average allocation behavior in the animacy condition (left panel) to the confederate agent on the high (black line), med-high (red line), med-low (green line), and low (blue line) interaction schedule.

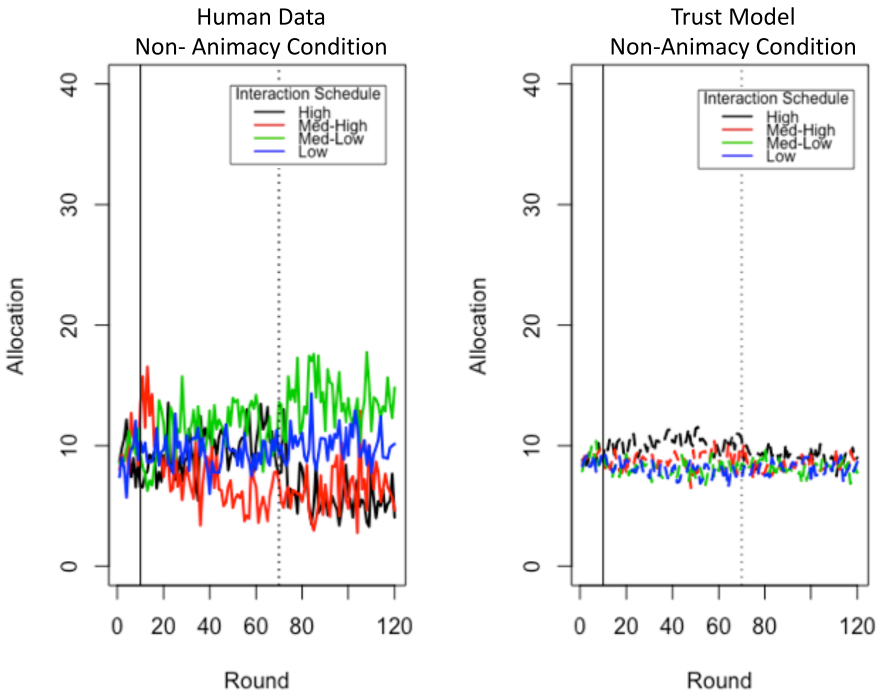


Figure 26. The trust model's fit (right panel) to the average allocation behavior in the non-animacy condition (left panel) to the confederate agent on the high (black line), med-high (red line), med-low (green line), and low (blue line) interaction schedule.

Overall the trust model fit the behavior of the participants in the animacy and non-animacy fairly well (*r* = .70, *RMSD* = 2.9). The trust model was found to better fit the behavior of participants in the animacy condition (*r* = .87, *RMSD* = 2.43) (Figure 7) compared to the non-animacy condition (*r* =.14, *RMSD* = 3.30) (Figure 8). The low correlation between the trust models fit of in the non-animacy condition and the model fit is due the nature of the participants' behavior during the non-animacy condition. In the non-animacy condition there was less differentiation in the participants' allocation between the different confederate agents based on their interaction schedule. The regular steady allocation makes it difficult for the trust model to fit the participants' average behavior during the non-animacy condition. However, it is seen that the trust model's fit to the non-animacy condition does account for some of the qualitative patterns seen the participants' behavior. Additionally, only the discounting parameter was allowed to vary across animacy and non-animacy condition. A better fit might have been able to be accomplished if we allowed more parameters to vary, but this would undercut our goal of model parsimony. The goal of fitting the trust model to the pilot data was to observe how well the trust model could account for behavior in both the animacy and non-animacy conditions.

An examination of the trust model's fit to the participants' behavior in the animacy condition shows that overall the trust model captures the behavior of the participants (Figure 9). The trust model allocates a majority of its per round endowment to the confederate agent on the high interaction schedule while decreasing its allocation to the confederate agents on the other interaction schedules, while the confederate agents use the high trustworthiness strategy. After the confederate agents change their strategy from high to neutral trustworthiness, the trust model also captures the participants' slight change in behavior. After the confederate agents' strategy change, the trust model slightly decreases its allocation to the confederate agent on the high and med-high interaction schedule while retaining the same allocation rate to the confederate agents on the med-low and low interaction condition. Little change is observed in the trust model's behavior after the confederate agents change in strategy. First, the models trust in the confederate agents on the med-low and low are already below T0, meaning that allocation behavior of the models is governed by noise (Figure 9 c, d). Second, the model's trust in the confederate agent on the high and med-high interaction schedule decreases once the confederate agents change their strategy, but reaches a new plateau at a new level above T0 (Figure 10). Due to the fact that the model's trust assessment plateaus to a new point, the model never loses trust in the confederate agents on the high and med-high interaction schedule.

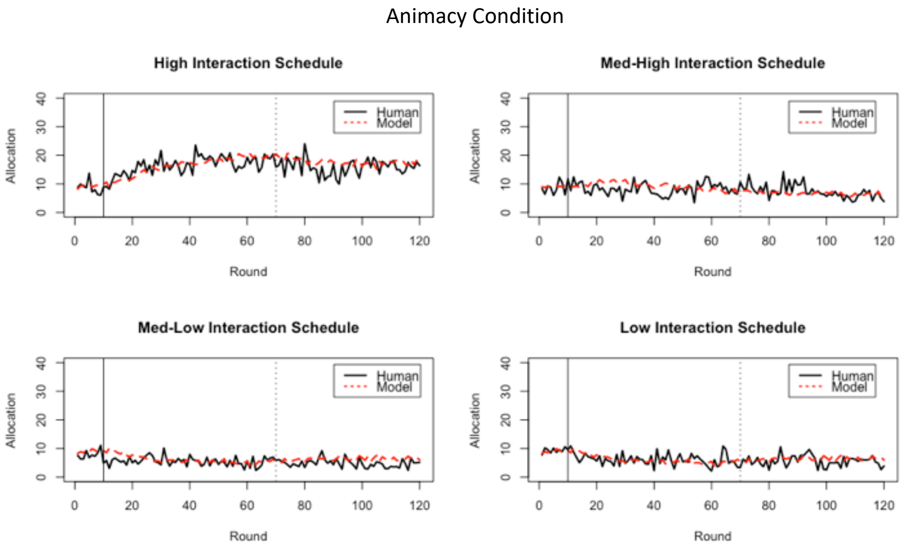


Figure 27. The trust model's fit (red line) to the average allocation behavior in the animacy condition (black line) to the confederate agent on the high (top left panel), med-high (top right panel), med-low (bottom left panel), and low (bottom right panel) interaction schedule.

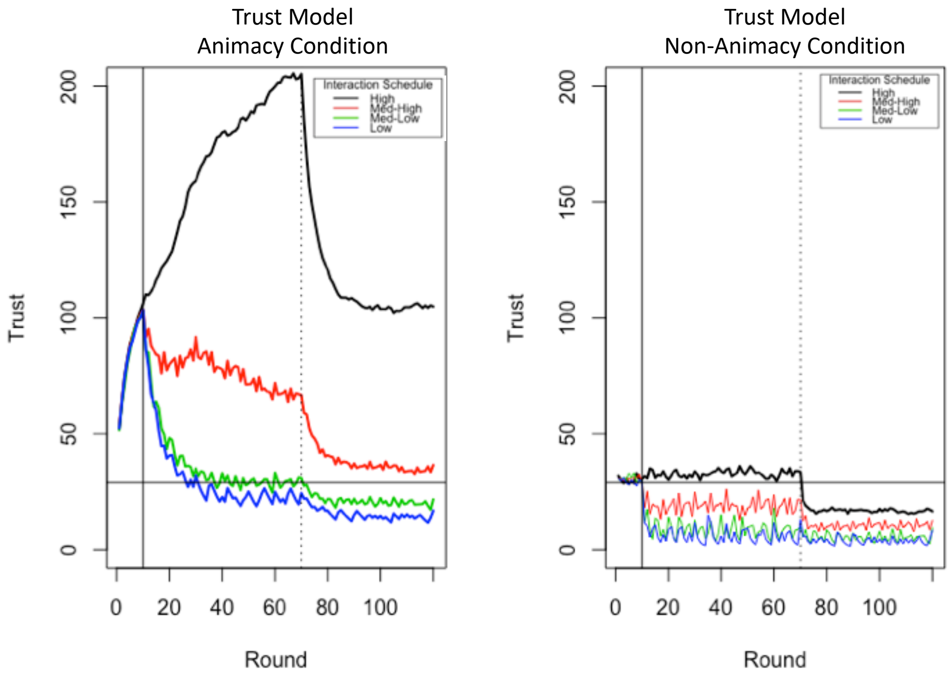


Figure 28. The model's trust assessment for the confederate agents on the high (black line), med-high (red line), med-low (green line), and low (blue line) interaction schedules.

During the non-animacy condition, it is again seen that the model is able to account for the participants' general pattern of behavior (Figure 11). Over the course of the experiment, the participants in the non-animacy condition initially increase their allocation across all four of the confederate agents over the course of the game. The same pattern of behavior is also observed in the trust model's behavior with one exception. Participants in the non-animacy condition allocated a greater amount of their per round endowment to the confederate agent of the med-low interaction schedule. Due to the fact that the trust is discounted during the periods when the confederate agents cannot interact with the trust model, the trust model is not able to allocate more of its endowment to the confederate agent on the med-low interaction condition. Second, the behavior of the trust model is determined by its trust in each of the confederate agents (Figure 10). Due to the fact that the trust model had a lower discounting parameter (a = .7), the trust models' trust in each of the confederate agents plateaus at a lower rate compared to the animacy condition (Figure 10).

Additionally, the trust model's lower discounting parameters leads to trust decreasing more during rounds when the confederate agents cannot interact with the trust model. These two factors lead the trust model in the non-animacy to quickly lose trust in the confederate agents on the med-high, med-low and low interaction schedule after they were placed on their unique interaction schedule. The quick loss of trust by the trust model means that the allocations are governed by noise. However, due to the consistent interaction with the confederate agent on the high interaction schedule, the model is able to maintain trust in the confederate agent on the high interaction condition. Although, because trust in the non-animacy condition plateaus at a lower rate, the trust model in the non-animacy condition does not increase its allocation to the confederate agent on the high interaction schedule as in the animacy condition.

Finally, the confederate agents' change from high to neutral trustworthiness was found only to affect the trust model's behavior towards the confederate agent on the high interaction condition. . This change is due to the fact that the trust change in the models behavior is generated by the confederate agents' behavior in strategy leading the trust model to distrust. As in the animacy condition, the trust model's trust in the confederate agent on the high interaction schedule decreases and then plateaus, but decreases enough for the model to lose trust in the confederate agent. It is the loss in trust that leads the trust model to decrease its allocation to the confederate agent on the high interaction schedule.

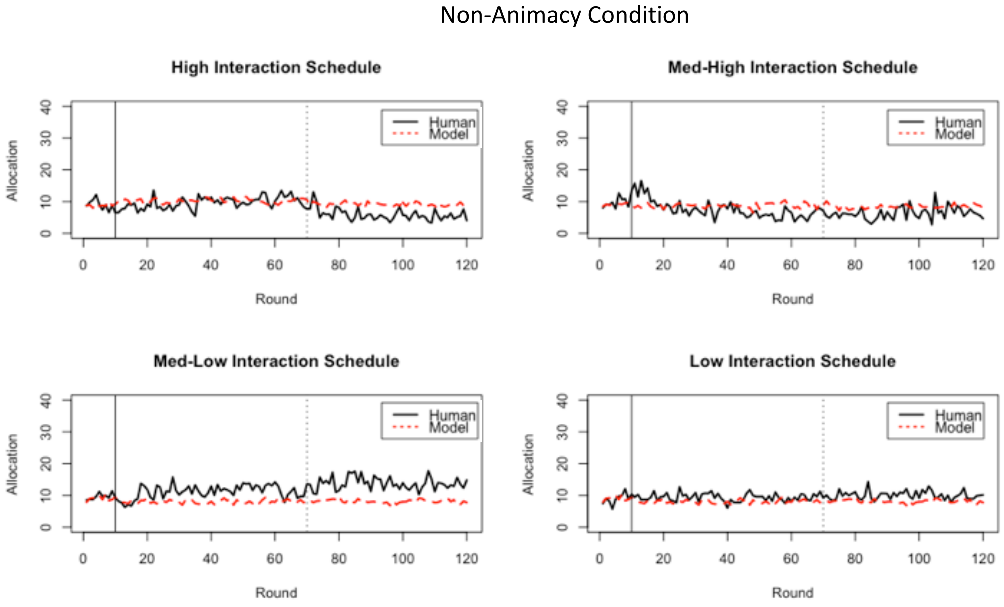


Figure 29. The trust model's fit (red line) to the average allocation behavior in the non-animacy condition (black line) to the confederate agent on the high (top left panel), med-high (top right panel), med-low (bottom left panel), and low (bottom right panel) interaction schedules.

Finally, an alternative model simulation was run to compare the fit of the animacy and non-animacy models (Figure 29). The alternative model was run in the same way as the other two models, but the trust model's discounting parameter (*a*) was removed (Figure 12). Under this construction the trust model does not discount is previous assessment of trust during each interaction with a confederate agent. Without the trust discounting parameter, the trust model's trust accumulates in a linear fashion and does not decrease under situations where the confederate agent cannot interact with the trust model. The alternative model was found to provide an overall worse fit to both the animacy (r = .68, *RMSD* = 3.8) and non-animacy (r = -.59, *RMSD* = 4.34) conditions. These results show the necessity of the trust model's discounting parameter and corroborate the predictions of the trust discounting.

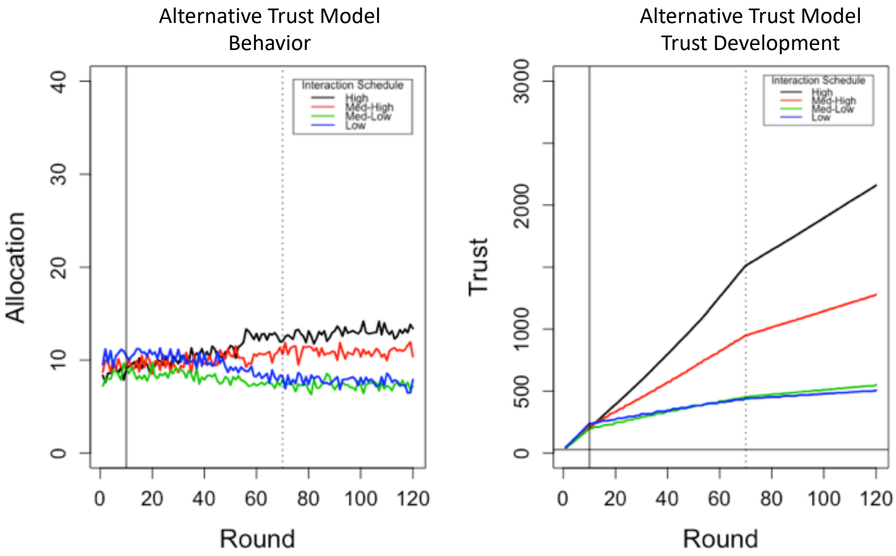


Figure 30. The average round by round allocation behavior of the alternative model (left panel) and the alternative models trust assessments (right panel) of the confederate agents on the high (black line), med-high (red line), med-low (green line), and low (blue line) interaction schedule.

Overall, it is observed that the trust model is able to account for the general trends in the behavior of both the animacy and non-animacy condition. The trust model's best fit to each of animacy condition was due to a difference in the trust models' discounting parameter. I hypothesized that the identity of the confederate agent would lead to participants developing more or less trust in the confederate agents, which in turn would lead to different behavior during the MATG. In line with this view the best fitting discounting parameter for the animacy condition (*a* = .96) was higher in the animacy condition compared to the non-animacy condition (*a* = .7). These differences in the discounting parameter lead the trust model to develop various levels of trust in the confederate agents, leading to behavioral differences in allocation.

Additionally, an alternative model with the trust model's discounting parameter removed was found to not fit either the animacy and the non-animacy condition as well as the trust model. Without the discounting parameter the trust model is less sensitive to the confederate agents' interaction schedule and strategy change. Overall, the model results are in line with predictions of the model and show the predicted necessity of the trust model's discounting parameter

XI. APPENDIX F: LINEAR MIXED EFFECTS MODEL FORMULA

Below is the R code for the three linear mixed effects models discussed in the Results section.

```
library(nlme )
1. Fixed model  = gls(DV ~ (Round/Strategy + I( (Round)^2) ) + Type + Arm +
   Strategy + Arm * Type  * (Round/Strategy + I( (Round)^2) ) * Strategy, data =
   Data)
2. random_intercept model    = lme(DV ~ (Round/Strategy + I( (Round)^2) ) + Type
   + Arm + Strategy + Arm * Type  * (Round/Strategy + I( (Round)^2) ) * Strategy,
   random = ~ 1|ID  , data = Data)
ctrl <- lmeControl(opt='optim');
3. Random intercept slope model   = lme(DV ~ (Round/Strategy + I( (Round)^2) ) +
   Type + Arm + Strategy + Arm * Type  * (Round/Strategy + I( (Round)^2) ) *
   Strategy, random = ~ Arm|ID, control = ctrl, data = Data)
```

To compare the three models.

```
anova( lm1.fixed, lm1.random_int, lm1.random_int_slope
```