Trust Violations in Human-Human and Human-Robot Interactions: The Influence of Ability, Benevolence, and Integrity Violations

Gene M. Alarcon
Air Force Research Laboratory
Wright Patterson AFB, OH
gene.alarcon.1@us.af.mil

August Capiola
Air Force Research Laboratory
Wright Patterson AFB, OH
august.capiola.1@us.af.mil

Justin Morgan
Wright State University
Dayton, OH
morgan.276@wright.edu

Izz aldin Hamdan
General Dynamics Information Technology
Dayton, OH
izzy.hamdan@gdit.com

Michael A. Lee
General Dynamics Information Technology
Dayton, OH
michael.lee@gdit.com

Abstract

The present work investigated the effects of trust violations on perceptions and risk-taking behaviors, and how those effects differ in human-human versus human-machine collaborations. Participants were paired with either a human or machine teammate in a derivation of a well-known trust game. Therein, the teammate committed one of three qualitatively different trust violations (i.e., an ability-, benevolence-, or integrity-based violation of trust). The results showed that ability-based trust violations had the largest impact on perceptions of ability; the other trust violations did not have differential impacts on self-reported ability, benevolence, or integrity, or risk-taking behaviors, and none of these effects were qualified by being partnered with a human versus a robot. Additionally, humans engaged in more risk-taking behaviors when paired with a robotic partner compared to a human over time.

1. Introduction

The military and industry view autonomous systems as key enablers for their respective future operations. However, the greater complexity of truly autonomous systems often means less control and predictability of the system's behavior [1], which may challenge human trust of the system. Autonomous systems capable of learning and adapting may have the capability to perform tasks themselves, but the operator must trust in the system to perform the task with little oversight in order for a "reliable" system to achieve maximum benefit. One approach to understanding human acceptance of autonomous systems is understanding what heuristics (i.e., biases) humans use to gauge the trustworthiness of machines (e.g., robots, automated assistants, decision support systems) relative to humans.

We structured the paper as follows. First, we discuss interpersonal trust. Next, we discuss how biases shape human perceptions of automated systems (e.g., decision support systems, robot partners). We discuss two frameworks that compare human-human and human-machine interactions and how they pertain to trust. From these competing models and extant work, we build a rationale for our hypotheses on how these biases lead to differential effects of violations of trustworthiness in human-human versus human-machine teams. We then test these hypotheses in an experiment comprising human-human and human-robot teams playing a trust game and discuss our findings through the lens of trust research.

1.1. Interpersonal Trust

Trust is an important aspect of human behavior and has been labeled a core social motive [2]. Certain situational aspects must be present for interpersonal trust to be relevant: there must be at least two parties (e.g., individuals, organizations, groups), the relationship between those parties must present some risk (i.e., there must be something at stake), and the parties must be dependent on each other to a degree [3].

To understand trust, one must delineate it from its antecedents (i.e., propensity to trust and trustworthiness; [4]). The Mayer et al. model of trust depicts trust beliefs (e.g., trustworthiness, propensity to trust) as predictors of trust intentions (i.e., a willingness to be vulnerable) and trusting actions (i.e., assuming vulnerability and engaging in a risk-taking behavior), with trust intentions mediating the relationship between trust beliefs and trust

URI: https://hdl.handle.net/10125/79412
978-0-9981331-5-7

actions. Although trust intentions and trust actions are closely related, they are theoretically distinct [4, 5].

When partners are making their decision to trust or not, biases may come into play. Humans use social cues to form impressions, often times in an automatic fashion (e.g., [6]). In interpersonal interactions, aspects of others such as race, gender, and ethnicity shape these initial impressions (e.g., [7]). Trust research has revealed several trust-based biases that influence interpersonal interactions (see [8, 9]). These tendencies have developed from strong evolutionary pressures to detect possible allies, enemies, predators, prey, or mates in the environment [10].

1.2. Human-Machine Trust

Trust in human-machine interactions can also be integrated into trusting actions (e.g., reliance on the robot or robotic system, monitoring behavior), trusting beliefs (e.g., perceptions of robot trustworthiness, automation schemas), and trusting intentions (a willingness to be vulnerable to the robot) [11]. In these contexts, trustworthiness is important because it leads to trust, and trust in turn leads to reliance behaviors [12]. Trust calibration is a function of perceived trustworthiness of and the reliance on the trustee or referent system. Properly calibrated trust (i.e., relying on the tool when reliance is warranted) leads to more appropriate human use of automated systems (p. 55).

Research has led to several competing models that compare human-human trust and human-machine trust. We focus on two: the computers as social actors (CASA; [13]) model and the unique-agent hypothesis [14]. The CASA model [13] postulates that people treat automated systems in much the same way as they do other humans, ascribing characteristics (e.g., personality, gender) to non-human systems such as computers [15] and website purchase assistants [13]. Humans at times anthropomorphize non-humans such as computers and trust computers in ways similar to people [13, 16]. The unique-agent hypothesis [14], in contrast, posits schemas, monitoring behaviors, trust judgements, and assessments of trust across human-human and human-machine interactions based on previous research. In part, the model postulates humans are more sensitive to performance-based errors made by automated systems than they are for humans [17, 11].

In the trust in automation literature [12], perceptions of an automated system's trustworthiness have been mapped on to antecedents proposed by Mayer et al. [4]. Specifically, Mayer et al. delineate perceptions of a referent's ability (e.g., are they competent?), benevolence (e.g., do they have my interest in mind?), and integrity (e.g., are their principles acceptable?) which predict trust in human-human interaction. Lee and See ([12], p. 59) leveraged Mayer et al.'s [4] model and explicated perceptions of an automated system's performance (e.g., is this system reliable?), purpose (what is this system's intent?), and process (is this system consistent?) that predict trust in the automated system. However, until recently [18], little research has systematically manipulated aspects of a system's trustworthiness and measured its effect on human trust toward automation, let alone those differences which may (or may not) arise in trust toward a human referent (see also [14]). Others have found that differences between relying on a human versus a robot to perform a task is largely influenced by the task type (i.e., dangerous to humans or not; [19]). A direct comparison while keeping the task environment neutral is necessary to effectively isolate bias involved in human trust of robots. Otherwise, the trust process may operate differently across two different task domains – which would make the study of bias contaminated by the influence of task-specificity.

As noted, researchers have begun to study the effects of trust violations on criterion and how they differ between human-human and human-machine interactions. de Visser et al. [14] instructed participants to guess what number would come next in a pattern of ten digits. Then, an agent (computer, avatar, and human) would make a suggestion with varying reliability. Participants had less trust toward less reliable agents over time, but anthropomorphism attenuated this effect: participants trusted anthropomorphic referents more than non-anthropomorphic referents. In two follow-on experiments, de Visser et al. found that anthropomorphism can have different effects on trust depending on whether it is a subjective or objective criterion, and contextual features can obfuscate the effects of anthropomorphism. Whereas de Visser et al. [14] focused on performance-based trust degradations, Alarcon et al. [18] investigated the effects of non-performance-based trust violations on trust toward a human versus a machine teammate in checkmate [20], a derivation of the Berg et al. [21] trust game. Results showed that when the referent did not return the amount of money expected, trust declined. However, there was no effect of referent type, nor an interaction between condition (trust/distrust) and referent type on criterion of interest. However, Alarcon et al. [18] noted several limitations of their study of which we discuss later.

1.3. Current Study

A recent meta-analysis [22] demonstrated robot performance (e.g., failure rate, reliability) related factors were the strongest predictors of trust, supporting the literature on automation schema [17, 23]. Merritt and colleagues [24] have researched the Perfect Automation

Schema (PAS) and found that it is related to higher trust in automated systems. The premise of the PAS is that humans may hold a general view of automation that they are near-perfect and error-free. However, PAS may make humans less forgiving of automated systems when they perceive performance errors that violate their existing expectations of technology [11, 17]. Given the importance of performance as a significant driver of trust of automated systems, it is expected that perceived ability will have a stronger effect on the perceived trustworthiness of the robot relative to the human. Thus:

Hypothesis 1: Prior to a trust violation, participants will perceive a robotic partner to have higher ability than a human partner.

Hypothesis 2: Prior to a trust violation, participants will wager more money with a robotic partner than with a human partner.

Hypothesis 3: Ability/performance-based trust violations will reduce ability perceptions toward a robotic partner more than a human partner.

Hypothesis 4: Ability/performance-based trust violations will reduce wagers toward a robotic partner more than a human partner.

Humans often ascribed intent to human and non-human agents. Humans do not approach robots "tabula rasa" but rather with default models of the robot's intent and knowledge [25]. When humans anthropomorphize robots, in-group and out-group biases may form. Research has demonstrated when participants were introduced to a robot with a name familiar within their country, they perceived it as warmer, psychologically closer, ascribed more of a mind to the robot, and had more contact than when the robot had a foreign sounding name [25]. Robots that display empathy may be more liked and trusted [26], suggesting the importance of intent from the robot. Robots can be believed to be responsible for mistakes, though not as much as a human in the same situation [27]. Thus, it is plausible that the impact of signaling intent may be weaker when directly comparing robots versus humans.

Unlike humans, robots behave in accordance with their programming. However, machines may someday have both decision authority and decision initiative to act autonomously. Then, a robot may be asked to "decide" what the best action might be for a set of stimuli and given set of rules of engagement. Thus, it is plausible that perceived intent from a robot matters, but we theorize that perceived intent will matter more for humans versus robots as it is difficult to separate intentionality of a robot (or any kind of automated system) from the designer of that system [12]. Thus:

Hypothesis 5: Prior to a trust violation, participants will perceive a human partner to have higher benevolence than a robotic partner.

Hypothesis 6: Benevolence/intent-based trust violations will reduce benevolence perceptions toward a human partner more than a robotic partner.

Hypothesis 7: Benevolence/intent-based trust violations will reduce wagers toward a human partner more than a robotic partner.

Consistency has demonstrated importance in human-automation collaboration as well as in interpersonal relationships. Consistency, (thought of as integrity – alignment to shared values in the interpersonal domain; see [4]) is a rational evaluation of past successes and failures and has been shown to be the most important driver of trust in high-stakes interpersonal situations [28]. Maintaining consistent, predictable behavior is a core antecedent to trust as this predictability allows one to forecast behavior in novel situations. Prior research has shown that consistency is key to trust of automation [29]. Predictability was considered a core antecedent for trust in automation in early studies of the construct [30]. The literature on HMT suggests the importance of shared mental models [16, 31], largely because shared mental models help humans anticipate the actions and needs of machine partners. Still, consistency/integrity of an automated system has been interpreted as a cue for intentionality which ought to be more strongly evoked from a human referent compared to a robot or other automated system [16, 18]. However, due to the asymmetry of perceived capability between humans and technology [17], we hypothesize that technology will likely pay a higher cost for deviations of consistency. Thus:

Hypothesis 8: Prior to a trust violation, participants will perceive a human partner to have higher integrity than a robotic partner.

Hypothesis 9: Integrity/consistency-based trust violations will reduce integrity perceptions toward a robotic partner more than a human partner.

Hypothesis 10: Integrity/consistency-based trust violations will reduce wagers toward a robotic partner more than a human partner.

In summary, we hypothesize humans have biases toward automated systems such as decision support systems and robots [17], and these biases influence how violations of system performance affects human trust [32]. However, few researchers have investigated the differential influence of trust violations on human and automated (i.e., robot) systems while controlling for contextual variability [14, 18] and also interpreting the effects of that violation through the lens of Mayer et al.'s

[4] trust model, which has been applied to human-automation contexts [12]. The present work takes care to address the limitations of past work of Alarcon et al. [18]. Specifically, we use a validated measure [33] to assess trustworthiness toward both human and robot referents, which corresponds directly to the trust manipulations in the present work. Secondly, we leveraged manipulations from [34] which differentiate between ability-, benevolence-, *and* integrity-based violations (the latter two of which Alarcon et al. [18] conflated) of trustworthiness and measure their effects on criterion of interest. In this way, we more fully test the assumptions of the unique-agent hypothesis [14] inspired by work on automation bias [11, 17]. Specifically, the assumption that machines are perceived to be more competent than humans before an error occurs, and the steeper decline in trust toward machines compared to humans after an error is perceived, is based upon biases toward perceptions of a machine *performance*. In their Limitations, Alarcon et al. [18] explicitly state that this was the precise factor they held constant in investigating the effects of *non*-performance-based trust violations on criterion of interest in human-human and human-robot partnerships, which limits the generalizability of their results in supporting the CASA model [13]. As such, we manipulate all three features of trustworthiness between-subjects to more fully test the assumptions of both the CASA and unique-agent hypothesis [14].

2. Method

2.1. Participants

A total of 44 participants completed the study; 57.76% were female, their ages ranged from 18 and 62 years (M = 30.69, SD = 12.21), consisting of 69.5% white/Caucasian, 19.05% black/African American, 9.52% Asian, and 2.38% other ethnicities. The study required participants to be at least 18 years old. Participants were recruited from a local university and online craigslist ads.

2.2. Task

Participants were instructed to complete the Checkmate task [20]. Checkmate is an augmented investor dictator trust game (also referred to as the trust game; [22]) in which a banker (investor) sends money to the runner (dictator). The runner uses this investment to complete a virtual maze running task, in which they navigate through a virtual maze and collect boxes. The number of boxes collected is directly related to the earnings. The runner was able to keep the earnings to themselves or share the earnings with the banker. Participants were either paired with a human (confederate) or a robot (NAO robot) partner. In this study, participants were always the banker and the confederate (human or robot) was always the runner.

Checkmate consists of a practice round and five rounds of the main task. In the practice round, participants were informed that money would not be gained/lost. Conversely, in the main task, money that was gained/lost was real and the final earnings were given to participants at the conclusion of the study. Participants began the task with $50 USD in their account (they were told that their partner began with the same amount as well). The banker loaned the money to the runner in hopes of the runner performing well in the maze running task and sharing the earnings with them. Before the banker sent their endowment, the runner notified the banker of the promised return and the risk level. Three risk levels: low (75–150%), moderate (50–200%), or high (0–300%) were offered to the runner, the possibility of reward and loss was related to the risk level. For instance, if "high risk" was selected and the runner performed poorly, the runner could lose the entire endowment (0%) with nothing to keep or share with the banker. For consistency, the runner's selection for risk level was always moderate (50–200%).

After the banker makes an endowment, the runner can change the risk level without informing the banker. Since the runner was a confederate, the risk level was not changed but the banker was informed of this possibility. The banker could loan anywhere from $1-$13. The runner and banker were given an aerial view of the maze, preceding each trial. Next, the runner had two minutes to collect as many boxes as possible. The banker was able to see the number of boxes the runner collected; however, the banker was not told how much money the runner earned. Once the trial was complete, the runner can send the banker their preferred amount; this can differ from the promised amount. The banker then receives the money from the runner and this process is repeated over the next few rounds until the conclusion of the study. Gains and losses from each of the five rounds carry over until the game is complete. For a full review of the task see Alarcon et al [20].

2.3. Manipulations

2.3.1. Partner. A confederate was used to simulate the role of the runner in all conditions. Depending on the condition participants were assigned to, participants were partnered with either a human (confederate researcher) or a robot (NAO robot). The confederate's actions were all pre-recorded and automated. Participants were presented with the pre-recorded

scenarios that corresponded to their condition through the platform.

2.3.2. Condition. We used the manipulations outlined by Alarcon et al [34]. Specifically, the *ability* manipulation was a degradation in the runner's performance such as running into buildings and going off into areas where there were no boxes. *Integrity* was manipulated by having the runner sending less money than promised back to the banker, with no discernable reasons for the lack of return (i.e., no performance degradations). Lastly, *benevolence* was manipulated by informing participants that the color of the boxes collected by the runner affect how earnings were distributed at the end of the round. Specifically, blue and white boxes generated earnings that could be shared between the runner and the banker, while red boxes only generated earnings for the runner. In the distrust rounds, the runner primarily collected red boxes that only benefitted the runner and could not be shared. Thus, participants viewed one of three manipulations of trust at rounds three and four of the experiment, a between-subjects manipulation. Importantly, each trust-violation led to the same return of money from the runner to the banker post-round. Thus, trustworthiness was manipulated in terms of the *reason* for the return in rounds three and four, but actual returns were equated across ability-, benevolence-, and integrity-based violations.

2.4. Measures

2.4.1. Trustworthiness. Trustworthiness was measured using Mayer and colleagues' [4] 17-item scale which measured ability, benevolence, and integrity. Ability was measured with six items (e.g., "I feel very confident about the runner's skills"). Benevolence was measured with five items (e.g., "The runner really looks out for what is important to me"). Integrity was measured with six items (e.g., "The runner tries hard to be fair in dealings with others"). Participants rated all items on a 5-point scale, ranging from 1 ("Strongly Disagree") to 5 ("Strongly Agree").

2.4.2. Risk-Taking. Trust behaviors were assessed using banker loan selection (i.e., dollar amount sent to the runner), as described in the task section above, with more money lent by the banker (participant) indicating higher trust in the runner (confederate).

2.5. Procedure

This study was advertised online through craigslist and word of mouth in a mid-western city. Participants were told they would work with a partner in a computer-mediated maze-running task. Once participants arrived, they were introduced to their partner (human confederate or NAO robot). In the human condition, the partner was taken to a different room. In the robot condition, the robot was placed near the participants on a desk. Once consent was attained, participants were administered several baseline surveys. Following the surveys, participants were instructed to complete an endowment task where they were given 10 minutes to complete five moderately difficult math problems. Participants were told that if they answered a minimum of three problems correctly, they would earn $50 USD to use in the task. Participants were given $50 USD regardless of their performance; this was done so that they perceived the money earned was by their efforts.

A PowerPoint tutorial was presented providing details on how to perform the maze running task and explaining each player's role. Following the training, participants were told that they were randomly assigned to be the banker. Deception was used since the participant was always the banker and the confederate was always the runner. The maze running task began with a practice round followed by five rounds of the task. Prior to each trial, participants sent a dollar amount to the runner. Following each trial, participants were to rate their partner's ability, benevolence, and integrity. Once the task was complete, workers were debriefed and received their earnings as well as a $30 USD gift card for participation.

3. Results

All analyses were conducted using SPSS version 23. A Greenhouse-Geisser correction was applied when sphericity was not met.

3.1. Trustworthiness

Although trustworthiness perceptions were recorded for each round, the current study only included the last four rounds, which centered on the trust violation. A repeated-measures (RM) multivariate analysis of variance (MANOVA) was conducted to test the effects of our between-subjects manipulations, Trust Violation type (Ability-based, Integrity-based, and Benevolence-based) and Partner (Robot and Human), and our within-subjects factor, Round, on the three self-report perceptions of trustworthiness: ability, integrity, and benevolence (dependent variables). The main effects of Violation Type [Pillai's $V = .5$, $F(6, 70) = 3.88$, $p < .01$, $\eta_p^2 = .25$] and Round [Pillai's $V = .61$, $F(9, 324) = 9.24$, $p < .001$, $\eta_p^2 = .2$] were significant, but Partner [Pillai's $V = .07$, $F(3, 34) = .79$, $p = .51$, $\eta_p^2 = .07$] was not significant. Round x Violation Type [Pillai's $V = .27$, $F(18, 324) = 1.81$, $p < .05$, $\eta_p^2 = .1$] was

the only significant multivariate interaction. While the Violation Type x Partner [Pillai's $V = .3$, $F(6, 70) = 2.07$, $p = .07$, $\eta_p^2 = .15$] interaction was only marginally significant, and the interactions of Round x Partner [Pillai's $V = .04$, $F(9, 324) = .42$, $p = .92$, $\eta_p^2 = .01$] and the three-way interaction between Round x Violation Type x Partner [Pillai's $V = .13$, $F(18, 324) = .83$, $p = .67$, $\eta_p^2 = .04$] were not significant. Next, we explored the repeated measures analysis of variance (RM ANOVA) for each outcome. Table 1 illustrates the results of the RM ANOVAs.

3.1.1. Ability Perceptions. The main effects of Round and Violation Type were statistically significant for ability perceptions. The main effect of Partner was not significant. The main effects were qualified by a Round x Violation Type interaction. All other interactions were not significant. Figure 1 (left) displays the estimated marginal means of ability perceptions over time by manipulation. As illustrated in Figure 1 (left), the Ability manipulation had the strongest effect on ability perceptions, whereas there was no difference between the Benevolence and Integrity manipulations.

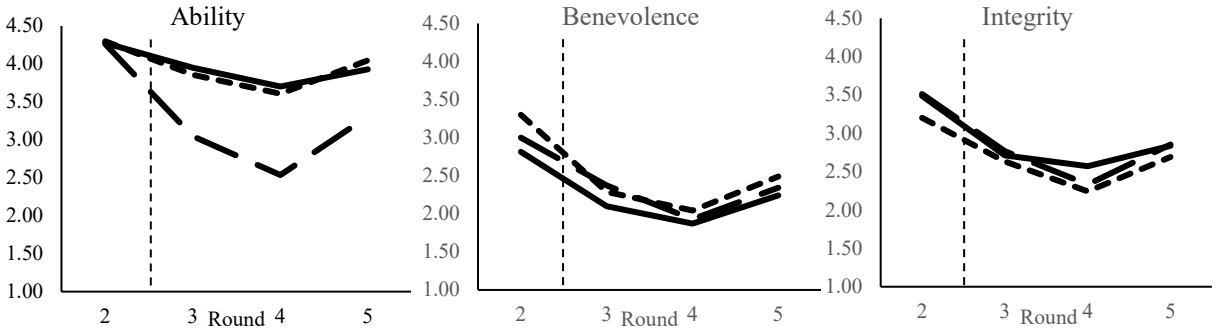
Table 1. RM ANOVA Results for All Outcome Variables

	Ability			Benevolence			Integrity			Risk-Taking Behavior		
	df	*F*	η^2_p	*df*	*F*	η^2_p	*df*	*F*	η^2_p	*df*	*F*	η^2_p
R	1.77, 63.69	25.39*	0.41	2.16, 77.68	29.2**	0.45	2.31, 83.18	39.93**	0.53	2.02, 72.77	21.97**	0.38
VT	2, 36	4.4**	0.20	2, 36	0.49	0.03	2, 36	0.56	0.03	2, 36	0.12	0.01
P	1, 36	0.01	0.00	1, 36	0.12	0.00	1, 36	0.27	0.01	1, 36	1.93	0.05
R*VT	3.54, 63.69	3.77*	0.17	4.32, 77.68	0.43	0.02	4.62, 83.18	0.5	0.03	4.04, 72.77	1.19	0.06
R* P	1.77, 63.69	0.30	0.001	2.16, 77.68	0.39	0.01	2.31, 83.18	0.44	0.01	2.02, 72.77	4.7*	0.12
VT*P	2, 36	1.05	0.06	2, 36	1.17	0.06	2, 36	3.14	0.15	2, 36	0.79	0.04
R*VT*P	3.54, 63.69	0.1	0.01	4.31, 77.68	1.05	0.06	4.62, 83.18	1.16	0.06	4.04, 72.77	0.69	0.04

Note. R = Round, VT= Violation Type, P = partner, *df* = degrees of freedom, η^2_p = partial eta squared, * = $p < .05$, ** = $p < .01$.

3.1.2. Benevolence. The main effect of Round was the only significant main effect on benevolence perceptions. No other main effects nor interactions were significant. Figure 1 (middle) illustrates the estimated marginal means of benevolence across time and manipulations. These scores also declined significantly following Round 3 and continued to remain low following the successive violation in Round 4. The increase after Round 5 is also present for these graphs after the Trust Repair in the final round. Thus, when trust is violated, benevolence perceptions decline, and when it is repaired, it increases but not to a full recovery of pre-violation levels.

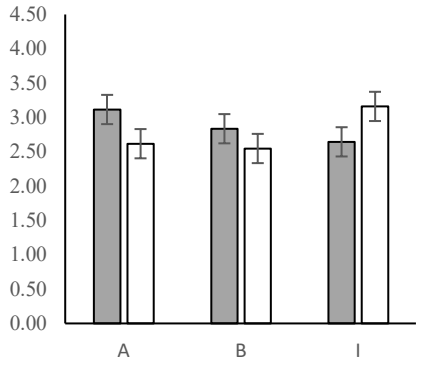
Figure 1. Time by Manipulation for Each Trustworthiness Outcome



Note. Ability, benevolence, and integrity outcomes for ability (long dash line), benevolence (short dash line), and integrity (solid line) manipulations. Vertical dash line depicts when a trust violation occurred.

3.1.3. Integrity. The main effect of Round was the only significant main effect. The Violation x Partner interaction was marginally significant. No other main effects nor interactions were significant. Figure 1 (right) illustrates the Ability, Benevolence, and Integrity manipulations on perceived integrity over time. Integrity perceptions declined significantly following Round three and continued to remain low following the successive violation in Round 4. The increase in Round 5, following the Trust Repair, is also present. Similar to the other conditions, trust did not recover to pre-violation levels. Thus, perceptions of integrity decreased following a trust violation, regardless of whether the violation was an Ability-, Benevolence-, or Integrity-based violation. Although the Manipulation x Partner interaction was only marginally significant, we plotted the interaction in Figure 2. As illustrated in Figure 2, overall integrity perceptions were higher for the human in the condition when Ability was manipulated. In contrast, integrity perceptions for the human were lower than the robot when Integrity was manipulated. There were no differences in the Benevolence condition.

Figure 2. Manipulation by Partner for Integrity Outcome

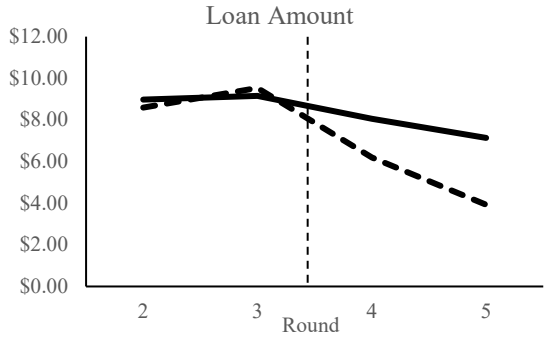


Note. Integrity outcomes for A = ability, B = benevolence, and I = integrity manipulations by partner (human = grey, robot = white).

3.2. Risk-taking Behaviors

Lastly, we conducted a RM ANOVA on risk-taking behaviors. The main effect of Round was the only significant main effect, while the Round x Partner interaction was the only significant interaction. The RM ANOVA results are reported in Table 1. As shown in Figure 3, risk-taking behaviors (or amount of money loaned) declined after each successive round. The Trust Repair behavior occurred in the last Round, after the risk-taking behavior took place. Thus, there was no opportunity for this behavior to recover like the perceptions of the trustworthiness dimensions did. What is unique about these results is that there was a significant difference between the overall risk-taking behavior over time across Partner types. Participants were more likely to risk more money with a Robot than a Human partner, following the trust violation.

Figure 3. Time by Partner for the Loan amount



Note. Loan amounts for human (dash line) and robot (solid line) conditions.

4. Discussion

The purpose of the current study was to examine perceived trustworthiness and risk-taking behavior in the context of human-human and human-robot interactions, and how this relationship is affected by different kinds of trust violations. We manipulated the partner and violation type in a trust game to determine the effect on the perceived trustworthiness of the partner (human or robot) and risk-taking behavior. Results indicate that participant's perceived trustworthiness of their partner decreased over time, as did their risk-taking behavior when trust was violated. In addition, participants' perceptions were most sensitive to ability/performance-based violations, and ability perceptions decreased over time during these violations. However, no differences between violations were found for manipulation type on behavior. In addition, we found some partner effects. There was a marginal interaction of manipulation and partner on integrity perceptions such that if the violation performed was an ability violation, the robot partner incurred a stronger decrease in integrity perceptions. In contrast, if the violation was integrity-based, the human partner incurred a stronger decrease in integrity perceptions. Lastly, participants performed more trusting behaviors with a robot partner than a human over the entire experiment.

Overall, these results support past meta-analytic findings that ability is the strongest driver of human-

human trust [35] and empirical results from Alarcon et al [34] showing that violations of ability perceptions are most impactful on trust criterion compared to benevolence- or integrity-based violations. Moreover, we replicate and extend Alarcon et al. [18] showing that there were minimal differences in how humans perceive violations of trustworthiness (of any kind) from human or robot referents, supporting the CASA model [13].

4.1. Trust Biases and Human Machine Teaming

We hypothesized that trust biases, such as automation bias [17], would influence perceptions and behaviors in the trust game. Specifically, these biases would lead to higher ratings on ability perceptions and lower perceptions of benevolence and integrity [H_1, H_5, H_8] prior to trust violations. We hypothesized greater risk-taking behavior [H_2,] for those paired with a robot partner than a human partner before a trust violation, due to the performance-based task and biases that robots should perform better than humans [17, 32]. Similarly, we expected that ability/performance- and integrity/consistency-based violations of trust would result in more detrimental effects in ability and integrity trustworthiness dimensions [H_3, H_9], and risk-taking behavior [H_4, H_{10}] for those paired with robot partners than human partners, but the opposite would be true for the effects of benevolence/intent-based trust violations [H_6, H_7]. Overall, we only found two interactions for partner effects which were supported statistically: the effect of partner on trust behaviors over time and the marginal interaction of manipulation type and partner on integrity perceptions.

We found that ability-based violations had the strongest impact on ability perceptions, providing evidence that this manipulation was perhaps strongest and more clearly defined from the benevolence and integrity manipulations. Benevolence and integrity perceptions were not differentially influenced by the quality of the trust violation. As Alarcon et al. [18] noted, it is difficult to tease apart benevolence- and integrity-based trust violations. However, ability-based violations may be most relevant on perceptions of ability, particularly in a task where runner competence may be the most meaningful aspect of trustworthiness.

In terms of partner effects, we found two interesting findings. First, participants were more likely to risk money with robot partners than human partners following the trust violation (regardless of the quality of that violation) over the entire experiment, complicating the interpretation that violations of ability (and integrity) should reduce trust more so when attributed to machines compared to humans [17, 32]. Our results demonstrated no partner differences prior to the trust violation. However, once trust was violated, participants performed less trust behaviors toward the human than the robot. This may be because we had three different conditions with different trust violations which either were biased towards the robot [H_4, H_{10}] or towards the human [H_7]. Thus, we may not have had the statistical power to differentiate between the conditions. Second, we found a marginal effect of manipulation and partner on integrity perceptions. Specifically, if the violation was ability-based, participants perceived higher integrity of the human than the robot. In contrast, if the violation was integrity-based, participants perceived the robot as having higher integrity. We can understand these findings in the context of the unique agent hypothesis. If a robot performs poorly (ability violation), participants may perceive the robot as having less integrity as it should perform without errors. In contrast, participants may have been more forgiving of a human running the task. When an integrity violation occurs, participants perceived higher integrity for the robot than the person. This may be due to the fact that robots lack true intentionality. In other words, the participants may not have perceived the robot as being able to betray, whereas a human has the intentionality to do so. Indeed, Alarcon et al. [18] noted that both benevolence *and* integrity violations may signal intentionality and may thus be more appropriate ascriptions of humans. However, this finding should be interpreted with caution as the marginal interaction is not statistically significant at the $p < .05$ cutoff.

4.2. Implications

Overall trust perceptions and behaviors decreased over time as expected. However, we found few effects of partner type. These findings are somewhat in line with mapping these trustworthiness dimensions on to the CASA [13], in that it appears that the participants attributed most of these trustworthiness dimensions to their robot partners, or at least in a similar fashion to how they were attributed to the human partners (see also [18]). However, we did find support for the unique agent hypothesis in the marginal interaction of the manipulation and partner on integrity and on the interaction of partner and time on behaviors.

Additionally, the ability-based trust-violation condition did result in lower ability perception ratings overall, and overtime, this is likely due to the nature of the task itself, rather than partner type. In fact, it appears that participants' trustworthiness perceptions, overall, did not vary across partner types over time. This implies that participants perceived the "humanlike" robot and the human partners to be, roughly, equally trustworthy and that the PAS bias did not differentiate the trust process in this task, with the marginal exception of

integrity. As similar findings were reported by Alarcon et al. [18], it may be that Checkmate relies heavily on performance perceptions to shape trust, regardless of whether the partner is a human or a robot. When the task is dependent on performance, it may be less relevant if the reason for the trust violation can be attributed to non-performance aspects (benevolence and integrity). Yet again, we see how important contextual constraints are at shaping the trust process in human-human and human-robot interaction (see also Sanders et al., [19]).

It was found that participants were more likely to risk more money with a robot partner than a human partner after a trust violation. This was unexpected in that the effect of trust violations were expected to be qualified by partner type, so that participants should have been more cautious with the robot partner (i.e., risked less money than the human partner) following ability and integrity (but not benevolence) trust violations. However, the results indicate participants trusted the robot partner's potential to provide them with greater monetary returns than the human partner regardless of the quality of the violation.

4.3. Limitations and Future Research

Although these results may seem at odds with the models that were originally proposed, there are several possible interpretations that could help to explain these differences. First, many of these models were built using data that involved human-computer/human-machine interactions and not human-robot teaming. As the NAO robot partner had more human characteristics (i.e. arms, legs, head, face, human name, and voice) than a computer or a less anthropomorphic robot, this could explain the lack of differences between partner types (see also [18]). Secondly, the participant was only introduced to their human partner at the very beginning of the study, before being isolated to separate rooms. There were no natural interactions between the participant and the human partner (e.g., no way to chat, or see each other) throughout the experiment, and participants could only see the result of the choices their confederate, human partner was "selecting" within in the trust game. Therefore, the interaction with the human partner is somewhat unnatural and heavily computer mediated. This contrasts with the robot partner being in the same room as the participant, making movements and vocalizations throughout the experiment. The human partner in a separate room was done to ensure the participant thought they were playing with a human. If they were in the same room, the participant might be able to glance over at the human confederate's screen and realize everything was pre-recorded. It would be interesting to see if these results would still hold true with more natural human-human teaming and/or less anthropomorphic robots. Third, the robot keeping part of the profits was done to instantiate risk, however this may have been confusing for participants as robots do not have use for monetary rewards. Future research should explore ability, benevolence and integrity violations in more ecologically valid scenarios to ensure the validity of the current study. Finally, our sample is underpowered which lead us to interpret our results with caution. However, at present, we are collecting data to bolster the power of this study and aim to present these data in a peer-reviewed manuscript.

10. References

[1] J.Y. Chen and M.J. Barnes, "Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues", IEEE Transactions on Human-Machine Systems, IEEE, New York, United States, 2014, pp. 13-29.

[2] S.T. Fiske, "or Minus Five", In Motivated Social Perception: The Ontario symposium, Psychology Press, United Kingdom, 2003, pp. 233

[3] I. Thielmann and B.E. Hilbig, "Trust: An Integrative Review from a Person–Situation Perspective", Review of General Psychology, SAGE Publications, California, United States, 2015, pp. 249-277.

[4] R.C. Mayer, J.H. Davis, and F.D. Schoorman, "An Integrative Model of Organizational Trust", Academy of Management Review, Academy of Management, United States, 1995, pp. 709-734.

[5] S.L. Jones and P.P. Shah, "Diagnosing the Locus of: A Temporal Perspective for Trustor, Trustee, and Dyadic Influences on Perceived Trustworthiness", Journal of Applied Psychology, American Psychological Association, United States, 2016, pp. 392.

[6] J.L.Wildman, M.L. Shuffler, E.H. Lazzara, S.M. Fiore, C.S. Burke, E. Salas, S. and Garven, "Trust Development in Swift Starting Action Teams: A Multilevel Framework", Group & Organization Management, SAGE Publications, California, United States, 2012, 137-170.

[7] C.B. Crisp and S.L. Jarvenpaa, "Swift Trust in Global Virtual Teams: Trusting Beliefs and Normative Actions", Journal of Personnel Psychology, Hogrefe Publishing Group, Boston, United States, 2013, pp. 45.

[8] M. Foddy, M.J. Platow, and T. Yamagishi, "Group-Based Trust in Strangers: The Role of Stereotypes and Expectations", Psychological Science, SAGE Publications, California, United States, 2009, pp. 419-422.

[9] M. Stirrat and D.L. Perrett, "Valid Facial Cues to Cooperation and Trust: Male Facial Width and Trustworthiness", Psychological Science, SAGE Publications, California, United States, 2010, 349-354.

[10] S.G. Guthrie, "Faces in the Clouds: A New Theory of Religion", Oxford University Press, England, 1993

[11] P. Madhavan, D.A. Wiegmann, "Similarities and Differences Between Human–Human and Human–Automation Trust: An Integrative Review", Theoretical

Issues in Ergonomics Science, Taylor & Francis, United Kingdom, 2007, pp. 277-301.
[12] J.D. Lee and K.A. See, “Trust in Automation: Designing for Appropriate Reliance”, Human Factors, SAGE Publications, California, United States, 2004, pp. 50-80.
[13] C. Nass and Y. Moon “Machines and Mindlessness: Social Responses to Computers”, Journal of Social Issues, Wiley-Blackwell, New Jersey, United States, 2000, pp. 81-103.
[14] E.J. de Visser, S.S. Monfort, R. McKendrick, M.A. Smith, P.E. McKnight, F. Krueger, and R. Parasuraman, “Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents”, Journal of Experimental Psychology: Applied, American Psychological Association, United States, 2016, pp. 331-349.
[15] C. Nass, J. Steuer, and E.R. Tauber, “Computers are Social Actors”, In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, United States, 1994, pp. 72-78.
[16] K.T. Wynne and J.B. Lyons, “An Integrative Model of Autonomous Agent Teammate-Likeness”, Theoretical Issues in Ergonomics Science, Taylor & Francis, United Kingdom, 2018, pp. 353-374.
[17] M.T. Dzindolet, L.G. Pierce, H.P. Beck, and L.A. Dawe, "The Perceived Utility of Human and Automated Aids in a Visual Detection Task”, Human Factors, SAGE Publications, California, United States, 2002, pp. 79-94.
[18] G.M. Alarcon, A.M. Gibson, S.A. Jessup, A. Capiola, A. “Exploring the Differential Effects of Trust Violations in Human-Human and Human-Robot Interactions”, Applied Ergonomics, Elsevier, Netherlands, 2021, 103350.
[19] T. Sanders, A. Kaplan, R. Koch, M. Schwartz, and P.A. Hancock, “The Relationship Between Trust and Use Choice in Human-Robot Interaction”, Human Factors, SAGE Publications, California, United States, 2019, pp. 614-626.
[20] G.M. Alarcon, J.B. Lyons, J.C. Christensen, S.L. Klosterman, M.A. Bowers, T.J. Ryan, S.A. Jessup, and K.T. Wynne, “The Effect of Propensity to Trust and Perceptions of Trustworthiness on Trust in Dyads”, Behavior Research Methods, Springer, New York, United States, 2018, pp. 1906-1920.
[21] J. Berg, J. Dickhaut, and K. McCabe, “Trust, Reciprocity, and Social History”, Games and Economic Behavior, Elsevier, Netherlands, 1995, pp. 122-142.
[22] P.A. Hancock, T.T. Kessler, A.D. Kaplan, J.C. Brill, and J.L. Szalma, “Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses”, Human Factors, Advanced Online Publication, SAGE Publications, California, United States, 2020, 0018720820922080.
[24] K.L. Mosier and L.J. Skitka, “Human Decision Makers and Automated Decision Aids: Made for Each Other? In R. Parasuraman & M. Mouloua (Eds.)”, Automation and Human Performance: Theory and Applications, Taylor & Francis, United Kingdom, 1996 pp. 201-220.
[25] A. Powers, A.D. Kramer, S. Lim, J. Kuo, S.L. Lee, and S. Kiesler, “Eliciting information from people with a gendered humanoid robot”, In ROMAN 2005 IEEE International Workshop on Robot and Human Interactive Communication, IEEE, New York, United States, 2005, pp. 158-163.
[26] I. Leite, A. Pereira, S. Mascarenhas, C. Martinho, R. Prada, and A. Paiva, “The Influence of Empathy in Human–Robot Relations”, International Journal of Human-Computer Studies, Elsevier, Netherlands, 2013, pp. 250-260.
[27] P.H. Kahn Jr, T. Kanda, H. Ishiguro, B.T. Gill, J.H. Ruckert, S. Shen, H.E. Gary, A.L. Reichert, N.G. Freier, R.L. Severson, “Do People Hold a Humanoid Robot Morally Accountable for the Harm it Causes?”, In Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, IEEE, New York, United States, 2012, pp. 33-40.
[28] J.A. Colquitt, J.A. LePine, C.P. Zapata, R.E., Wild, “Trust in Typical and High-Reliability Contexts: Building and Reacting to Trust Among Firefighters”, Academy of Management Journal, Academy of Management, United States, 2011, pp. 999-1015.
[29] R. Parasuraman, R. Molloy, and I.L. Singh, “Performance Consequences of Automation-Induced Complacency”, The International Journal of Aviation Psychology, Taylor & Francis, United Kingdom, 1993, pp. 1-23.
[30] B.M. Muir, “Trust in Automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems”, Ergonomics, Taylor & Francis, United Kingdom, 1994, pp. 1905-1922.
[31] S. Ososky, E. Philips, D. Schuster, and F. Jentsch, “A Picture is Worth a Thousand Mental Models: Evaluating Human Understanding of Robot Teammates”, In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications, California, United States, 2013, pp. 1298-1302.
[32] S.M. Merritt, J.L. Unnerstall, D. Lee, and K. Huber, “Measuring Individual Differences in the Perfect Automation Schema”, Human Factors, SAGE Publications, California, United States, 2015, pp. 740-753.
[33] R.C. Mayer, and J.H. Davis, “The Effect of the Performance Appraisal System on Trust for Management: A Field Quasi-Experiment”, Journal of Applied Psychology, American Psychological Association, United States, 1999, pp. 123.
[34] Alarcon, G.M., Capiola, A., Lee, M. A. and Jessup, S.A. “Comparing Decision-making Models in a Trust Game: The Effects of Trustworthiness Manipulations on Perceptions and Risk-taking Behaviors.” Manuscript under review.
[35] J.A. Colquitt, B.A. Scott, and J.A. LePine, “Trust, Trustworthiness, and Trust Propensity: A Meta-Analytic Test of their Unique Relationships with Risk Taking and Job Performance”, Journal of Applied Psychology, American Psychological Association, United States, 2007, pp. 909-927.