AFRL-RH-WP-TR-2021-0076

AUTOMATION BIASES IN HUMAN-ROBOT TRUST INTERACTIONS

Dr. Gene Alarcon
711th HPW/RHWC
Wright-Patterson AFB, OH 45433

April 2021

FINAL REPORT

AIR FORCE RESEARCH LABORATORY
711th HUMAN PERFORMANCE WING
AIRMAN SYSTEMS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the AFRL Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2021-0076 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

ALARCON.GENE.MICHAEL.1139820521
Digitally signed by ALARCON.GENE.MICHAEL.1139820521
Date: 2022.01.26 11:34:28 -05'00'

GENE ALACRON, DR-III, Ph.D.
Program Manager
Collaborative Interfaces & Teaming Branch
Airman Systems Directorate
711th Human Performance Wing

WEBB.TIMOTHY.S.1085032239
Digitally signed by WEBB.TIMOTHY.S.1085032239
Date: 2022.01.28 10:57:56 -05'00'

TIMOTHY S. WEBB, DR-IV, Ph.D.
Chief, Collaborative Interfaces & Teaming Branch
Airman Systems Directorate
711th Human Performance Wing

CARTER.LOUISE.ANN.1230249128
Digitally signed by CARTER.LOUISE.ANN.1230249128
Date: 2022.02.09 08:53:35 -05'00'

LOUISE A. CARTER, DR-IV, Ph.D.
Chief, Warfighter Interactions and Readiness Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
April 2021	Final	7 June 2018	7 April 2021

3. TITLE AND SUBTITLE
Automation Biases in Human-Robot Trust Interactions

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
In House		
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
		H0WG

6. AUTHOR(S)
Dr. Gene Alacron

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Air Force Material Command Air Force Research Laboratory 711th Human Performance Wing Airman Systems Directorate Warfighter Interactions and Readiness Division Wright-Patterson AFB, OH 45433	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'SREPORT NUMBER(S)
Air Force Material Command Air Force Research Laboratory 711th Human Performance Wing Airman Systems Directorate Warfighter Interactions and Readiness Division Wright-Patterson AFB, OH 45433		AFRL-RH-WP-TR-2021-0076

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A. Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
AFRL-2021-3692; Cleared 22-Oct-2021

14. ABSTRACT
This final report summarizes the examination of possible human biases in trust between human-human and human-robot teaming. The studies found humans do hold biases against robot partners. Specifically, robot partners were suffered greater trustworthiness declinations after a trust violation than did human partners but only when lower trustworthiness could be ascribed to specific aspects. That is, there were differences between the types of manipulation, such that performance/ability violations were much stronger than benevolence or integrity violations. Additionally, the perfect automation schema (PAS) construct significantly predicted slope variance in benevolence perceptions following a trust violation.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	37

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Dr. Gene Alacron	

TABLE OF CONTENTS

LIST OF FIGURES

SUMMARY

This final report summarizes the examination of possible human biases in trust between human-human and human-robot teaming. The studies found humans do hold biases against robot partners. Specifically, robot partners were suffered greater trustworthiness declinations after a trust violation than did human partners but only when lower trustworthiness could be ascribed to specific aspects. That is, there were differences between the types of manipulation, such that performance/ability violations were much stronger than benevolence or integrity violations. Additionally, the perfect automation schema (PAS) construct significantly predicted slope variance in benevolence perceptions following a trust violation.

1.0 THEORETICAL BACKGROUND OF BIASES BETWEEN HUMAN-HUMAN AND HUMAN-ROBOT TEAMING

Robots are becoming more prevalent in many life contexts ranging from manufacturing (Davies et al., 2000), surgery (Schäfer et al., 2019), education (Belpaeme et al., 2018), and therapy (Chang & Kim, 2013). Robots have increasingly been used in a variety of social interactions (Young et al., 2009), departing from the more traditional view of robots as machines. For example, robots are being used for companionship in nursing homes (Bemelmans et al., 2012) and in autism therapy (Pennisi et al., 2016). There are burgeoning theories on human-machine teaming, which explore the machine agent as a teammate rather than a tool (Wynne & Lyons, 2018). However, the benefits of robot assistance are realized only when robots are accepted by the population. Trust undoubtedly plays a role in Human-Robot Interaction (HRI), and the last two decades have seen extensive research on trust in HRI (see Hancock et al., 2011, 2020). However,experiments investigating the role of anthropomorphism on trust are not numerous (see Hancock et al., 2020). Although much research has examined how humans interact with automation, machines, and decision support systems (DSS), more research is needed to determine whether findings from trust in automation research generalize to research investigating trust toward anthropomorphized robot partners in more social contexts (rather than those contexts where the robot is simply used as a tool). Furthermore, the majority of the trust in automation, machines, and DSS research has focused on performance manipulations, largely ignoring other theoretically relevant aspects of trustworthiness posited by researchers (Lee & See, 2004).

1.1 Models of Human-Machine Interactions

Two dominant paradigms have attempted to understand how humans interact with machines and how this interaction may or may not differ from human-human interactions. First, the computers as social actors ([CASA]; Nass & Moon, 2000) paradigm (also referred to as media-equation; Nass & Moon, 2000) hypothesizes humans attribute human-like characteristics (e.g., personality, gender) to machines such as computers (Nass et al., 1994) and automated website purchase assistants (Nass & Lee, 2001) in much the same way as they do toward other humans. The CASA paradigm proposes people treat computers as social actors. In this paradigm, the constructs that govern human-human interactions (e.g., similarity-attraction; Nass & Lee, 2001) are the same constructs that govern human-machine interactions. In their review of more than a decade of research, Nass and Moon (2000) noted that participants "mindlessly" apply social expectations and rules to computers.

Across CASA studies, participants reacted to gender and ethnic stereotypes, politeness, and reciprocity in non-anthropomorphic computers, even though in post-experimental surveys the participants note computers do not have any feelings or a sense of self (Nass et al., 1994; Nass & Moon, 2000). The researchers attribute these findings to participants relying on "scripts drawn in the past" (p. 83). These scripts for social interactions drive the relationships of both human-human and human-machine interactions. Although the CASA paradigm was developed based on how humans socially respond to computers, research has shown that the model can also be applied to other types of machines such as robots. Lee et al. (2006) demonstrated that participants applied personality-based social rules (e.g., complementary attraction) to a robotic dog, providing evidence that the CASA paradigm can be expanded to HRI. Despite the vast

amount of research supporting their theory, other research and models postulate differences between human-human and human-machine interactions.

In contrast to the CASA model (Nass & Moon, 2000), a growing breadth of research has demonstrated significant differences between how one interacts with a machine and with a human (e.g., de Visser et al., 2016; Madhavan & Wiegmann, 2007; Smith et al., 2017), and this has been referred to as the unique-agent hypothesis (de Visser et al., 2016). The unique-agent hypothesis posits that people view machines differently than humans based on the characteristics of the machine. As such, each referent (i.e., human or machine, respectively; Hoff & Bashir, 2015) and situation is unique, and may or may not evoke similar mental models used during human-human interactions. Madhavan and Wiegmann (2007) emphasized that humans are biased to perceive DSS to perform at near perfect levels (i.e., automation bias), which may significantly reduce trust in the system following a perceived DSS error (see Dzindolet et al., 2002). Their model integrated theoretical differences in trust judgements, monitoring behaviors, and schemas of automation (Dzindolet et al., 2001, 2002, 2003). The model hypothesizes that when humans are partnered with a DSS, users will be more critical of its failures compared to failures from human partners (see Dzindolet et al., 2002). These differences have been discussed in relation to several constructs such as trust (e.g., de Visser et al., 2016; Madhavan & Wiegmann, 2007; Xie et al., 2019), moral accountability (e.g., Kahn et al., 2012), and social interactions (Wang & Quadflieg, 2015). Compared to people in human-human social interactions, people view robots in HRI as less believable, less intelligent, less capable of experiencing emotions, and more eerie (Wang & Quadflieg, 2015). Additionally, people believe that humans should be held morally accountable for their actions to a higher degree than robots (Kahn et al., 2012). Researchers have also found that increased anthropomorphism of an agent activates neural areas that are attributed to more human-like capabilities such as theory of mind (Krach et al., 2008).

Prior research has delineated how people view and treat machines into one of two models (e.g., CASA, unique-agent), finding empirical support for both. Research supporting the CASA model (Nass & Moon, 2000) has shown that people interact with machines in a similar way as they do with human counterparts during interactions, often applying social norms to interactions with machines (e.g., Lee et al., 2006; Nass et al, 1994). However, with technology advances—such as machines with increased capabilities and diversity (e.g., therapy, military assistance, social companion)—researchers (e.g., Wang & Quadflieg, 2015) have found differences in how humans treat machines during social interactions, compared to other humans, which supports the unique-agent model (de Visser et al., 2016). One factor that has demonstrated influence on humans' perceptions of machines (both similar *and* different to perceptions of humans) is trust.

1.2 Trust in Machines vs Trust in Humans

At its most basic, trust is conceptualized as the willingness to rely on another (Mayer et al., 1995). Though trust was first investigated as an important construct in organizational and interpersonal contexts (e.g., Mayer & Davis, 1999; McAllister, 1995; Rotter, 1980; see also Colquitt et al., 2007), trust research has been expanded to the automation (Hoff & Bashir, 2015; Lee & See, 2004) and robotics (Hancock et al., 2011; Lee & Seppelt, 2009) literatures. Trust toward machines (i.e., automation, robots, computers, etc.), similar to interpersonal trust (Colquitt et al., 2007), has been delineated into trustor beliefs (e.g., automation schemas, trustworthiness perceptions; Hoff & Bashir, 2015), trust intentions (i.e., a willingness to be

vulnerable; Calhoun et al., 2019), and behavioral manifestations of trust (e.g., reliance, trust behaviors, monitoring behaviors; Lee & See, 2004; Madhavan & Wiegmann, 2007; Parasuraman & Riley, 1997; Sanders et al., 2019). For the purpose of this paper, we focus on trustworthiness perceptions, trust intentions, and trust behaviors.

Trust intentions and trust behaviors are the same theoretically across the interpersonal (Mayer et al., 1995) and machine trust literatures (Lee & See, 2004). Trust intentions are the willingness of the trustor to be vulnerable to the referent (i.e., human or machine, respectively; Hoff & Bashir, 2015). Trust behaviors are the subsequent observable risk-taking actions of the trustor. In contrast, there are differences in conceptualizations of trustworthiness across the interpersonal and machine trust literatures. Lee and See (2004) proposed that perceptions of automation performance, purpose, and process are antecedents to user trust toward automated systems. These antecedents were informed by Mayer and colleagues' (1995) organizational trust model, where perceptions of a human referent's ability (how competent are they?), benevolence (do they have my best interest in mind?), and integrity (do they have similar principles as I?) inform a trustor's willingness to be vulnerable (i.e., trust) and ultimately their decision to act (i.e., risk-taking behavior). In a similar way, the trust in automation literature proposes performance (what does the automation do and how well does it do that task?), purpose (why was the automation developed?), and process (how does the automation do its task?), respectively, are informative antecedents to trust automation and overlap conceptually with antecedents highlighted in Mayer et al.'s model of trust (see Lee & See, 2004, p. 59). However, the empirical data on automation bias and trust that inform Madhavan and Wiegmann's (2007) theoretical model has focused predominantly on perceptions of automation performance, omitting potentially important attributions of purpose and process perceptions on user's trust (pp. 280-281).

1.3 HRI

Although there has yet to be a universally agreed upon definition of robots, the Institute of Electrical and Electronics Engineers (IEEE) defined robots as, "autonomous machine[s] capable of sensing its environment, carrying out computations to make decisions, and performing actions in the real world" (Guizzo, 2020). Additionally, in comparing robots to automation, Hancock et al. (2011) noted that "robots differ from most automation in that they are mobile…sometimes built in a fashion that approximates human or animal form and are often designed to effect action at a distance" (pp. 518-519). These sentiments were also echoed by Salem et al. (2015). As robots are adapted to resemble humans more accurately, humans' biases toward automation may be less applicable to humanized robot partners (see Salem et al., 2015). There is a growing literature on trust in robots, indicating its importance for a variety of fields. Specifically, meta-analytic research has proposed a particularly relevant aspect of robots in relation to trust is anthropomorphism (Hancock et al., 2011, 2020).

1.3.1 Anthropomorphism.

Anthropomorphism is the degree to which an agent expresses human-like characteristics, often leading to humans ascribing intent, motivation, goals, or a sense of self to the agent (Epley et al., 2007). Advances in technology and robotics have led to the advent of humanoid social robots such as the Nao robot and pet robots such as Aibo, the robotic dog. Research in social robotics has demonstrated humans prefer anthropomorphic robots more than non-anthropomorphic robots. For example, robots with a humanoid face are more preferred than non-humanoid robots

(Broadbent et al., 2013; Hertz & Wiese, 2017; Smith & Wiese, 2016). When a robot is viewed as more mechanistic than human-like, humans tend to act less politely to the agent, treat it as more subservient, and lower their expectations of the agent (Hinds et al., 2004). Humans form a mental model of humanoid robots by making unconscious assumptions about them and use these assumptions to gauge knowledge, skills, and performance expected from the robot. However, people do not treat robots exactly the same way they treat other living things. For example, Friedman et al. (2003) found participants did not treat a robotic dog, Aibo, the same as a real dog. They found participants did not view the robotic dog as responsible for its actions as a real dog. Researchers have suggested it is unlikely humans will treat robots the same as other humans and living creatures because of the different mental models and lack of intention on part of the robot (Dautenhahn, 2002; Dautenhahn et al., 2006). However, the CASA model has demonstrated individuals do treat machines the same as humans, as they unknowingly rely on the social interaction norms, despite stating they should not (Nass & Moon, 2000). Therefore, the key question is: are there differences in human-human and human-robot trust when the latter interaction comprises an anthropomorphized agent?

1.3.2 Anthropomorphism and Trust

Research on the role of anthropomorphism in trust in robots has increased in recent years, but findings from the research has supported both the CASA (Nass & Moon, 2000) paradigm and the unique-agent hypothesis (de Visser et al., 2016). Furthermore, the previous research has almost exclusively explored the robot as a tool rather than as a teammate in a social context. Xie et al. (2019) examined trust in a human partner (an experimental confederate—that is, someone who is privy to the aim of the experiment unbeknownst to the participants) and a non-anthropomorphic robotic partner in an unmanned aerial vehicle (UAV) task, in which both the human confederate and robot ran the same software algorithms. Participants played various UAV tasks (e.g., searching, firefighting) with a partner whom they were told was a human or a robot. Their results showed that participants reported higher trust toward the simulated human rather than the simulated robot partner. However, they did not find any differences in trust behaviors.

A study by Sanders et al. (2019) found participants were more likely to choose to use a robot in a dangerous context, which in their study was an improvised explosive device disposal task. In contrast, participants stated that “being human” or “having a brain” makes the human less equipped for engaging in the dangerous task. In both instances, trust predicted use (i.e., trust behavior). However, the study did not directly compare trust assessments between the human referent and the robot referent, nor was the robot anthropomorphic. As anthropomorphism becomes more prevalent in robots, it is important to understand how humans trust them, and what aspects of trustworthiness influence this trust (i.e., purpose, process, or performance). As such, research directly comparing humans to anthropomorphic robots remains sparse.

Anthropomorphism of a partner has also been shown to influence initial trustworthiness perceptions. For instance, people trusted and complied more with a computer agent, compared to an avatar or human agent, in a digit pattern recognition task where the referent offered suggestion of what number ought to come next in the pattern, demonstrating automation bias (de Visser et al., 2016). However, when reliability of the computer decreased, trust rapidly declined compared to the decreased reliability of the avatar and human. Indeed, they found that following a trust violation, participants declined less in their trust perceptions toward an anthropomorphic robot compared to their non-anthropomorphic counterparts. Somon and colleagues (2019) also

found the same neurological responses to an automated agent (e.g., a computer) in the N2-P3 components as when monitoring a human. However, cluster analysis demonstrated a decrease in signal in the P3 component (response clusters associated with detected novelty) when monitoring a human, indicating mixed neurological results. That is, differing data analytic techniques in the same study yielded results supporting similar or differing neurological responses toward a human versus a robot error.

Although research exists on HRI in performance- and reliability-based scenarios as in Sanders et al. (2019), little experimental research has been conducted on human-robot interaction in social contexts. Researchers (Mota et al., 2016) conducted a pilot experiment on humans' perceptions of robots during the Trust Game (see Berg et al., 1995), where participants gave an allotment of money to a robot, the allotment was tripled, and the robot subsequently returned an amount higher or lower than expected by the human. Note that in scenarios such as the Trust Game, there is little chance for the robot to display varying task competence (i.e., performance), yet violations of trust may be attributed to the robots perceived purpose or process when a return fails to match participants' expectations. However, due to a small sample size ($N = 5$), Mota et al. were unable to investigate quantitative effects of non-performance-related trust violations on human perceptions to the robot referent.

Other researchers (Tulk & Weise, 2018) have investigated the effects of physical human-likeness (a component of anthropomorphism) on trust toward a robot during the Ultimatum Game (see Güth et al., 1982)—where participants simply decided whether to accept or reject the monetary offer from robot—as well as the Trust Game. Though the researchers found that fairer offers in both trust games were associated with higher acceptance rates and return offers in the Ultimatum Game and Trust games, respectively, they found no effect of human-likeness on behavioral trust (i.e., accepting offers, returning more of the endowment). However, Tulk and Weise (2018) found that human-likeness was positively related to perceived approachability of the robot in the Ultimatum Game and self-reported trust in the Trust Game. These subjective assessments mediated the relation between perceived human-likeness and behavioral trust. However, Tulk and Weise were unable to investigate the differences between non-performance-related trust violations from humans versus robots on trust because their experiment(s) did not comprise a human condition for comparison. Moreover, the robot was not physically present but was manipulated to display more/less visible human-likeness and displayed on a computer monitor. As such, it remains to be seen how such trust violations from a human or a physically present robot differentially affect trust.

2.0 METHOD

2.1 Task

Participants were asked to play Checkmate (Alarcon et al., 2018), a two-player modified investor/dictator game. Participants completed the study with a confederate acting as the second player. Participants were assigned the role of "banker" (investor), while the confederate was assigned the role of "runner" (dictator). The experiment was only available weekdays during normal business hours to ensure participants believed they were playing the game with a person in the laboratory. All behaviors carried out by the runner were automated, pre-recorded, and only varied by participant condition and manipulation. The game was played over the course of one practice round (Round 0) and five actual rounds (Rounds 1-5), with the banker tasked with loaning money to the runner each round and the runner tasked with collecting as many boxes as possible in a virtual maze each round.

Following an initial practice round to introduce the participant to the game, they were informed that any money exchanged over the course of the task represented real money. The amount of money the player had after the final round was paid out via Amazon Mechanical Turk (MTurk). Participants were initially given a certain amount of money ($50 in person, $25 MTurk) in their account, which they used to loan money to the runner before each round. Money loaned to the runner had the potential to gain interest or be lost based on the runner's performance in the upcoming round. The runner selected a promised return and a promised risk level for the task. A higher risk level meant they could win more money if the runner performed well, but could lose more money if the runner performed poorly. The selected risk level could be low (75–150%), moderate (50–200%), or high (0–300%). For example, if the risk level was moderate, then the maximum loss would be 50% (100%–50% = 50%) if the runner performed as poorly as possible, while the maximum gain would be 100%. The participant/banker then chose a risk level for their investment. The participant was able to select a loan amount using a slider, which they could adjust to select a loan amount in one cent increments within an available range ($1–$7).

After the banker made their decisions, the runner selected an actual risk level. This process was automated so that a moderate (50–200%) risk level would always be selected. Afterwards, the runner would complete the maze running task while the banker observed the performance. Prior to the beginning of the task, both banker and runner were able to preview the maze from a top-down perspective. The goal of the round was to collect boxes of varying colors distributed throughout the maze within a time limit. If the runner successfully obtained a sufficient number of boxes, the runner would receive earnings based on the banker's loaned amount and the risk level decided at the beginning of the round. If the runner performed poorly, the runner would lose a portion of the banker's loan based on their performance and risk level. While the runner is completing the task, the banker is able to see the runner's performance from a top-down perspective where the participant can view the runner's position on the map and the number of boxes collected by the runner, but the banker is not given information as to how much money the runner earned in the round. After completing the task, the runner then sends a desired amount back to the banker. The banker is then informed of how much money was received from the runner. This process is carried out over five rounds, with earnings and losses being carried across all rounds.

2.2 Manipulations

Participants were randomly assigned to either a control condition or one of three trust violation conditions depending on the study. In the control condition, the runner successfully completed the maze and returned the promised amount for all five rounds. In the trust violation conditions, the runner successfully completed the maze and returned the promised amount for the practice session and first two rounds as well as the final found, but the runner committed a trust violation in the third and fourth rounds, which are outlined below. Importantly, across trust violation conditions, the banker received the same proportion of money back when the runner displayed untrustworthy behavior (i.e., less money than previous rounds), but the reason for this was designed to be due to lowered ability, integrity, or benevolence. Specifically, when the runner performed a trust violation, the runner returned only 50% of the loaned amount. After the first iteration of the Checkmate protocol, we manipulated ability, benevolence, and integrity with the following experimental manipulations in the task. Note that not all manipulations were conducted in every study. When we discuss the studies we discuss the manipulations employed.

2.2.1 Ability

In order to manipulate perceptions of ability, we degraded the performance of the runner in rounds three and four. During the rounds in which the runner displayed ability-based trust violations, the runner performed poor enough in the maze running task that there was a resulting loss in the banker's investment. To make the ability manipulation salient, the runner performed behaviors such as running into buildings and going off into areas where there were no boxes to collect. As a result, the runner was unable to return the amount they promised to the banker. During all rounds, the banker was able to watch the runner's performance via the overhead view as well as monitor the box count in the upper right-hand corner of the display. However, the banker was unaware of the exact dollar amount earned by the runner. Thus, the banker inferred the ability of the runner through various stimuli (i.e., overhead view of performance, box count), but the banker was unable to verify the runner's *actual* behavior (i.e., the banker was not able to see how much was earned and thus what percentage was returned). The runner committed the same violation again in the fourth round, then performed trustworthy behaviors in the fifth round.

2.2.2 Integrity

As mentioned earlier, integrity is a perception the trustee will fulfill their promise or adhere to an ethical standard the trustor finds acceptable. If the banker is promised a certain payment and the runner's performance is maintained, it can be reasoned that the runner has earned the promised amount to send back to the banker. A less than expected amount of money sent back from the runner to the banker can thus be viewed as an integrity violation (i.e., a lack of fulfilling the word-action promise). In order to manipulate perceptions of integrity, we manipulated the amount returned to the runner as an integrity-based trust violation while keeping performance consistent. Performance remained consistent throughout all rounds with the runner collecting approximately the same number of boxes in every round. However, during the third and fourth rounds, the runner chose to renege on their promised return to the banker, keeping the majority of the loaned and earned amount for themselves. Given that performance remained the same across all rounds, the runner sending less money back would indicate a lack of fairness and/or honesty, indicating an integrity-based trust violation.

2.2.3 Benevolence

In order to manipulate perceptions of benevolence, we modified both initial instructions for the task and the performance of the runner. Participants received additional instructions for how earnings were calculated in the game. Specifically, they were told that the color of the boxes collected by the runner affect how earnings were distributed at the end of the round. 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 other words, any money earned by collecting red boxes cannot be sent to the banker. In the first two rounds and the final round of the game, the runner primarily collected boxes that benefited both players. During rounds three and four, the runner chiefly collected red boxes that only benefited themselves. The number of boxes collected was similar across all six rounds, with the main difference being what color of boxes were targeted by the runner. Thus, we were able to maintain performance so that any differences in the amount of money returned when there was a benevolence-based trust violation were due to the collection of boxes that could not be shared, and not performance itself.

3.0 HYPOTHESES

Studies in human-robot trust in teams (Human-Machine Teams [HMT]) are relatively nascent (Ososky, Schuster, Phillips, & Jentsch, 2013). The literature has explored some human-robot trust interactions. A recent review of the HMT literature found two types of errors are prominent: evaluation errors where the operator follows the wrong guidance from the agent and intent errors where the operator is aware that the agent has a superior solution but the operator fails to follow it (Chen & Barnes, 2014). A recent meta-analysis (Hancocket al., 2020) demonstrated robot performance (e.g. failure rate, reliability) related factors were the strongest predictors of trust. This not to say attribute related factors (e.g., personality, anthropomorphism) did not relate to trust, just not as strongly as performance. This is supported by the literature on automation schema.

Merritt and colleagues (2015) have researched the notion of the PAS and found that it is related to higher trust in automated systems. The premise of the PAS is that humans may hold a view of automation that they are close to perfect and error-free. PAS may make humans less forgiving of automated systems because performance errors may violate their existing expectations of technology. Research has shown that PAS can help individuals be more sensitive to positive changes in reliability, however it can be detrimental to individuals when reliability is perceived to decrease (Pop, Shrewsbury, & Durso, 2015). Dzindolet et al. (2003) also found humans have a positive bias toward automation due to better perceived reliability. In contrast, a meta-analysis on interpersonal trust illustrated ability, benevolence, and integrity all had similar effects on one's willingness to be vulnerable (Colquitt et al., 2007). Given the importance of performance as a significant driver of trust of robotic systems, it is expected that perceived ability will have a stronger effect on the perceived trustworthiness of the robot relative to the human.

- Hypothesis 1: Perceived initial ability/performance in robotic partner will be higher than with a human partner (intercept differences) and initial wagers will be higher with a robotic partner.
- Hypothesis 2: Initial wagers with a robotic partner will be higher than with a human partner (intercept differences).
- Hypothesis 3: Ability/performance will have a stronger effect on perceived trustworthiness of a perceived robotic partner than a human partner, over time (slope differences).
- Hypothesis 4: Ability/performance will have a stronger effect on change in wagers of a perceived robotic partner than a human partner, over time (slope differences).

As mentioned above, 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 (Powers, 2005). 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 (Powers, 2005). Robots that display empathy may be more liked and trusted (Leite et al., 2013), suggesting the importance of intent from the robot. Intelligent agents that are more cooperative independent of reliability were more trusted by individuals in a shared resource

management task (Chiou & Lee, 2016), again suggesting that intent-based inference can be applied to technology. Yet, the state of the art in machine transparency focuses on state awareness and projection for tasks (Mercado et al. 2016) and providing rationale for recommendations (Lyons et al., 2016). More research is needed for understanding transparency of intent (Lyons 2013). There have been no studies to date (that the authors are aware of) which have directly compared intent from a human versus intent from a robot using a common task. The authors contend that signaling intent from a robot will influence trust perceptions but not as much as intent from another person. Some support for this notion can be drawn from research on how humans attribute blame for mistakes. Specifically, robots can be believed to be responsible for mistakes, though not as much as a human in the same situation (Kahn et al., 2012). Thus, it is plausible that the impact of signaling intent may be weaker when directly comparing robots versus humans.

Unlike humans, robots lack free will and hence, behave in accordance with their programming. However, machines may someday have both decision authority and decision initiative to act autonomously on the battlefield. In such a situation, 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. The idea of autonomous robots may invoke fear as individuals endorse the zeitgeist of "killer robots" (Lyons & Grigsby, 2016), yet we must consider the fact that autonomous robots may support non-combat roles such as information analysis and dissemination, decision aides, mission planning, aeromedical evacuation, inspection, among other areas. One notable safety system, the Automatic Ground Collision Avoidance System, is a highly-advanced form of automation that has the decision authority to take control away from a pilot when an imminent collision with the ground is detected. This fielded system has already saved multiple lives and notably, the pilot's perceived "benevolence" of the system is a significant driver of trust in the system (Lyons et al., 2016). Thus, it is plausible that intent from a robot matters, we just believe that it will matter more for humans versus robots.

- Hypothesis 5: Perceived initial benevolence/intent in robotic partner will be lower than a human partner (intercept differences).
- Hypothesis 6: Initial wagers in a robotic partner will be lower than a human partner (intercept differences) in the benevolence/intent conditions.
- Hypothesis 7: Intent perceptions will change faster with a robotic partner than with a human partner, over time (slope differences).
- Hypothesis 8: Wagers will change faster with a robotic partner than with a human partner, over time (slope differences) in the intent conditions.

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) is rational evaluation of past successes and or failures and has been shown to be the most important driver of trust in high-stakes interpersonal situations (Colquitt et al., 2011). Maintaining consistent, predictable behavior is a core antecedent to trust as this predictability allows one to forecast potential behavior in novel situations. Prior research has shown that consistency is a key to trust of automation (Parasuraman, Molloy, & Singh, 1993). The idea of predictability was one of the core antecedents for trust in automation during the initial studies of the construct (Muir, 1994). The literature on HMT suggests the importance of

shared mental models (Ososky et al., 2013; Wynne & Lyons, 2018) largely because shared mental models will help humans anticipate the actions and needs of their machine partners. Due to the asymmetry perceived capability between humans and technology, technology will likely pay a higher cost for deviations of consistency. Biases such as PAS may play a role in consistency as there may be a bias for automation to be perfectly consistent, along with perfect performance. Once the cognitive bias has been formed, any deviation from the bias greatly impacts trustworthiness perceptions.

- Hypothesis 9: Perceived initial consistency/integrity in robotic partner will be lower than a human partner (intercept differences).
- Hypothesis 10: Initial wagers in a robotic partner will be lower than a human partner (intercept differences) in the consistency/integrity conditions.
- Hypothesis 11: Reliability perceptions will decline faster with a robotic partner than with a human partner, when consistency/integrity is not stable over time (slope differences).
- Hypothesis 12: Wagers will decline faster with a robotic partner than with a human partner, when consistency/integrity is not stable over time (slope differences).

4.0 ACCOMPLISHMENTS

We finished data collection of in-person and online data collections. A total 953 participants were collected via Amazon MTurk for various iterations of the current Long-Range Imaging Radar (LRIR). A total of 204 participants were collected from in-person data collections. We outline results from each of the studies below. Overall, results have demonstrated a significant difference between human-human and human-robot interactions. Additionally, these differences span beyond performance manipulations and extend to benevolence and integrity manipulations. Furthermore, in the online data collections we found significant differences between rate of change (i.e., slopes) for the different trust manipulations, and PAS predicted the change in the slopes.

In our first study, we explored differences between human-human and human-robot teaming while holding performance constant. We used a mixed factorial design to examine the effects of trust and trust violations on human-human and human-robot interactions over time with an emphasis on anthropomorphic robots in a social context. We found consistent and significant effects of partner behavior. Specifically, partner distrust behaviors led to participants' lower levels of trustworthiness perceptions, trust intentions, and trust behaviors over time compared to partner trust behaviors. We found no significant effect of partnering with a human versus an anthropomorphic robot over time across the three dependent variables, supporting the computers as social actors (CASA; Nass and Moon, 2000) paradigm. However, we note that trust was manipulated by the partner simply not sending back as much as promised, without any information as to why. Results were published in *Applied Ergonomics*. As such, out next study was an attempt to manipulate ability, benevolence, and integrity through partner actions.

In Study 2, we created ability, benevolence, and integrity manipulations. For Study 2, we only explored human-human interactions as we were concerned with verifying the manipulations first, before attempting to assess their effects on human-robot teaming. We experimentally manipulated ability-, integrity-, and benevolence-based trust violations and measured trust-related criteria (i.e., trustworthiness perceptions and risk taking). We found differences between the ability-based trust violation and the integrity- and benevolence-based trust violations on both trustworthiness perceptions and risk-taking. However, no significant differences were found for differences in integrity- or benevolence-based trust violations. Figures 1, 2, 3 and 4 illustrate manipulation effects on ability perceptions, benevolence perceptions, integrity perceptions, and risk-taking behavior, respectively. As illustrated in the figures, ability had the strongest effect on all criteria. There were no discernable differences between the benevolence and integrity manipulations. Given that Study 2 was conducted online we theorized the lack of differences were due to information saliency, as participants may have rushed through training or forgotten the caveats to the benevolence condition. As such, we modified the protocol so that in the benevolence condition there was a reminder that the box colors have different meaning for future studies.

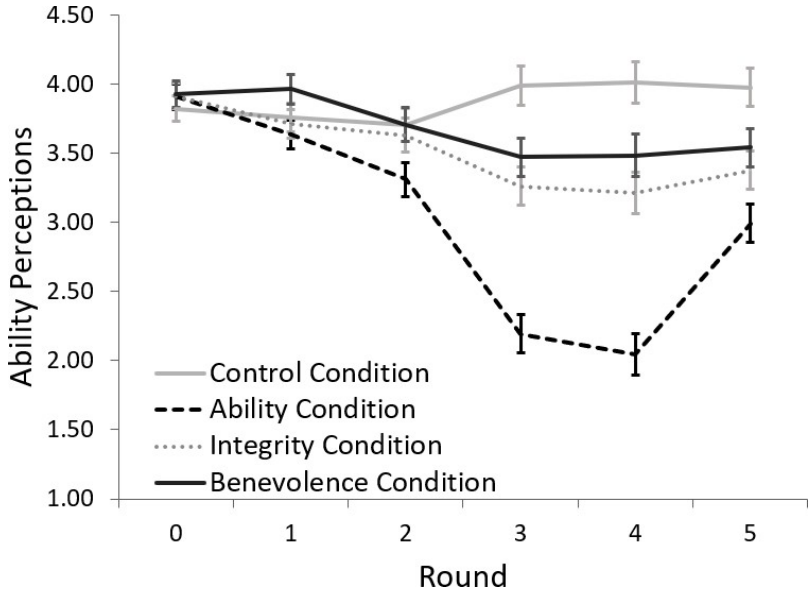


Figure 1. Ability Perceptions in Each Manipulated Condition Compared to the Control Condition.

Note. Ability Condition = Ability-based trust violation condition; Integrity Condition = Integrity-based trust violation condition; Benevolence Condition = Benevolence-based trust violation condition.

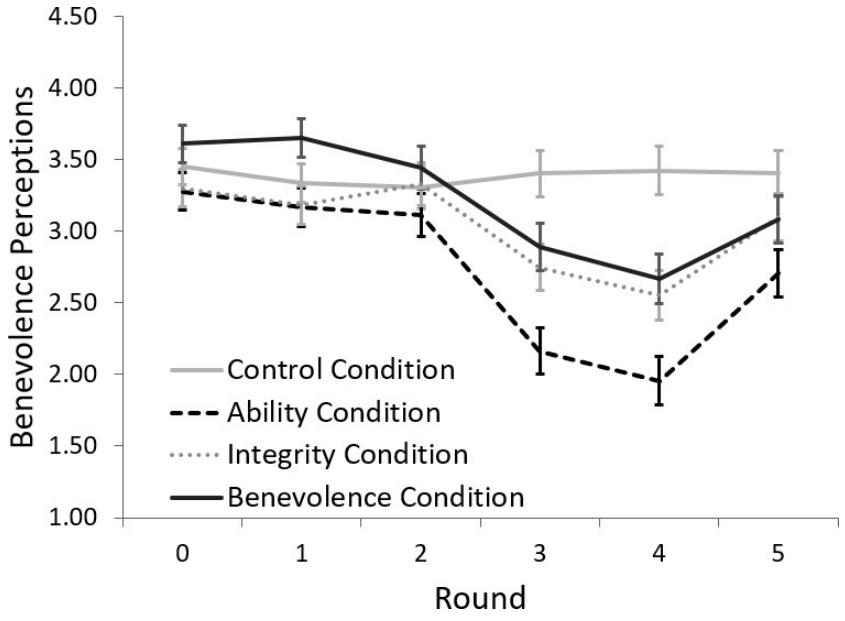
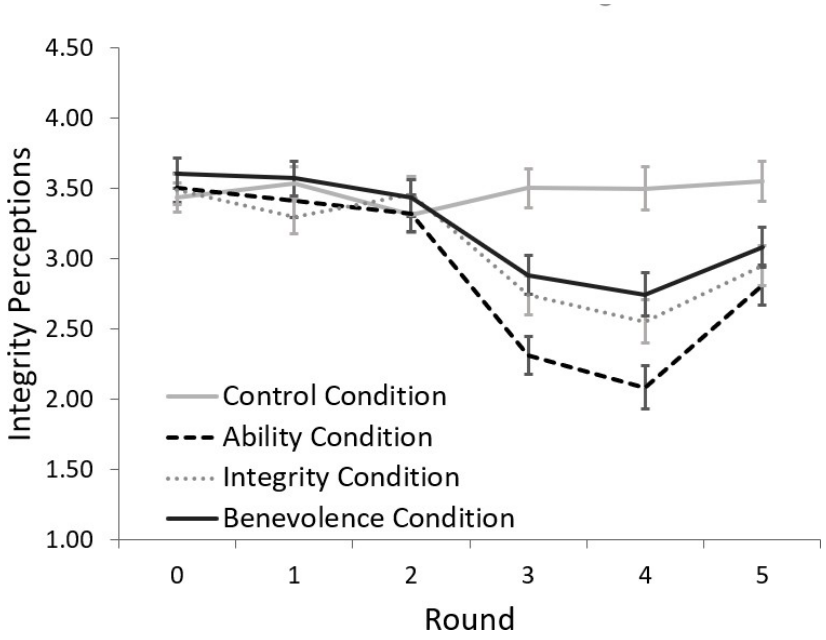


Figure 2. Benevolence Perceptions in Each Manipulated Condition Compared to the Control Condition.

Note. Ability Condition = Ability-based trust violation condition; Integrity Condition = Integrity-based trust violation condition; Benevolence Condition = Benevolence-based trust violation condition.



.

Figure 3. Integrity Perceptions in Each Manipulated Condition Compared to the Control Condition

Note. Ability Condition = Ability-based trust violation condition; Integrity Condition = Integrity-based trust violation condition; Benevolence Condition = Benevolence-based trust violation condition.

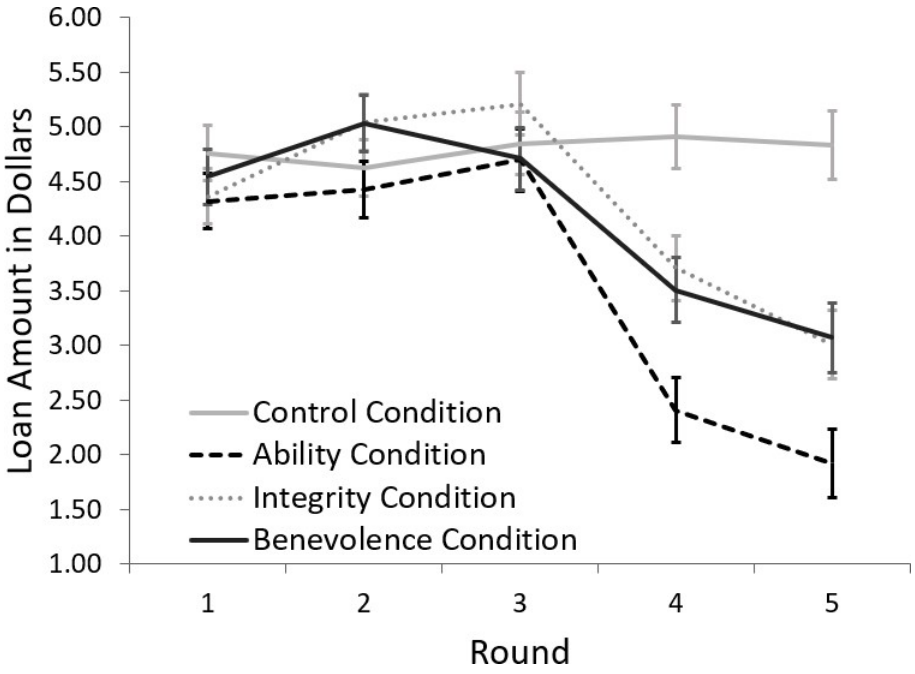


Figure 4. Risk-taking Behaviors in Each Manipulated Condition Compared to the Control Condition.

Note. Ability Condition = Ability-based trust violation condition; Integrity Condition = Integrity-based trust violation condition; Benevolence Condition = Benevolence-based trust violation condition.

Next, in Study 3, we tested our manipulations of ability, benevolence, and integrity between human-human teaming and human-robot teaming on an in-person sample. We experimentally manipulated ability, benevolence, and integrity in both human-human and human-robot teams. Using repeated measures multivariate analysis of variance we found several interesting findings. First, there were no mean differences between any of the conditions (i.e., partner or manipulation type) on risk taking behaviors. All risk-taking behaviors declined significantly following a trust violation. In contrast, overall trustworthiness declined significantly more for the robot than for the human following an ability violation, as illustrated in Figure 5. Interestingly, when a benevolence violation occurred the robot partner also suffered a higher decrease in trustworthiness perceptions. When an integrity violation occurred, there were no differences in overall trustworthiness between the human-human team and the human-robot team. This latter result is similar to the finding in study 1, as they were basically the same manipulation.

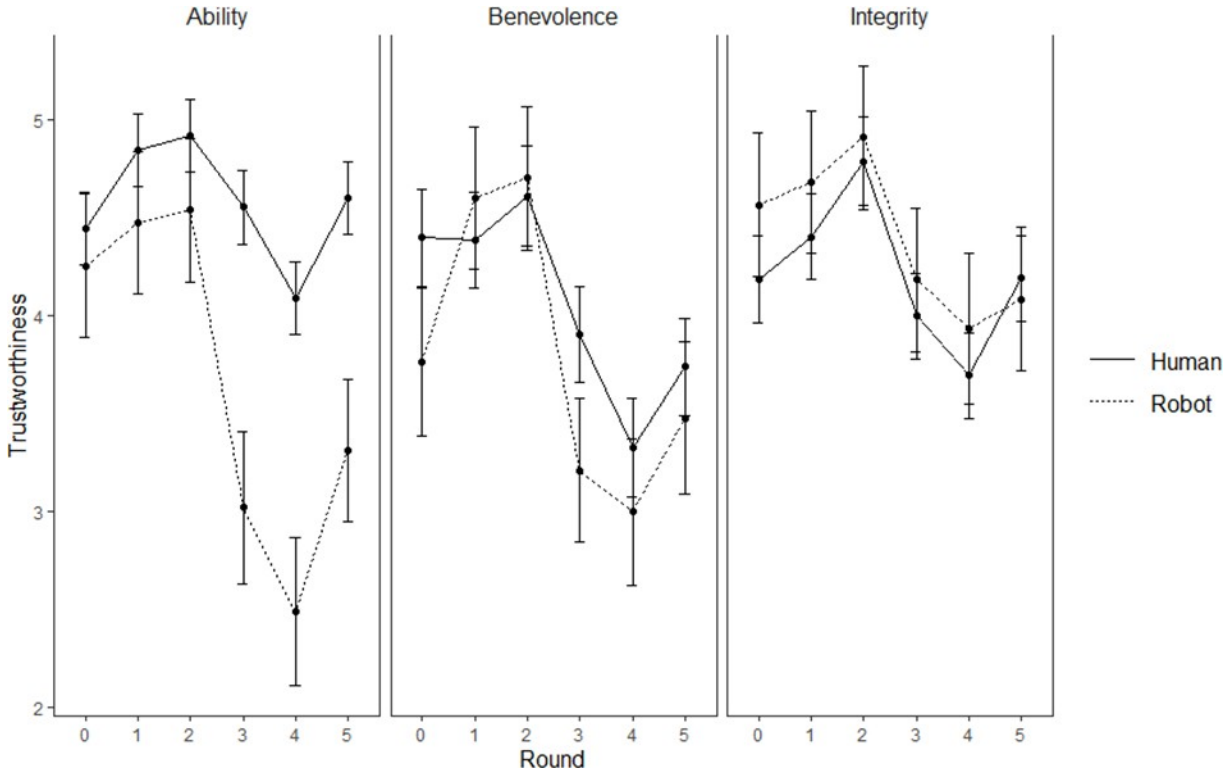


Figure 5. Overall Trustworthiness Perceptions in Each Manipulated Condition.
Note. Ability Condition = Ability-based trust violation condition; Integrity Condition = Integrity-based trust violation condition; Benevolence Condition = Benevolence-based trust violation condition.

In Study 4, we explored the role of PAS in the rate of change for each slope. In other words, does PAS predict the rate of change in trustworthiness perceptions after a trust violation? We hypothesized that trust violations committed by a robot would lead to steeper decreases in trust behaviors and trustworthiness perceptions for performance violations, compared to violations from a human partner. To assess the influence of PAS, we restructured the data so that the manipulations were dummy coded, so as to include them as first level variables in a regression

equation (see Tabachnick and Fidell, 2009). We then conducted latent growth modeling using mixed effects models. We treated the ability condition as the referent condition for analyses as the previous literature has focused on decrements of trust based on performance in the previous literature. We used the same manipulations as outline in Study 3. We found partial support for the unique agent hypothesis. As with Study 3 there were significant differences between human-human and human-robot teaming. Specifically, when modeling the rate of change the human partner actually suffered a larger decrease in overall trustworthiness perceptions than the robot partner did in the online study, as illustrated in Figure 6. Additionally there were no differences between the human and robot condition in the integrity condition. There was also a significant interaction of the benevolence manipulation and partner type, as illustrated in Figure 7.

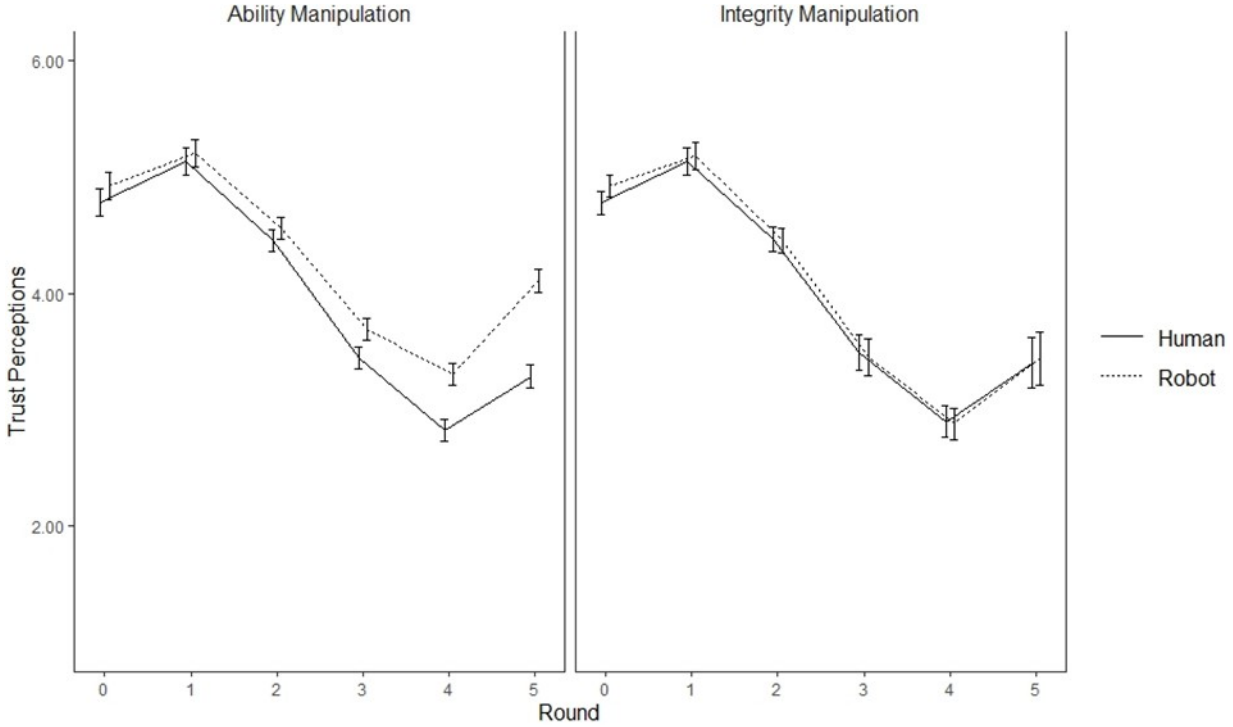


Figure 6. Three-way Interaction of Partner*Time*Manipulation for Latent Growth Model of Overall Trustworthiness Perceptions (Ability vs Integrity Conditions).

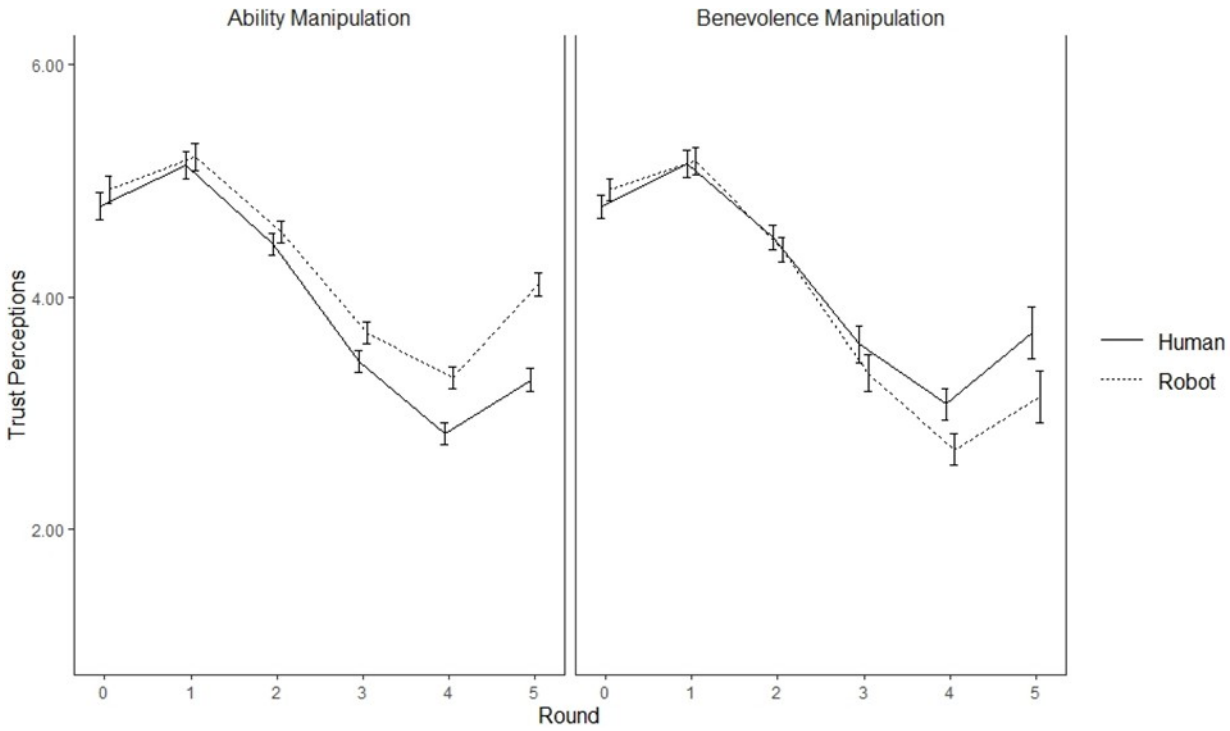


Figure 7. Three-way Interaction of Partner*Time*Manipulation for Latent Growth Model of Overall Trustworthiness Perceptions (Ability vs Benevolence Conditions).

We added the PAS scale to the equation to determine if PAS predict the rate of change in the slopes following a trust violation. We found the all-or-none thinking facet of PAS significantly predicted variance in the slope of benevolence perceptions, such that participants higher in all-or-none thinking viewed the robot as significantly less benevolent following a trust violation if they were higher in the construct. Figure 8 illustrates the results of PAS predicting the slope.

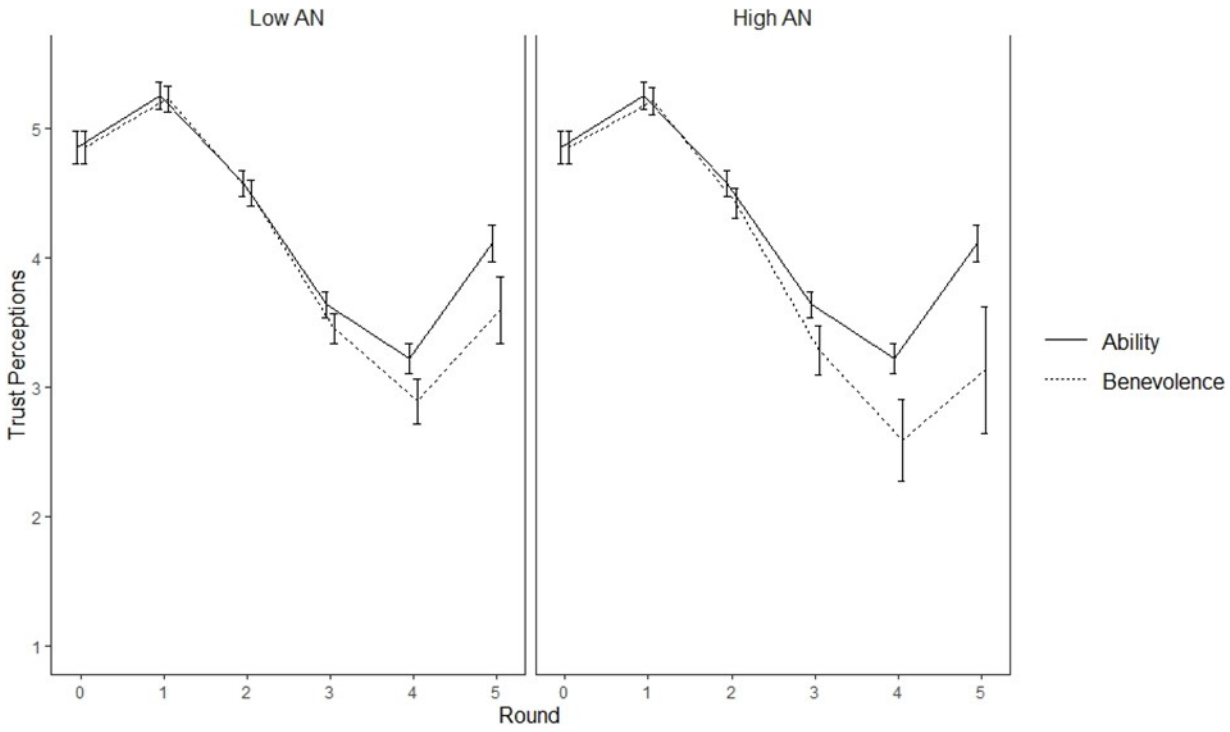


Figure 8. Three-way Interaction of Time*Manipulation*All-or-Nothing Thinking for Latent Growth Model of Overall Trustworthiness Perceptions

Note. Data is only plotted for Robot Condition.

Lastly, we collected data with three levels of anthropomorphism. We used the Nao Humanoid Robot, the Cue Robot which has some anthropomorphic features, and an arm robot with no anthropomorphic features. We collected data on ability, benevolence, and integrity conditions across the three robot types. We have not finished cleaning and analyzing the data. Future data analysis will include anthropomorphism as a factor across the ability, benevolence, and integrity conditions. We expect that robots that are higher in anthropomorphism will show less decrements in trust perceptions and trust behaviors than robots lower in anthropomorphism.

5.0 DISSEMINATION OF RESULTS

Results have been disseminated through conference proceedings and journal publications. The LRIR has resulted in one journal publication and five conference proceedings. There are currently three journal articles being prepped for publication. References are:

Alarcon, G. M., Gibson, A. M., Jessup, S. A., & Capiola, A. (in press). Exploring the Differential Effects of trust Violations in Human-Human and Human-Robot Interactions. *Applied Ergonomics*, *93*, Article 103350.

Capiola, A., Alarcon, G. M., Gibson, A. M., Jessup, S. A., & Hamdan, I. (in press). The Same or Different? Investigating whether Trust and Distrust are orthogonal Constructs or Span a Continuum. *Paper submitted to Hawaii International Conference on System Sciences.*

Alarcon, G. M., Capiola, A., Morgan, J., Hamdan, I., & Lee, M. (in press). Trust Violations in Human-Human and Human-Robot Interactions: The Influence of Ability, Benevolence and Integrity Violations. *Paper submitted to Hawaii International Conference on System Sciences.*

Gibson, A. M., Alarcon, G. M., Jessup, S. A., & Capiola, A. (January, 2020). Do You Still Trust Me? Effects of Personality on Changes in Trust During an Experimental Task with a Human or Robot Partner. *Proceedings of the Hawaii International Conference on System Sciences Annual Meeting*, Maui, HI.

Jessup, S. A., Gibson, A. M., Alarcon, G. M., & Capiola, A. (January, 2020). Investigating the Effect of Trust Manipulations on Affect over Time in Human-Human versus Human-Robot Interactions. *Proceedings of the Hawaii International Conference on System Sciences Annual Meeting*, Maui, HI.

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019, July). The Measurement of the Propensity to Trust Automation. In *International Conference on Human-Computer Interaction* (pp. 476-489). Springer, Cham.

6.0 SCIENCE, TECHNOLOGY, ENGINEERING AND MATHEMATICS (STEM)-RELATED ACTIVITIES

There were several STEM-related activities associated with the project. First, Ms. Sarah Jessup utilized data from the current project to complete her Master's Thesis. This thesis was later presented at the Human-Computer Interaction Conference and published in the proceedings. Additionally, Ms. Jessup utilized data from the current project to present two additional papers at Human-Computer Interaction and Hawaii International Conference on System Sciences conferences. Dr. Anthony Gibson utilized the data to publish a paper regarding personality and the changes of the experimental manipulations on trust and suspicion while serving as a post-doctoral student in the Consortium Research Fellows Program (Consortium of Universities of the Washington Metropolitan Area).

7.0 IMPACTS

In Study 1, 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 (Nass & Moon, 2000), 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 Alarcon et al., 2021). 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. Importantly, Study 1 did not have any ability or benevolence violations. As such, we implemented those in Study 2.

Study 2 expanded the understanding of trust games and their role in the psychological literature. As mentioned above, the trust literature has faced a dearth of studies that are able to separate and manipulate all aspects of the proposed trustworthiness model by Mayer et al. (1995), and to a degree McAllister (1995). With respect to economic game theoretical investigations on trust utilizing extant trust games (e.g., prisoner's dilemma, trust game), we have noticed that it is difficult to classify a trustee's violation of trust as an integrity-based or benevolence-based violation. Additionally, the most notable aspect of trustworthiness that has been missing from trust games is ability (see Alarcon et al., 2021). In order to examine these concerns, we leveraged a trust game (Alarcon et al., 2018) that incorporates a performance component and ambiguity over the course of multiple trials and implemented novel manipulations of ability, integrity, and benevolence in three experiments to examine the effects of these theoretically relevant trust violations on perceptions of trustworthiness. We manipulated trust by instantiating ability, integrity, and benevolence trust violations in the distrust conditions and measured the influence of these manipulation on participants' perception of their partners' trustworthiness and the effect of these manipulations on participants' risk-taking behaviors in subsequent rounds of the game. Our work provides a starting point for future investigations to delve into the nuance of trustworthiness manipulations and test their effects on criterion of interest. Additionally, it supports the theoretical distinction between can do aspects of trustworthiness (ability) and will do aspects of trustworthiness (benevolence and integrity).

Study 3 explored the differential effects of ability, benevolence, and integrity violations onto trustworthiness perceptions between humans and robots. Results illustrated a clear bias against robots when a trust violation occurred. However, not all trust violations were viewed the same. The ability violation had the strongest influence on trustworthiness perceptions and risk-taking behavior. These results are supported by the literature on perfect automation schemas (PAS; Madhavan & Wiegman, 2007; Merritt et al., 2015). However, the vast majority of research on PAS has focused on performance degradations. Study 3 found the robot suffered greater declines in trustworthiness, compared to a human, following ability and benevolence trust violations. As such, these perfect automation schemas may influence more than performance perceptions.

Indeed, Study 4 found PAS predicted the change in ability and benevolence perceptions, but not integrity perceptions. The reason for this may be that trust is an information processing activity, rather than a rational choice. The integrity violations did not provide any information as to why the violation occurred, as such biases, or schemas, organize information and relationships among them; however if no information is salient, the schemas may not be activated. This explains why Study 1 and the integrity condition in Studies 3 and 4 demonstrated no differences between

humans and robots, as there was too little information to understand why the trust violation had occurred.

The current project has several implications for the Air Force. First, engineers will want to be cognizant of the information displayed to the user. It may not be possible to create displays or automation that do not have a performance aspect. Indeed, performance helps to close the feedback loop for the operator to understand how the automation/robot is working. However, too much information can unduly bias users against automation. Second, engineers and the research community have traditionally been focused on performance declinations in the automation/robot. However, the current study indicates benevolence information may also be of concern to the user. Engineers that can find a way to display benevolence of the system, even when the system errs can reduce the impact of the PAS on the perception of the err. As noted above, when participants viewed the violation as an ability violation their risk-taking behavior declined much more rapidly then when it was a benevolence violation. As such, researchers would do well to include benevolence as an important aspect of their systems when thinking about transparency.

8.0 CHANGES TO PROTOCOL

Changes were made to the protocol due to the Corona Virus Disease (COVID) 19 pandemic. The pandemic delayed the ability to collect in person data, as such much of the data was collected online using Amazon MTurk. We utilized Amazon MTurk to test manipulations for Ability, Benevolence and Integrity instead of performing the study in-person. Additionally, we utilized the online platform to collect data on any differences in the rate of change between conditions. The collection of data using an online platform necessitated a change in reimbursement to approximately half of what was proposed in the study. This was to avoid any undue influence by offering larger sums of money participation in an online experiment.

9.0 REFERENCES

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social Robots for Education: A Review. *Science Robotics*, *3*(21), Article eaat5954.

Bemelmans, R., Gelderblom, G. J., Jonker, P., & De Witte, L. (2012). Socially Assistive Robots in Elderly Care: A Systematic Review into Effects and Effectiveness. *Journal of the American Medical Directors Association*, *13*(2), 114-120.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, *10*(1), 122-142.

Broadbent, E., Kumar, V., Li, X., 3rd, J. S., Stafford, R. Q., MacDonald, B. A., & Wegner, D. M. (2013). Robots with Display Screens: A Robot with a more Human-Like Face Display is Perceived to have more Mind and a Better Personality. *PLoS ONE*, *8*(8), Article e72589.

Calhoun, C. S., Bobko, P., Gallimore, J. J., & Lyons, J. B. (2019). Linking Precursors of Interpersonal Trust to Human-Automation Trust: An Expanded Typology and Exploratory Experiment. *Journal of Trust Research, 9*(1), 28-46.

Chang, W. H., & Kim, Y.-H. (2013). Robot-Assisted Therapy in Stroke Rehabilitation. *Journal of Stroke*, *15*(3), 174-181.

Chen, J.Y.C., & Barnes, M.J. (2014). Human-Agent Teaming for Multi-Robot Control: A Review of the Human Factors Issues. *IEEE Transactions on Human-Machine Systems, 44*, 13-29.

Chiou, E. K., & Lee, J. D. (2021). Trusting Automation: Designing for Responsivity and Resilience. *Human Factors*, 00187208211009995.

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), 909-927.

Colquitt, J. A., & Rodell, J. B. (2011). Justice, Trust, and Trustworthiness: A Longitudinal Analysis Integrating Three Theoretical Perspectives. *Academy of Management Journal*, *54*(6), 1183-1206.

Dautenhahn, K. (2002). Design Spaces and Niche Spaces of Believable Social Robots. *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication, Germany,* 192-197.

Dautenhahn, K., Walters, M., Woods, S., Koay, K. L., Nehaniv, C. L., Sisbot, A., ... & Siméon, T. (2006). How May I Serve You? A Robot Companion Approaching a Seated Person in a Helping Context. *Proceedings of the ACM SIGCHI/SIGART Conference on Human-Robot Interaction, USA,* 172-179.

Davies, B., Starkie, S., Harris, S. J., Agterhuis, E., Paul, V., & Auer, L. M. (2000). Neurobot: A Special-Purpose Robot for Neurosurgery. *Proceedings of the International Conference on Robotics and Automation*, *USA*, *4*, 4103-4108.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents. *Journal of Experimental Psychology*: *Applied*, *22*(3), 331-49.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The Role of Trust in Automation Reliance. *International Journal of Human-Computer Studies, 58*(6), 697-718.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors*, *44*(1), 79-94.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting Misuse and Disuse of combat Identification Systems. *Military Psychology*, *13*(3), 147-164.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review*, *114*(4), 864-886. https://doi.org/10.1037/0033-295X.114.4.864

Friedman, B., Kahn Jr, P. H., & Hagman, J. (2003). Hardware Companions? What Online AIBO Discussion Forums Reveal about the Human-Robotic Relationship. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, USA,* 273-280.

Guizzo, E. (2020, May 28) *What is a Robot*? Robots IEEE. https://robots.ieee.org/learn/what-is-a-robot/

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367-388.

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.

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2020). Evolving Trust in Robots: Specification through Sequential and Comparative Meta-Analyses. *Human Factors*, Advanced Online Publication.

Hertz, N., & Wiese, E. (2017). Social Facilitation with Non-Human Agents: Possible or Not? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *USA*, *61*(1), 222-225.

Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose Job is it Anyway? A Study of Human-Robot Interaction in a Collaborative Task. *Human-Computer Interaction*, *19*(1-2), 151-181.

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors that Influence Trust. *Human Factors*, *57*(3), 407-434.

Kahn, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., Gary, H. E., Reichert, A. L., Freier, N. G., & Severson, R. L. (2012). Do People Hold a Humanoid Robot Morally Accountable for the Harm it Causes? *Proceedings of the International Conference on Human-Robot Interaction*, *USA*, 33-40. https://doi.org/10.1145/2157689.2157696

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. *PLoS ONE*, *3*(7), Article e2597.

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50-80.

Lee, J. D., & Seppelt, B. D. (2009). Human Factors in Automation Design. In S. Y. Nof (Ed.), *Springer Handbook of Automation* (pp. 417-436). Springer.

Lee, K. M., Peng, W., Jin, S. A., & Yan, C. (2006). Can Robots Manifest Personality? An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction. *Journal of Communication, 56*(4), 754-772.

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., & Paiva, A. (2013). The Influence of Empathy in Human–Robot Relations. *International Journal of Human-Computer Studies*, *71*(3), 250-260.

Lyons, J. B. (2013, March). Being Transparent about Transparency: A Model for Human-Robot Interaction. In *2013 AAAI Spring Symposium Series*.

Lyons, J. B., & Grigsby, M. A. (2016). Acceptance of Advanced Autonomous Systems: A Call for Research. *Military Psychology*, *31*, 11-15.

Lyons, J. B., Ho, N. T., Koltai, K. S., Masequesmay, G., Skoog, M., Cacanindin, A., & Johnson, W. W. (2016). Trust-Based Analysis of an Air Force Collision Avoidance System. *Ergonomics in Design*, *24*(1), 9-12.

Lyons, J. B., Koltai, K. S., Ho, N. T., Johnson, W. B., Smith, D. E., & Shively, R. J. (2016). Engineering Trust in Complex Automated Systems. *Ergonomics in Design*, *24*(1), 13-17.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and Differences between Human-Human and Human-Automation Trust: An Integrative Review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277-301.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *Academy of Management Review*, *20*(3), 709-734. https://doi.org/10.2307/258792

McAllister, D. J. (1995). Affect-and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *Academy of Management Journal*, *38*(1), 24-59.

Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management. *Human Factors*, *58*(3), 401-415.

Merritt, S.M., Unnerstall, J.L., Lee, D., & Huber, K. (2015). Measuring Individual Differences in the Perfect Automation Schema. *Human Factors*, 57, 740-753.

Mota, R. C. R., Rea, D. J., Le Tran, A., Young, J. E., Sharlin, E., & Sousa, M. C. (2016). Playing the 'Trust Game' with Robots: Social Strategies and Experiences. *Proceedings of the International Symposium on Robot and Human Interactive Communication* (RO-MAN), *USA*, 519-524.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems. *Ergonomics*, *37*(11), 1905-1922.

Nass, C., & Lee, K. M. (2001). Does Computer-Synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity-Attraction, and Consistency-Attraction. *Journal of Experimental Psychology*: *Applied*, *7*(3), 171-181.

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81-103.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are Social Actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, *USA*, 72-78.

Ososky, S., Schuster, D., Phillips, E., & Jentsch, F. G. (2013, March). Building Appropriate Trust in Human-Robot Teams. In *2013 AAAI Spring Symposium Series*.

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, *39*(2), 230-253.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance Consequences of Automation-Induced'complacency'. *The International Journal of Aviation Psychology*, *3*(1), 1-23.

Pennisi, P., Tonacci, A., Tartarisco, G., Billeci, L., Ruta, L., Gangemi, S., & Pioggia, G. (2016). Autism and Social Robotics: A Systematic Review. *Autism Research*, *9*(2), 165-183.

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual Differences in the Calibration of Trust in Automation. *Human Factors*, *57*(4), 545-556.

Powers, A., Kramer, A., Lim, S., Kuo, J., Lee, S. L., & Kiesler, S. (2005). Common Ground in Dialogue with a Gendered Humanoid Robot. *Proceedings of RO-MAN 2005*.

Rotter, J. B. (1980). Interpersonal Trust, Trustworthiness, and Gullibility. *American Psychologist*, *35*(1), 1-7.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would You Trust a (Faulty) robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. *Proceedings of the International Conference on Human-Robot Interaction*, *USA*, 141-148.

Sanders, T., Kaplan, A., Koch, R., Schwartz, M., & Hancock, P. A. (2019). The Relationship Between Trust and Use Choice in Human-Robot Interaction. *Human Factors*, *61*(4), 614-626.

Schäfer, M. B., Stewart, K. W., & Pott, P. P. (2019). Industrial Robots for Teleoperated Surgery-A Systematic Review of Existing Approaches. *Current Directions in Biomedical Engineering*, *5*(1), 153-156.

Smith, S. J., Stone, B. T., Ranatunga, T., Nel, K., Ramsoy, T. Z., & Berka, C. (2017). Neurophysiological Indices of Human Social Interactions between Humans and Robots. In C. Stephanidis (Ed.), *Extended abstract in communications in computer and information science*: *Vol. 713. International conference on human-computer interaction* (pp. 251-262). Springer.

Smith, M. A., & Wiese, E. (2016). Look at Me Now: Investigating Delayed Disengagement for Ambiguous Human-Robot Stimuli. In A. Agah, J. J.Cabibihan, A. Howard, M. Salichs, & H. He (Eds.), *Lecture Notes in Computer Science*: *Vol. 9979*. *Social robotics* (pp. 950-960). Springer.

Somon, B., Campagne, A., Delorme, A., & Berberian, B. (2019). Human or Not Human? Performance Monitoring ERPs during Human Agent and Machine Supervision. *Neuroimage*, *186*, 266-277.

Tulk, S., & Wiese, E. (2018). Trust and Approachability Mediate Social Decision Making in Human-Robot Interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *USA*, *62*(1), 704-708.

Wang, Y., & Quadflieg, S. (2015). In our Own Image? Emotional and Neural Processing Differences when Observing Human-Human vs Human-Robot Interactions. *Social Cognitive and Affective Neuroscience*, *10*(11), 1515-1524.

Wynne, K. T., & Lyons, J. B. (2018). An Integrative Model of Autonomous Agent Teammate-Likeness. *Theoretical Issues in Ergonomics Science*, *19*(3), 353-374.

Xie, Y., Bodala, I. P., Ong, D. C., Hsu, D., & Soh, H. (2019). Robot Capability and Intention in Trust-Based Decisions across Tasks. *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, *Korea*, 39-47.

Young, J. E., Hawkins, R., Sharlin, E., & Igarashi, T. (2009). Toward Acceptable Domestic Robots: Applying Insights from Social Psychology. *International Journal of Social Robotics*, *1*, 95-108.

10.0 LIST OF ACRONYMS, ABBREVIATIONS AND SYMBOLS

CASA	Computers as Social Actors
COVID	Corona Virus Disease
DSS	Decision Support Systems
HMT	Human-Machine Teams
HRI	Human-Robot Interaction
IEEE	Institute of Electrical and Electronics Engineers
LRIR	Laboratory Research Initiation Request
MTurk	Amazon Mechanical Turk
PAS	Perfect Automation Schema
STEM	Science, Technology, Engineering and Mathematics
UAV	Unmanned Aerial Vehicle