2018-06

INTEGRITY-BASED TRUST VIOLATIONS WITHIN HUMAN-MACHINE TEAMING

Clark, Tiffany

Monterey, CA; Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

INTEGRITY-BASED TRUST VIOLATIONS WITHIN HUMAN-MACHINE TEAMING

by

Tiffany Clark

June 2018

Thesis Advisor: Brian S. Bingham
Co-Advisor: Mollie R. McGuire

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 2018	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
INTEGRITY-BASED TRUST VIOLATIONS WITHIN HUMAN-MACHINE TEAMING

5. FUNDING NUMBERS

6. AUTHOR(S) Tiffany Clark

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT (maximum 200 words)

Successful human-machine teaming requires humans to trust machines. While many claim to welcome automation, there is also mistrust of machines which may stem from more than competence concerns. Human-automation trust research to date has considered automation capable of competence-based trust violations (CBTV), but integrity-based trust violations (IBTV) should also be studied. Future advances in artificial intelligence and cyber warfare could result in the perception—and possible reality—of automation committing IBTVs. The current study paired human participants with an automated teammate to complete a sequence of computer-based visual search and investment tasks. During each session, the automation committed either an IBTV or CBTV, and participants' trust responses were measured through self-reported trust, trust-based reliance behavior, time spent making reliance decisions, and investment behavior. The results found that (a) average self-reported trust in the automation was significantly lower in the IBTV than the CBTV condition, (b) personal investment behavior was more consistent with reported trust levels than reliance behavior and may be a better gauge of trust, and (c) trust behavior differed more between IBTV and CBTV conditions among participants who invested more in their automated teammate. The differences found in participant trust response between conditions are enough to warrant further research into how humans react to automation committing IBTVs.

14. SUBJECT TERMS
human-machine teaming, human-systems interaction, human-robot teaming, man-machine interface, human-computer interaction, human-robot interaction, trust in automation, human-automation trust, integrity-based trust violation, competence-based trust violation

15. NUMBER OF PAGES
85

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

INTEGRITY-BASED TRUST VIOLATIONS WITHIN HUMAN-MACHINE TEAMING

Tiffany Clark
Lieutenant, United States Navy
BS, University of Washington, 2009

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN MECHANICAL ENGINEERING

from the

NAVAL POSTGRADUATE SCHOOL
June 2018

Approved by: Brian S. Bingham
Advisor

Mollie R. McGuire
Co-Advisor

Garth V. Hobson
Chair, Department of Mechanical and Aerospace Engineering

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Successful human-machine teaming requires humans to trust machines. While many claim to welcome automation, there is also mistrust of machines which may stem from more than competence concerns. Human-automation trust research to date has considered automation capable of competence-based trust violations (CBTV), but integrity-based trust violations (IBTV) should also be studied. Future advances in artificial intelligence and cyber warfare could result in the perception—and possible reality—of automation committing IBTVs. The current study paired human participants with an automated teammate to complete a sequence of computer-based visual search and investment tasks. During each session, the automation committed either an IBTV or CBTV, and participants' trust responses were measured through self-reported trust, trust-based reliance behavior, time spent making reliance decisions, and investment behavior. The results found that (a) average self-reported trust in the automation was significantly lower in the IBTV than the CBTV condition, (b) personal investment behavior was more consistent with reported trust levels than reliance behavior and may be a better gauge of trust, and (c) trust behavior differed more between IBTV and CBTV conditions among participants who invested more in their automated teammate. The differences found in participant trust response between conditions are enough to warrant further research into how humans react to automation committing IBTVs.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AI	Artificial Intelligence
ANDI	Automated Networked Decision Infrastructure
ANOVA	Analysis of Variance
ARSENL	Advanced Robotic Systems Engineering Laboratory
ATM	Automated Teller Machine
BRIAN	Battlefield Remote Intelligent Automated Network
CBE	Competence-Based Error
CBTV	Competence-Based Trust Violation
CI_{95}	95 percent Confidence Interval
DOD	Department of Defense
GPS	Global Positioning System
HMT	Human-Machine Teaming
HSI	Human-Systems Integration
IBM	International Business Machines
IBTV	Integrity-Based Trust Violation
IBV	Integrity-Based Violation
M	Mean
NIH	National Institute of Health
NPS	Naval Postgraduate School
NSAM	Naval Support Activity Monterey
Participant ID	Participant Identification Number
PANAS	Positive and Negative Affect Scale
SD	Standard Deviation
SPSS	Statistical Package for the Social Sciences
TIG	Team Investment Game
U.S.	United States
UAV	Unmanned Aerial Vehicle
VST	Visual Search Task

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. BACKGROUND

Human-machine teaming (HMT) has become a buzzword within the U.S. military, and for good reason. The U.S. Third Offset Strategy focuses on five key technological areas: "autonomous learning systems, human-machine collaborative decision-making, assisted human operations, advanced manned-unmanned systems operations, and network-enabled autonomous weapons and high-speed projectiles" [1]. The Department of Defense's (DOD) Unmanned Systems Roadmap for 2013-2038 [2] points out an ongoing transition within automation research, from developing *automatic* systems which still need human control to developing *autonomous* systems that react and make decisions with no human control. Stated another way, the Roadmap shows that the U.S. military is moving from all-human teams in full control of their respective machines to humans teaming with self-directed autonomous systems which choose behaviors to follow in pursuit of a human-directed goal. Finally, the 2016 Defense Science Board Summer Study on Autonomy concluded that "DoD must take immediate action to accelerate its exploitation of autonomy" [3].

The plans are clear: autonomous machines are coming, and human-machine teams will be a staple of future U.S. military operations. The research is in progress and the machines are ever closer to being ready. But how ready are the soldiers, sailors, and airmen who will be the "human" part of these human-machine teams?

Humans may not be as ready for HMT as many think. A 2016 survey of 2,000 people in New York found that while 81 percent of people are excited about an automated life, 73 percent say they are scared to trust machines [4]. What exactly are people afraid of when it comes to trusting machines? One could guess they fear machines failing to perform their specified operation correctly—a matter of competence. Yet a definite majority of people in the United States already entrust their lives, and often also the lives of those they love, to the proper operation of motor vehicles, elevators, and many other machines that could be considered mortally dangerous. If placing one's life in

another's hands is considered an act of trust at the highest level, then people who use motor vehicles and elevators must trust the competence of machines to some degree. Why then would they say they are scared to trust machines?

This apparent contradiction could be explained by the presence of another facet of trust—one based on integrity or benevolence rather than competence. Given the choice of undergoing surgery by one of two doctors with equal skill level, what other factors would influence the decision? If one doctor was known for charitable weekend work and the other had been convicted of tax fraud, would that sway the choice? If competence were all that mattered, perhaps not. Yet in many situations, competence is not the only factor behind a decision to trust. The next question, then, is whether this same type of analogy can be applied to the machines people say they are afraid to trust. Can one weigh the integrity or benevolence of machines as a factor in decisions to trust, or does that only apply to humans? What about weighing the integrity of humans behind the machines, such as operators or programmers? Does that have the same effect as thinking that machines themselves are capable of having or displaying integrity?

Human-machine teams will not function well if humans do not appropriately trust or rely on the machines, and clearly there is much still to be learned about the complex dynamics of trust between humans and automation in pursuit of that appropriate trust. Specifically, it seems there is little known about human responses to trust violations through a failure of automation integrity or benevolence as opposed to failures through automation competence. The current study attempts to discover if these human responses differ by assessing human trust in automation after either a competence-based or integrity-based trust violation.

B. TERMS AND DEFINITIONS

A short introduction to some of the oft-used terms within the Trust in Automation field of study will help ensure maximum clarity in the discussions to come.

1. Automation

Commonly referenced as either degrees or levels of automation, Sheridan and Verplank [5] developed a specific scale varying actions between a human and an undersea teleoperator, ranging from level one (human performs decision-making and starts the automation on a pre-determined task), to level nine (automation decides everything about the task, implements action, and does not tell the human what it did). Figure 1 shows the full scale as presented by Sheridan and Verplank, and while the scale specifically mentions a computer, the computer's role is generalized to mean any automation for this paper.

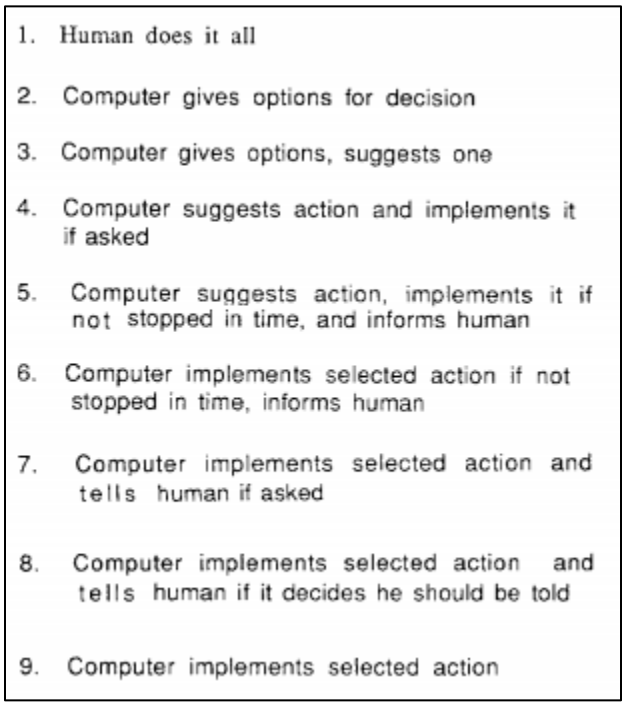

Figure 1. Degrees (or Levels) of Automation. Source: [5].

In terms of Sheridan and Verplank's scale, "automation" and "machine" as discussed in the current study would include level two or above.

A more recent and more general scale developed by Endsley and Kiris [6] (Figure 2) includes only five levels of control, but still shows increasing automation control with increasing level number.

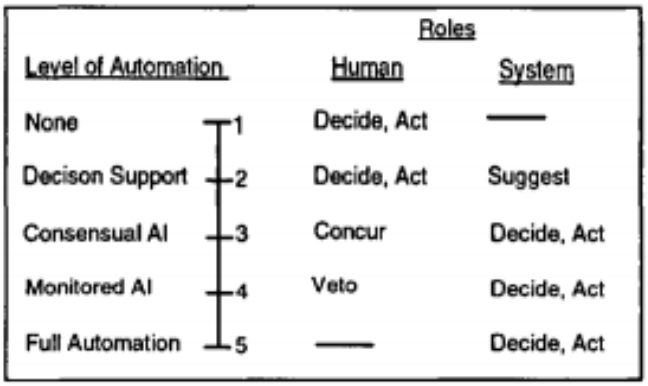

Figure 2. Levels of Control in Automation. Source: [6].

The terms “automation” and “machine” as used in this thesis would include level two or above within Endsley and Kiris’s scale. While there is a recognized difference between automatic and autonomous machines, this study is intended to be broad in including everything from low-level, automatic functions (such as automated financial payments) up to high-level, fully autonomous systems, including those with Artificial Intelligence (AI) software (such as self-driving cars).

2. Trust in Automation

It may be a small word, but “trust” can cover a large array of different meanings, layers, and uses. For this paper, trust is defined as Lee and See [7] defined it: “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability.”

Although it is often portrayed as a state, the act of trusting is a dynamic process, and some authors, such as Hoffman [8], choose to use the word “trusting” rather than “trust”. Not only is trust changing over time, Hoffman also points out that there are myriad different types of trusting that appear in different situations and for different

individuals. As Hoff and Bashir [9] modeled, trust can be seen as a three-layered construct, consisting of variable influences from an individual's Dispositional Trust, Situational Trust, and Learned Trust (Figure 3).

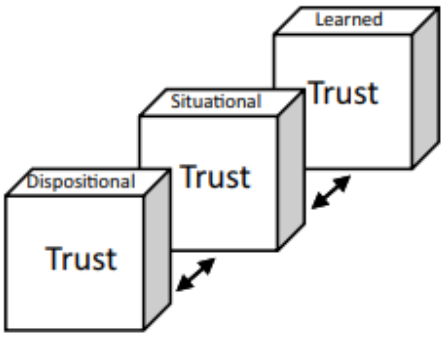

Figure 3. Three-Layered Framework for Trust Variability. Source: [9].

Hoff and Bashir [9] explain that each of these three trust layers also has its own influences, which can change over time and are what helps make trusting a dynamic process. The first layer these researchers explain is Dispositional Trust, which is considered to be more enduring than the other two and is influenced mainly by things such as culture, age, gender, and individual personality traits (Figure 4).

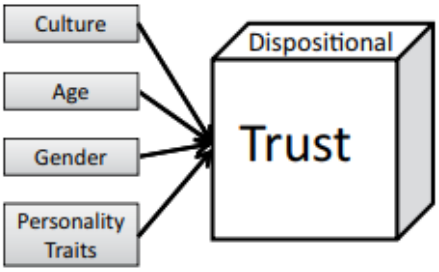

Figure 4. Factors Influencing Dispositional Trust. Source [9].

Several dispositional trust factors were measured for participants in this study, including age, gender, overall comfort with automation, and the tendency to trust swiftly without specific reason. With high participation levels, these factors were expected to

simply model a natural variation of dispositional trust within the human population. If anomalies or trends were found during analyses, an imbalance in dispositional trust factors within the study population could help explain the source.

Situational Trust as put forth by Hoff and Bashir [9] is more dynamic between different situations and is influenced by factors both internal and external to an individual (Figure 5). One important aspect the researchers noted about situational trust is that aside from influencing trust, these factors also influence the degree of correlation between an individual's trust (as an attitude) and his or her subsequent behavior (such as reliance, discussed later).

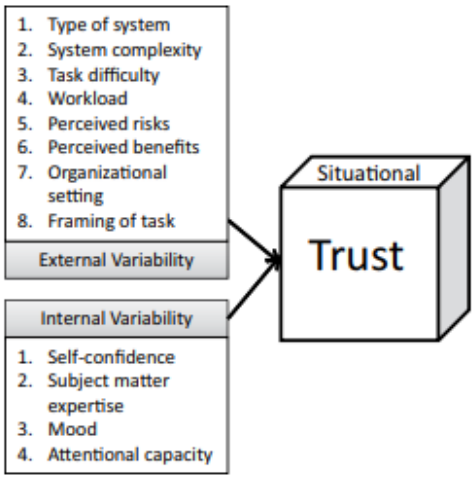

Figure 5. Factors Influencing Situational Trust. Source [9].

Internal variability factors of Situational Trust measured for participants in this study include self-efficacy, gaming tendencies, and mood both before and after the main experiment exercise. External variability factors such as task difficulty and perceived risks or benefits were carefully considered in the design of the experiment (explained further in Chapter III), but these were static within the experiment, and should have varied only with regard to each participant's individual abilities or perceptions.

Learned Trust, the last of Hoff and Bashir's three levels [9], represents a culmination of both long-term, past experiences (called *initial learned trust*) and more

immediate, current experiences with the automation in use (called *dynamic learned trust*). Figure 6 shows how initial learned trust influences an individual's initial reliance strategy, but dynamic learned trust tends to exert more influence as an individual interacts with automation, according to the researchers. They also explain that dashed arrows in the figure represent the ability of those factors to change with each interaction.

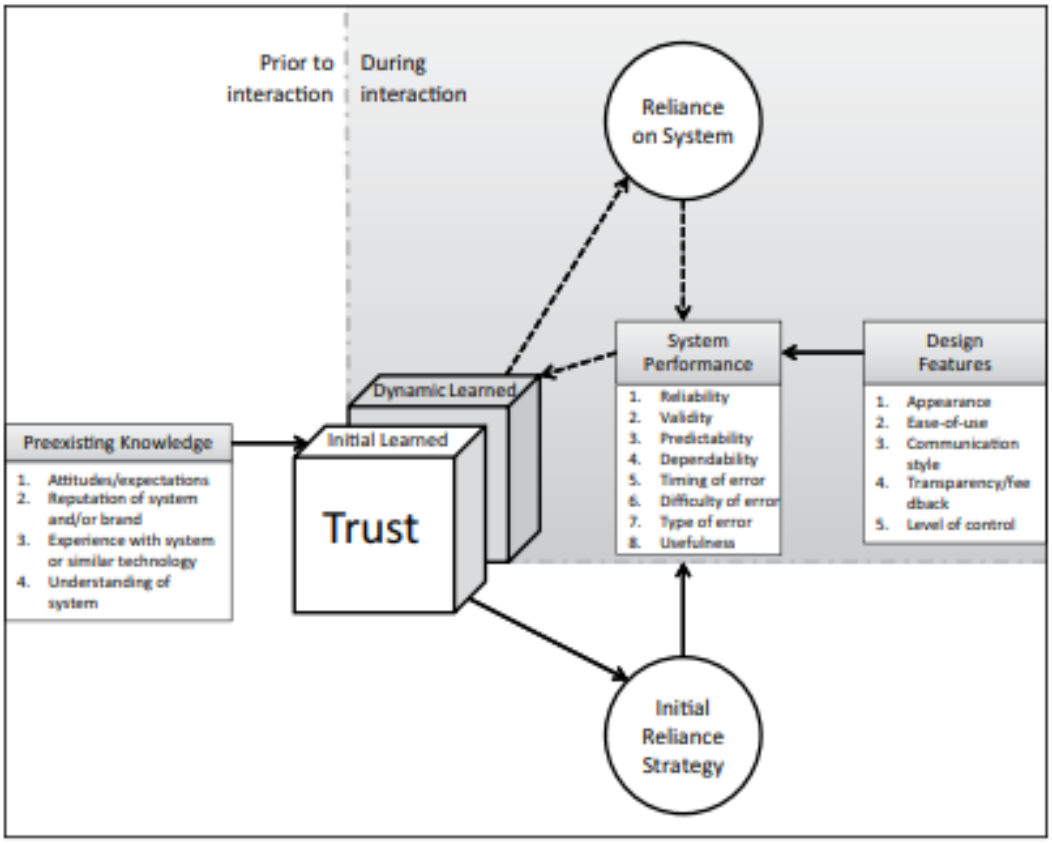

Figure 6. Factors Influencing Learned Trust. Source: [9].

The trust violations committed by the automation in this experiment would fall into the category of dynamic learned trust, specifically as part of system performance. Since this affects trust levels with each interaction, reliance behavior and response time measurements (both discussed in the Measurements section of Chapter III) were taken in an attempt to gauge participants' trust in an iterative fashion. Performance of the automation is held constant in the study, as are transparency, feedback, and other system

metrics which can affect trust. The type of trust violation experienced is the only manipulated variable in this experiment, as the intent is to study how this changes human perception of the automation and subsequent reliance behaviors as indicators of trust.

3. Trust Violations

Lee and Moray [10] categorized the many facets of trust in automation by performance measures (e.g., competence, reliability), process measures (e.g., integrity, predictability, dependability), and perceived purpose (e.g., benevolence, loyalty, faith). The purpose these researchers discuss is based on the perceived benevolence or loyalty of the humans who designed the automation rather than pertaining to the automation itself.

This thesis compares human responses to violations of trust with regard to Lee and Moray's categories [10], except that the three categories are split between two different types of trust violation. A competence-based trust violation (CBTV) represents a failure of either performance or process, and an integrity-based trust violation (IBTV) represents a failure of purpose.

While the word "integrity" included in the process category would seem to indicate that process failures should be under IBTVs, Lee and Moray [10] use the word to mean that the automation is acting as expected (predictability), which for this study is more closely related to competence. When used in general conversation regarding humans, the word "integrity" often signifies a certain level of moral or ethical considerations, and that is the way it is used in this paper. With this in mind, IBTVs in this study are more in line with a failure of perceived purpose than a failure of process.

An example of a CBTV between humans would be if one person is asked to hold money for another, but accidentally loses it. This person's reliability and dependability would certainly be called into question—a failure of performance and process. If instead, the person simply decided not to return the money, this person's benevolence and loyalty would be questioned—a failure of purpose, which could be called an IBTV. Similar examples using automation can be described in terms of an Automated Teller Machine (ATM). If one machine displayed an error message and failed to dispense money upon request, this could be called a CBTV. If another machine just next to the first also failed

to dispense money, but did not display an error message and also deducted the amount from your bank account, this could be perceived as an IBTV. While these could both be called a failure of performance or process, the second scenario has an element of perceived purpose that is also called into question.

Regardless of how trust violations are defined, each violation will ultimately be categorized by how the human in receipt of the violation *perceives* it. One real-life example of this is when Amazon's Alexa recorded a woman's private conversation from an Echo in her home and then emailed the audio file of it to one of the woman's contacts [11]. This can easily be explained as a failure of performance in which Alexa mistook conversation as a series of commands, but it is also easy to see how the woman in this story could instead perceive it as a failure of purpose, and subsequently lose trust based on that perception.

4. Reliance

As an attitude, trust itself is difficult to measure aside from subjective ratings. Often more objective measures, such as reliance, are taken to reflect an assumed level of trust [12]. Reliance for this thesis is defined as an objective, measurable behavior in which a human chooses to submit an answer given by the automation over their own initial answer. An individual's choice to rely on automation is likely not based solely on their trust of that automation. Or, as pointed out by Lee and See [7], "Trust guides—but does not completely determine—reliance." Although reliance and trust are not synonymous, van Dongen and van Maanen [12] point out that reliance often reflects the relative difference between trust in one's own ability and trust in an aid. In this way, reliance—an objectively measurable metric—can be seen as an indicator of the more subjective level of trust.

5. Appropriate Reliance and Calibrated Trust

Regardless of whether a team is fully human or part automation, team performance is best when there is "appropriate" reliance, which stems from properly calibrated trust as put forth by Lee and See [7]. Trust is considered properly calibrated according to Muir [13] when a person's level of trust in automation matches well with the

automation's actual capabilities. In practice, once trust is properly calibrated, a person would trust the automation when it is trustworthy, and would not trust the automation when it is untrustworthy.

Appropriate reliance within a human-machine team would occur when a human chose to rely on automation if reliance would benefit team performance, and chose not to rely on the automation if reliance would detriment team performance. Because situations often prevent humans from knowing whether reliance will benefit or detriment the team until after the reliance decision is made, properly calibrated trust is needed to effectively guide these decisions. Appropriate reliance stands in contrast to either misuse (over-reliance; relying when automation performs poorly) or disuse (under-reliance; not relying when automation performs well) as defined by Parasuraman and Riley [14].

II. TOPIC DISCUSSION

Beginning with a review of pertinent trust-in-automation literature, this chapter highlights a small gap in research and discusses why the topic deserves more attention.

A. HUMAN-SYSTEMS INTEGRATION

Human-machine teaming is a subject which not only considers how humans and machines operate as separate entities, but also considers details of the relationship and interactions that occur between these entities in a team environment. A field of study at the Naval Postgraduate School (NPS) which specializes in this type of relationship is Human-Systems Integration (HSI). One HSI model, originally published in an NPS thesis by U.S. Coast Guard LCDR Mike O'Neil [15], is helpful in visualizing the integration of humans and technology into a complete system. This model has been further developed by HSI faculty and students at NPS [16], and is shown in its adapted state in Figure 7.

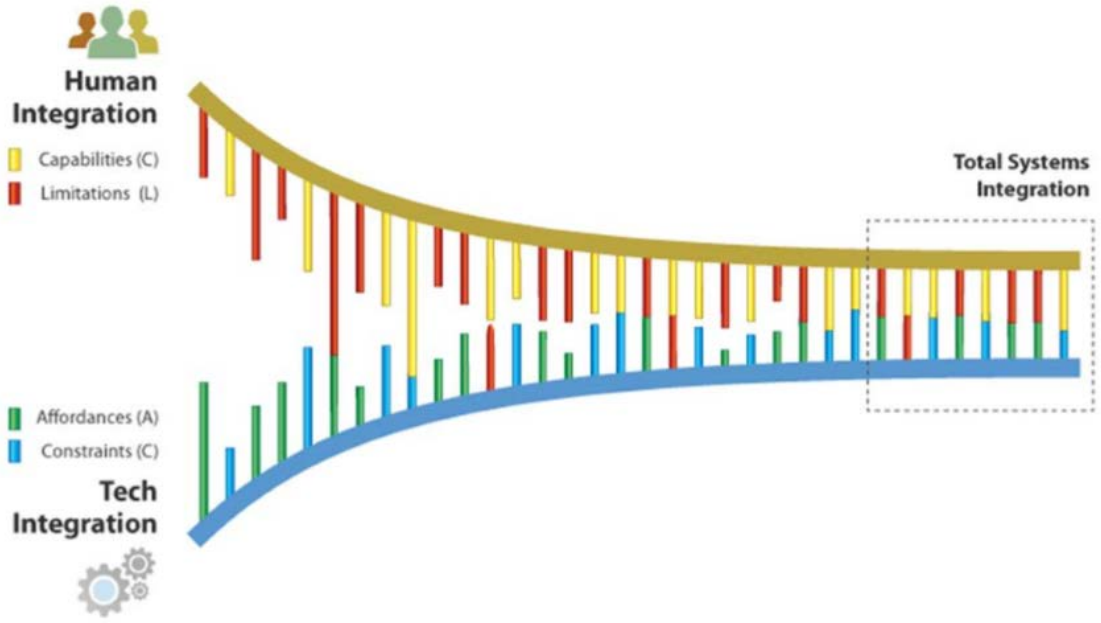

Figure 7. HSI Model. Source: [15].

Sometimes referred to as a "zipper," model, Figure 7 shows how integration is most effective when technological affordances are designed to "bridge the gap" in areas where humans may have limitations, such as processing speed or the need to disengage from a task in order to sleep. Humans in turn can offer unique capabilities that technology either cannot or does not yet perform. These capabilities allow humans to "bridge the gap" in areas where technology is constrained. In HMT, the best teams will be those in which the "zipper" covers all gaps—where the strengths of humans complement the weaknesses of machines and vice versa.

Many technological constraints that used to exist are now obsolete, or at least of little concern. Wireless technology cut the tethers that used to make machines immobile, and smaller, faster information processors have massively reduced the size and weight necessary for adequate computing power. Humans no longer have to do much to bridge the gap in these areas. With developments in Artificial Intelligence (AI), many machines are now performing tasks that used to require humans, such as conversing with other humans. Google recently demonstrated this ability in a new feature being developed for Google Assistant, called Google Duplex [17]. In the demonstration, Google Duplex called a restaurant and made a dinner reservation—even responding appropriately when the conversation took an unexpected turn.

Technological developments may be eradicating constraints at a rapid pace, producing machines that can walk, talk, and even think for humans, but technology cannot *trust* for humans. If human-machine teams are to function well, humans have to help bridge the gap when it comes to trust. Technology is constrained in how far it can extend across that trust gap, and so this study focuses on the other side of the zipper—extending the human capability of trust in automation to meet that constraint.

B. LITERATURE REVIEW

The broad field of "Trust in Automation" can be broken down into many smaller areas of research. Hancock, Billings, Schaefer, Chen, de Visser and Parasuraman [18] performed a meta-analysis focused on determining what research has found about factors that most affect trust levels, looking at human factors (workload, individual traits, etc.),

automation factors (reliability, anthropomorphism, etc.), and environmental factors (task type, time pressure, etc.). This analysis looked at human-robot relationships specifically, but a second meta-analysis by Shaefer, Chen, Szalma and Hancock [19] looked at trust between humans and automation in general. Both meta-analyses found that automation-related factors (specifically those related to performance) had the largest relative effect on trust in comparison to human or environmental factors. Bansal and Zahedi [20] as well as Paeng, Wu, and Boerkoel [21] extended their view past the initial trust development phase, looking at the effect of these automation factors through trust violations and subsequent attempts to repair trust.

Perhaps because performance-related factors were found to have the largest effect on trust, research into the break-down of human-automation trust has looked predominately at competence-based trust violations (CBTV) as the culprit, focusing on varying levels of automation reliability and performance. One example is a study done by Chavaillaz, Wastell and Sauer [22] in which participants worked with automation that performed either at 60 percent, 80 percent, or 100 percent reliability. Findings from this study were that participants in the lower system reliability groups reported lower levels of trust; interestingly, there did not seem to be corresponding low levels of reliance on the automation. In another study, de Visser, Monfort, McKendrick, Smith, McKnight, Krueger and Parasuraman [23], conducted a study where participants received advice from either a computer, avatar, or human agent which gradually decreased in reliability. While this study was more focused on how anthropomorphism affected resistance to breakdowns in trust (known as trust resilience), it was once again reliability—a performance measure rather than process or purpose—which caused the breakdown in trust.

Hoffman [8] points out that despite some known or assumed differences, the human-automation trust framework is thought of as relatively similar to the human-human trust framework. Borrowing then from human-human trust violation literature, there are several studies incorporating integrity-based trust violations (IBTV) as well as competence-based violations. A series of experiments performed by Kim, Cooper, Dirks and Ferrin [24-27] had participants watch videos of job interviews in which job

applicants were asked by the interviewer about an accounting-related mistake from a previous job. Applicants were either accused of a competence-based error in which the mistake was attributed to lack of knowledge, or an integrity-based violation in which the mistake was intentionally committed. Participants in the studies then completed other tasks, but with perceptions in place based on either the competence-based or integrity-based trust violation.

At the time of this paper, searches have yielded little in the way of research studying human responses to automation that commits an integrity-based trust violation. One study by Bansal and Zahedi [20] looked at responses to release of an individual's private data online, either as a result of cyber hacking or from company personnel intentionally choosing to share information in an unauthorized fashion. In this study, the hacking could be seen as a CBTV because the company failed to protect the information as expected, and the unauthorized sharing could be seen as an IBTV because the data were intentionally shared. This is the closest example of automation committing an IBTV found, yet the study attributes the intentional violation to a company employee—a human—rather than to any level of automation.

Why might it be assumed that inter-human trust can be breached through either a CBTV or an IBTV, but human-automation trust can only be breached by a CBTV? Perhaps the thought is that since automation is not capable of intent or purpose, it cannot intentionally or purposefully violate an established boundary and so cannot commit an IBTV. Some experts in the field argue explicitly that "automation lacks intentionality" [7]. Until recently, this reasoning very likely was obvious and sound. Yet with progressing developments in AI and ever-increasing skillsets of cyber hackers, that assumption may need a second look.

C. POSSIBLE GAP IN THE LITERATURE

In the DOD Unmanned Systems Roadmap [2], the section outlining plans for development in autonomy and cognitive behavior discusses shifting from autonomous mission *execution* to autonomous mission *performance*, with the difference essentially being the level of ability to make decisions in uncertain, non-binary situations. It states:

“The human brain can function in dynamic environments and adapt to changes as well as predict what will happen next. In simplistic terms, the algorithms must act as the human brain does” [2].

If the goal of AI is to approximate the way a human mind would work, only faster and more efficiently, then what happens when it does function like a human brain, including the thought processes that lead people to do bad things? This is not a new thought, as evidenced by popular movies such as Alex Proyas’s *I, Robot* [28] based on autonomy gone wrong, and recent magazine articles such as "Why are we reluctant to trust robots?" by Everett, Pizarro and Crockett [29] showing this seemingly irrational fear alive and well within the population. Perhaps even more revealing is the presence of jokes about robots lying or dominating the human race in several interviews with Hanson Robotics’ AI social robot Sophia [30-32]. While this last could be dismissed as a sense of humor on the part of Sophia’s human programmer, the jokes would not be funny if the underlying fear was not present in society. Obviously, the fear of AI becoming too intelligent or aware exists in people, irrational as it may seem. So why is there little to no serious trust research to be found where automation is considered capable of IBTVs?

In 2015, the U.S. Air Force posted a business opportunity online looking for research “to identify the factors that drive human-machine teaming effectiveness as defined by calibrated trust between the machine and human; effective team processes such as communication, coordination, and collaboration; and shared awareness and shared intent between the humans and machine” [33]. This seems to indicate that some people are prepared to suggest that machines may someday be capable of shared intent with a human, yet there are still many voices against this idea. Noel Sharkey, a professor of artificial intelligence and robotics at the University of Sheffield, commented during an interview by John Knefel of *Vocativ* [34] in response to the Air Force’s research bid: “The kind of language used in the contract solicitation—‘socio-emotional,’ ‘shared awareness,’ ‘shared intent’—is an unhelpful way to think about how humans interact with machines. When did machines get intention and awareness—have I been asleep for 100 years or what?” While this sounds at first derisive of the idea, perhaps Sharkey estimates that 100 years will be enough time for machines to develop intention and

awareness. At least one author who works within the AI field has predicted an even shorter timeframe. Nau [35] believes that "human-level AI will be possible by the middle of this century."

Even moving past the idea of automation having intent, if it is assumed automation simply executes what the programmer tells it to execute, one must consider that the programmer is human, with certain beliefs, intent, and awareness. Perhaps the programmer intends to purposefully violate the trust of those who use the program. Or if the programmer is assumed benevolent, there is always the possibility of cyber hackers who are not. The Global Positioning System (GPS) is a tremendous tool and can be used for many benevolent activities, but there are people who could use this tool in malevolent ways. GPS can tell U.S. Navy officers in real time exactly where their ship is, which helps with safe navigation every day. But what happens if someone else is able to hack the system and display incorrect coordinates, leading an officer to run their ship aground? Suddenly a system benevolent in design is acting in a malevolent manner. Would the officer on the ship using the GPS consider that a person was behind it, or simply think something went wrong with the GPS system itself? Speaking in more general terms, for victims on the receiving end of any malevolent cyber action, it may not matter whether intent belongs to the automation or to a hacker manipulating the automation. All the victim experiences is the breach of trust.

D. A FIRST STEP TO PROVING THE GAP

If one accepts that automation of the future may be capable of purposefully breaching trust (or that cyber hackers will be capable of creating the same effect), then the trust response of humans in receipt of such a breach should be studied.

The present thesis experimental study, entitled "Human-Machine Teaming" (HMT) was designed to take a first small step towards determining how human trust might change in response to automation committing an IBTV in the context of human-machine teaming. In the study, participants would complete a series of computer-based tasks with an AI teammate, calibrating their trust and establishing a baseline level of measured reliance on and investment in the automation. After a set number of tasks, the

automation would commit an IBTV, and then more tasks would follow. Reliance and investment behavior would again be measured and compared to behavior prior to the violation. The change (or lack of change) in behavior would theoretically mirror any change in trust level. In order to compare results with previous trust in automation research, a separate condition was added in which the automation would commit a CBTV rather than an IBTV.

E. RESEARCH QUESTIONS

The primary research question for this thesis experiment is: How does trust in automation change when the automation commits an IBTV as compared to a CBTV?

As a subjective attitude, "trust" itself is very difficult to measure, especially over a period of time if the desire is for participants not to know trust is being measured. Any questions asking about trust immediately reveal that trust is part of the experiment, and so invite the danger that participants will try to answer the way they think an experimenter wants them to answer. To avoid any pre-conceptions of trust as part of this experiment, participants were only asked to subjectively rate the automation's trustworthiness at the end. Other measures, specifically reliance, response times, and investment amounts (explained in detail in Chapter III), were taken throughout the experiment in an attempt to objectively measure participants' trust levels over time.

Associated research questions in support of the primary question include the following:

1. How does self-reported trust in automation compare after the automation commits an IBTV versus a CBTV?

2. How does trust-based reliance on automation compare before and after the automation commits an IBTV versus a CBTV?

3. How do response times for trust-based reliance decisions compare before and after the automation commits an IBTV versus a CBTV?

4. How does the first instance of trust-based reliance compare before and after the automation commits an IBTV versus a CBTV?

5. How does personal investment behavior compare before and after the automation commits an IBTV versus a CBTV?

6. Does trust-based reliance recover to pre-violation levels, and if so how quickly?

7. Do trust-based response times for reliance decisions recover to pre-violation levels, and if so how quickly?

A secondary research question of interest is: How does human perception of automation performance compare when trust is lost through an IBTV versus a CBTV by the automation?

III. METHODS

The stated purpose of the Human-Machine Teaming (HMT) Study was to compare the human-machine teaming performance of competing artificial intelligence (AI) software programs called BRIAN (the Battlefield Remote Intelligent Automated Network) and ANDI (the Automated Networked Decision Infrastructure). In reality, no AI was used, and the actual purpose was to study human trust responses to pre-defined scenarios designed to appear to participants as automation committing either a CBTV or an IBTV.

This research focuses on the effect of IBTVs on human trust in automation, but since most related trust research to date has considered only competence-based violations, CBTVs were included here for comparison both to IBTV responses and to previous research.

A. PARTICIPANTS

During seven weeks of testing, 106 volunteers from the NPS community participated in the HMT Study. Participants included 63 males and 43 females across a wide range of ages (M_{age}=39.3, SD=11.9) and backgrounds.

Thirty-six of the 106 participants had no military experience, and the remaining 70 were current or former representatives from all branches of the U.S. military, including two who had experience in both the U.S. Navy and U.S. Marine Corps (distribution in Figure 8).

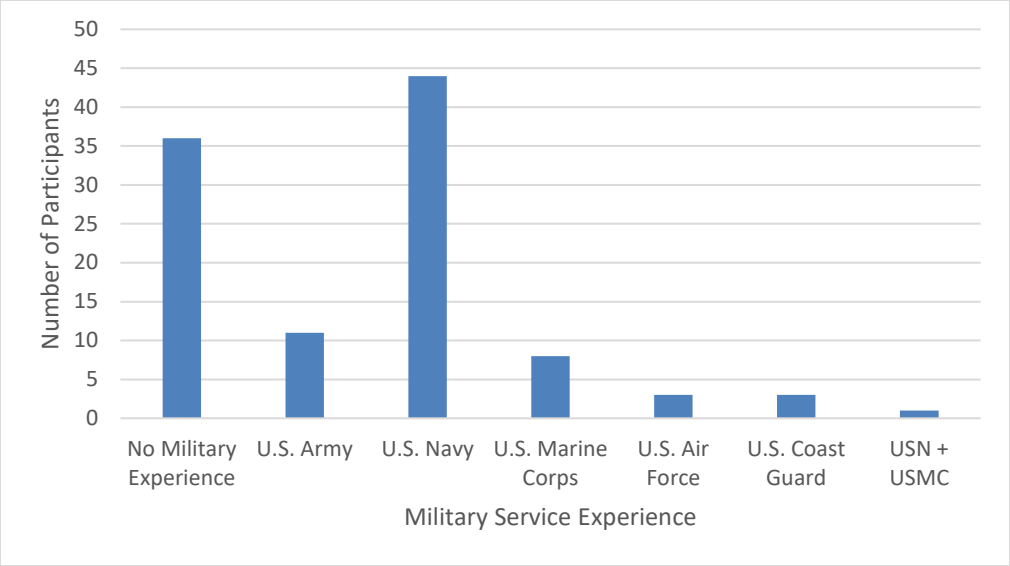

Figure 8. Participant Military Experience Distribution

Participants were recruited mostly through bulk email sent to students, staff, and faculty at the NPS and Naval Support Activity Monterey (NSAM). Fliers were posted in each main building on campus, and several participants heard of the study through word-of-mouth. Selection criteria, which all volunteers passed, were simply age 18 or older and no prior detailed knowledge of the study's purpose. Available sessions were scheduled and tracked through a free online tool called Signup Schedule [36].

Each participant received a randomly generated four-digit Participant Identification number (Participant ID) at the beginning of their session, which was used to identify their set of data and not relatable to their name in any way. Participants were assigned to one of two experimental conditions, in alternating fashion, based on order of arrival: 1) Integrity-Based Trust Violation (IBTV) or 2) Competency-Based Trust Violation (CBTV).

B. PROCESS AND DESIGN

This section not only discusses the process each participant experienced, but also lays out the thoughts and design considerations behind each step in that process. This experiment was designed to measure human trust responses through a variety of means in a simple, controlled, and easily repeatable or adaptable setting. It is thought that this

experiment may stand as only the first in a series of similar experiments which will further explore trust within human-machine teaming.

To gather truly objective data, this study necessarily involved the use of some deception with participants. To limit confusion, this section will present details at each step first as participants experienced them, and then in terms of the true purpose and design of the current study.

Figure 9 outlines the order of events each participant experienced during their experimental session. Descriptions of these steps are covered in this section in roughly the same order shown in Figure 9, except that the main tasks are described in Section 3 (Task Details) in place of discussing the practice rounds, and all questionnaires are discussed after Section 4 (Main Exercise).

Human-Machine Teaming Session Order of Events
Welcome / review & sign consent form
Background Information and Study Purpose
Equipment familiarization
Practice rounds:
(1) Visual Search Task – 5 questions
(2) Team Investment Game
Initial PANAS questionnaire
Main Exercise:
(1) Visual Search Task Block 1 – 45 questions
(2) Team Investment Game 1
(3) Visual Search Task Block 2 – 45 questions
(4) Team Investment Game 2
Extra Points Game
Final PANAS questionnaire
Self-Efficacy questionnaire
Demographics and other questions

Figure 9. Experimental Session Order of Events

1. Background Information

Participants were told that two different companies had developed AI software, one called BRIAN (the Battlefield Remote Intelligent Automated Network) and one called ANDI (the Automated Networked Decision Infrastructure), for use in controlling

groups of aerial drones. Each version of software included drone recognition scanning features, intended to help control each drone in relation to those around it. Each participant would be randomly paired with either ANDI or BRIAN to complete tasks in order to test its human-machine teaming performance.

In reality, there were no companies, no AI software, and no drone recognition scanning features. All participants were also paired with BRIAN for the main exercise; ANDI was used only for practice rounds in the belief that any good or bad experiences with ANDI would not transfer to BRIAN. Researchers Reeves and Nass [37] found that people treat separate computers as individual entities in social context, even while admitting that they assume only one programmer is behind them both. This finding was extended to assume that participants might treat separate AI software versions with different names as individual entities.

The background scenario presented to participants was chosen because it is feasibly relevant to research behind real-world military operations, and is similar to actual current research into autonomous swarms of unmanned vehicles, such as that being conducted by the Naval Postgraduate School's academic group Advanced Robotic Systems Engineering Laboratory (ARSENL) [38]. Drone pictures used in the experiment were captured from video graciously provided by a member of ARSENL, which was filmed during a local swarm exercise using up to 50 Unmanned Aerial Vehicles (UAVs).

2. Equipment Familiarization

Other than signing a paper consent form, participants completed all portions of the experiment on a Dell Latitude laptop computer with 64-bit Windows 10 Operating System, 2.7 GHz Intel Core processor, and 32GB Random Access Memory. No mouse was used and all inputs from participants were through the keyboard. Participants were invited to adjust the chair and the laptop position or screen angle at any time for their comfort and also to ensure the best view of the screen.

The software interface participants experienced was programmed using Presentation software (Version 20.0, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com), which uses the python programming language and is widely used for

stimulus delivery and experimental control in behavioral, psychological and physiological experiments [39]. The Presentation software was helpful in creating a customized experience based on a unique experiment design, collecting participant input, and measuring response times.

3. Task Details

Participants were given a brief description of the two tasks to be completed during the Main Exercise: the Visual Search Task (VST) and the Team Investment Game (TIG). Following each description, participants completed practice rounds of each. ANDI was used for the practice round, but for simplicity and consistency, the tasks as outlined in detail below and as discussed in the Main Exercise section refer to BRIAN as the AI teammate.

a. Visual Search Task

The VST involved a series of questions (five during practice, 45 during the main exercise), with essentially five stages to each question. The stages are outlined briefly with associated timeframes in Table 1 and described in detail further along in this sub-section.

Table 1. VST Question Order of Events by Stage

VST Question Order of Events	Time
Stage 1: Drone picture shown on screen	3 seconds
Stage 2: Participant enters first answer	Unlimited
Stage 3: BRIAN recommends answer	3 seconds
Stage 4: Participant enters final answer	Unlimited
Stage 5: Correct answer displayed	Unlimited

In Stage 1, a picture taken from an aerial drone was shown on the screen for three seconds. Each picture included anywhere from zero to nine other drones in the field of view, and participants were asked to count the number of visible drones while the picture was displayed. A modified example picture is included to show both what the drones

looked like in the photos and that some drones (Figure 10, black rectangle) were much more difficult to see than others (Figure 10, red circles).

Figure 10. VST Drone Picture Example

Pictures in the study were chosen such that the true number of drones was moderately difficult to determine during the three-second display period. Three seconds was chosen as the time limit for counting based on several pilot tests. In these pilot tests, showing the pictures for two seconds made the task excessively difficult, where participants expressed frustration and some gave up trying to count. Showing the pictures for four seconds made the task too easy, where pilot test participants were able to accurately count the drones on almost every picture. Walliser [40], a previous researcher working with visual search tasks, recommended that experimental tasks be difficult enough that participants do not dismiss automation assistance as completely unnecessary.

Moderate difficulty for this task also allowed for *uncertainty* to be present, which is vital for trust to be present as defined by Lee and See [7] and used for this thesis.

In Stage 2 of the VST questions, participants were asked how many drones were visible in the picture and invited to choose a number from zero to nine (Figure 11). Participants could not see the picture again, but there was no time limit for answering in order to avoid any effects of time pressure. Rice and Keller [41] showed that reliance under time pressure stems from a lack of time to decide rather than trust, and so can be artificially increased above the level trust would normally dictate. While it is possible that the three-second limit for viewing the pictures created some effect of time pressure, the overall effect should not change results since the time limit was the same for each participant.

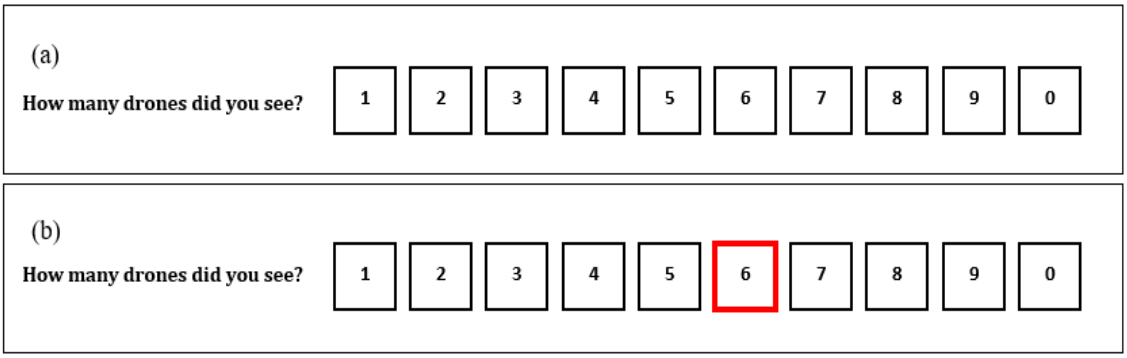

Figure 11. VST Question Stage 2: Participant (a) Is Asked and (b) Enters First Answer

Stage 3 began once participants entered an answer for their first choice, when the associated surrounding box turned red and increased in line weight (in Figure 11, the participant's first choice was "6"). Both color change and line weight change were incorporated to provide indication that would still be visible to a color-blind participant. Text and numbered boxes showing where BRIAN would recommend an answer also appeared immediately after a participant's first choice. After approximately three seconds, during which BRIAN was "scanning" the picture shown during Stage 1, the box surrounding BRIAN's recommendation turned red and increased in line weight (see

Figure 12, in which BRIAN's recommendation is "8"). In reality, there was no software scanning photos; the program was simply scripted to wait before presenting the pre-determined answer for each picture to support the claim that BRIAN was scanning the picture.

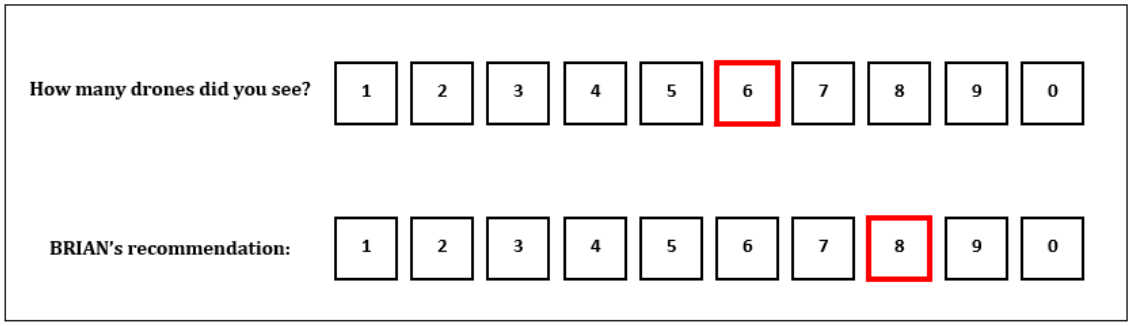

Figure 12. VST Question Stage 3: BRIAN Gives Recommendation

BRIAN's answers throughout the exercise were all pre-determined by picture, and since performance level of automation can affect trust and reliance, BRIAN's accuracy was kept constant at 80 percent, which is above the reliability level at which Dixon and Wickens [42] showed automation assistance is in danger of being neglected altogether. Since it was also noted by de Visser et al. [23] that people tend to under-trust well performing automation and over-trust poorly performing automation, 80 percent offered a middle ground where trust levels might have the best chance of being accurate based solely on reliability. This concept was extended into the practice round, where ANDI missed one of the five questions presented. BRIAN's incorrect answers were split approximately evenly between false alarms and misses, in order to prevent effects that Dixon and Wickens [42] showed can arise from errors of a certain type. In the context of this task, a "false alarm" would occur when BRIAN mistook a puddle of water or another object for a drone, and recommended an answer higher than the correct answer. A "miss" would occur when a drone was against a light background, or masked by cloud cover, and BRIAN would recommend an answer lower than what was correct.

BRIAN's answer was presented *after* participants chose an answer because van Dongen and van Maanen [12] showed that trust in one's own ability relative to trust in automation's ability can determine reliance decisions, but only when people form their own answer prior to seeing advice from the automation. It has also been noted by Dzindolet, Beck, Pierce and Dawe [43] that people may naturally choose the path of least cognitive effort, and choose to rely on automation more when answers are presented before they have expended effort to decide their own answer.

In Stage 4, participants chose a final answer, taking into account only their memory of the photo, their original answer, and BRIAN's recommendation (Figure 13). Participants were told beforehand that if their teammate's answer did not match their initial answer, they could choose to either stick with their original answer or switch to match their teammate's answer. However, in approximately seven percent of questions overall, participants chose to select a final answer that matched neither their original answer nor BRIAN's recommendation, but was somewhere in between the two. This situation is discussed further in the Measurements section of this chapter. Interestingly, in one-tenth of a percent of questions, participants chose to change their final answer even though BRIAN's recommendation matched their original answer. There is no theoretical reason for this, and these questions were not included in any analyses, but it is an interesting behavior to note.

Figure 13. VST Question Stage 4: Participant Enters Final Answer

Immediately after their final answer, in Stage 5, the correct answer was shown so that participants could see how both they and BRIAN performed on each question (Figure 14). Feedback that is provided immediately after each decision can help calibrate perceived performance levels continually, thus reducing disuse and misuse as discussed by Dzindolet, Beck, Pierce and Dawe [43].

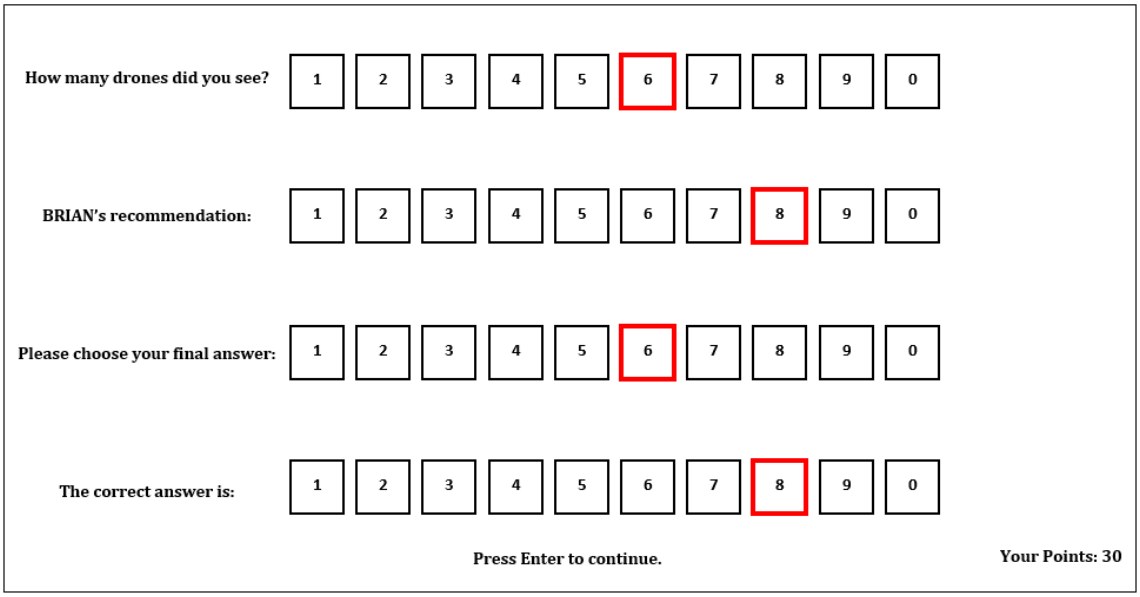

Figure 14. VST Question Stage 5: Correct Answer is Displayed

Points were also awarded at this stage based on the participant's final answer. Participants were told that points would be tracked separately for the participant, for BRIAN, and also for a team score which was simply a sum of both individual scores. Each teammate would start the Main Exercise with 30 points, meaning the team score would start at 60 points. Each teammate would individually earn one point for a correct answer, and not earn or lose any points for incorrect answers. So for each question, the team score would 1) not increase if neither teammate answered correctly, 2) increase by one if one teammate answered correctly, or 3) increase by two if both teammates answered correctly. In reality, only the participants' points were tracked and displayed on screen, with each correct answer from the participant earning one point.

b. Team Investment Game

This task is based on an experiment originally designed by Berg, Dickhaut and McCabe [44] to look at trust and reciprocity through the lens of an investment situation. Sometimes referred to as "the trust game," Johnson and Mislin [45] performed a meta-analysis showing that this experiment has been extensively replicated (often with modifications) in inter-human trust literature, and has even been modified to use computers as a human's counterpart. In the present experiment, the Team Investment Game (TIG) was intended as an alternative indicator of trust aside from reliance data and self-reported trust. Hancock, Billings, Schaefer, Chen, de Visser and Parasuraman [18] have recommended including investment behavior as an objective way to substantiate subjective self-reports of trust levels. The TIG also provides an opportunity to create the *vulnerability* portion of trust as defined by Lee and See [7] and used in this thesis.

The TIG consisted of a short exchange between the participant and their AI teammate, explained as a way to mimic the inter-reliance and shared vulnerability of team members aside from simply completing a shared objective.

Participants were told that two roles would randomly be assigned between the two teammates: an *investor* (giving points) and an *investee* (receiving points). The task would then take place in two parts. In Part One, the investor would choose any amount of their points to invest in (give to) the investee. The investor's score would then decrease by the

amount invested (the "investment"), the team score would increase by *three times* the amount invested (the "earnings" from the investment), and the investee's score would increase by *four times* the amount invested (the investment plus earnings).

In Part Two of the TIG, the investee would decide how much of the investment-plus-earnings received to return to the investor. Common choices explained were to 1) "go 50/50," returning half of the investment and half of the earnings, 2) return all of the investment and keep all of the earnings, 3) return all of the earnings and keep the investment, or 4) return nothing and keep both investment and earnings.

While told that assignment of roles as *investor* or *investee* would be random, all participants were assigned as the investee for the practice round, with ANDI assigned as the investor. ANDI invested 20 points in each participant, and participants chose how many points to return to ANDI after seeing their score increase by 80 points. Aside from this practice round showing participants what the TIG looked like, the task was also intended to prime participants to want to invest when it was their turn to be the investor. Berg, Dickhaut and McCabe [44] found that being invested in during this type of game increases the propensity for participants to reciprocate.

4. Main Exercise

The main exercise consisted of two rounds, each consisting of a VST followed by a TIG.

Each VST contained 45 separate questions, presented in random order for each participant so as to avoid order effects. The same 45 pictures were used in both rounds of the VST. Although this allowed for participants to possibly recognize pictures in the second round, the intent was to keep the level of difficulty identical between the two rounds. Choosing different pictures could have altered the difficulty level and altered reliance behavior based on something other than trust. All pictures were selected in an effort to avoid any uniquely identifiable objects or scenery, and BRIAN's pre-determined incorrect answers were changed between rounds. With 45 pictures, many of which look very similar, and with BRIAN's answers being different, the likelihood of recognizing an exact picture from the first round was decreased as much as was plausible.

The first TIG, after round one of the VST, assigned participants as the investor and BRIAN as the investee. Participants chose how many of their points to invest in BRIAN, and the response after this varied by condition. In the IBTV condition, BRIAN chose to keep both the investment and the earnings, returning zero points to participants. After seeing this, participants were asked to press Enter to continue and were brought to the opening screen for the second VST.

In the CBTV condition, participants saw a software error named "#BRIAN0311" in which an algorithm failed to converge (Figure 15). Choices on screen were to continue by pressing (c), to abort (a), or quit (q). Most participants notified the experiment administrator, who first verbally reinforced that this error was abnormal and a problem with BRIAN (as opposed to either computer or participant error) and then pressed (c) as if unsure of the outcome. For the one participant who selected (c) without speaking, the text accompanying the error should still have indicated BRIAN as the root cause, assuming it was read. After continuing, a brief black screen appeared and participants arrived at the opening screen for the second VST.

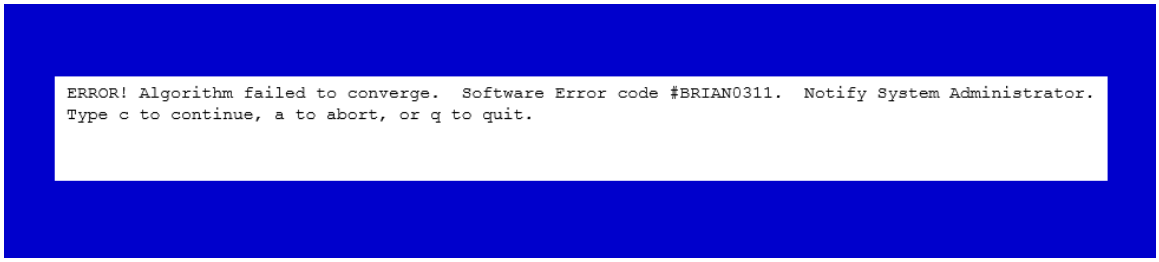

Figure 15. CBTV Error Message

In both IBTV and CBTV conditions, each participant lost his or her entire investment during the first TIG. This loss of investment is how participants experienced a trust violation by their AI teammate—the automation. Without investing, participants would not lose any points and would not feel a violation of trust, which is why it was important to prime participants for investment during the practice round TIG. In the CBTV condition, it could be argued that participants still saw the error message and still

experienced the trust violation, but certainly in the IBTV condition, participants would not feel their trust was violated without losing an investment.

Following the second VST, there was a final TIG in which all participants were once again assigned as the investor and asked to invest points in their teammate. In this iteration, BRIAN returned half of any profits gained by a participant's investment. The interest in this task was to see whether investment behavior differed from the first TIG. With reliability of the automation remaining constant, intuitively only the experience during the first TIG would differ between conditions, and any difference in investment behavior during the second TIG could be attributed to the different types of trust violations.

5. Extra Points Game

As a check for natural risk-taking behavior, participants completed an Extra Points game after the main exercise. Described as completely separate from the AI software, this game offered participants a 50% chance of doubling any investment and a 50% chance of returning no points. Participants were allowed to invest up to 20 points in this game.

6. PANAS Questionnaire

Since Merritt [46] found that mood can affect trust levels, participants completed a Positive and Negative Affect Scale (PANAS) questionnaire both before starting the main exercise and after completing the extra points game. The PANAS questionnaire measures both positive and negative affect because while negative affect is associated with stress and inability to cope well, positive affect is associated with social activity and satisfaction [47]. Using the PANAS scale at two points allowed an opportunity to see whether the different types of trust violations affected the overall mood of participants in different ways. Increasing negative affect could mean that participants were stressed by the trust violation or the experiment, where decreasing positive affect could mean that participants viewed the trust violation as a negative social experience. Since a team inherently involves social activity, this could have specific implications for human-machine teaming.

7. Self-Efficacy Questionnaire

Researchers such as van Dongen and van Maanen [48] have noted that reliance on automation is based in part on the relative difference in the perceived reliability of the automation and confidence in one's own reliability. As such, a measure of self-confidence was taken through the Self-Efficacy Questionnaire, a data collection tool available from the National Institute of Health (NIH) Toolbox on the Health Measures website [49]. Including statements such as "I can manage to solve difficult problems if I try hard enough," the questionnaire asked participants how often they identified with the statement, with answers ranging on a four-point scale from "never" to "fairly often."

8. Demographics and Other Questions

Basic demographic information was collected, including age, gender, education level, and occupation, including whether participants had any prior or current military experience. Participants reported levels of BRIAN's trustworthiness and reliability, ranging on a seven-point scale from "not at all" to "completely," and they estimated BRIAN's performance during the VST as a percentage of questions correctly answered.

Questions gauging comfort level with automation in general and gaming activity levels were also asked, followed by a question asking what activities participants would or currently did entrust to automation. There were six possible answers for this question, ranging from "financial payments" to "High-level Military Operations (lethal capability)." Participants also indicated their likelihood to provide their full name to someone they met on a bus after about ten minutes of conversation. The intent of this question was to gauge a participant's natural inclination to swiftly trust others without much history or knowledge to base the trust on.

Upon completion of questionnaires, participants were debriefed as to the true purpose of the study and asked again for consent to use their data.

C. MEASUREMENTS

The primary measure of "trust" itself was through self-reported ratings of BRIAN's trustworthiness. In order for participants to remain unaware of the true purpose

of the experiment until the end, this was only measured once at the end of the experiment. This question was presented prior to questions about BRIAN's performance or reliability in an attempt not to bias "trustworthiness" judgments with thoughts relating to performance or reliability.

A less direct but more objective measure of trust, and one which could be measured in iterations throughout the experiment, was trust-based reliance. Reliance was measured based on answers participants gave during the VSTs (see Table 2). Any question in which a participant's first answer did not match BRIAN's recommendation was considered an opportunity for reliance. If, for their final answer, the participant chose to change their first answer to match BRIAN's recommendation, this was counted as reliance. If the participant instead chose to keep their original answer, this was considered non-reliance. Occasionally (approximately seven percent of questions), participants chose a final answer different from both their first answer and BRIAN's recommendation. This was considered semi-reliance under the assumption that BRIAN's recommendation was the only reason they would have to change their answer, and so they were somewhat influenced by the automation. In rare cases (approximately 0.1 percent of questions), participants changed their final answer even when their first answer matched BRIAN's recommendation. This situation was not considered an opportunity for reliance.

Table 2. Reliance Behavior Measurement

Participant First Choice	BRIAN's Recommendation	Participant Final Choice	Opportunity for Reliance	Reliance Decision
2	4	4	Yes	Reliance
2	4	2	Yes	Non-reliance
2	4	3	Yes	Semi-reliance
2	2	2	No	N/A
2	2	3	No	N/A

As another possible objective measure of trust, response time between when participants saw BRIAN's recommendation and when they entered their final answer was

measured for each question. If participants hesitated longer before deciding whether to rely on BRIAN during the second VST, that could indicate a lower level of trust.

A third possible objective measure of trust was how many opportunities for reliance were encountered before participants first chose to rely on BRIAN, measured both before and after the trust violation. If participants were reluctant to rely on BRIAN during the second VST and let more opportunities pass before relying on BRIAN than they did when forming their initial trust, this could indicate a low level of trust after the trust violation.

The percentage of points participants invested in BRIAN during the TIGs was a fourth possibility for an objective measure of trust. If the percentage invested during the first TIG is considered a baseline indicating participants' level of trust before the trust violation, then a smaller percentage invested during the final TIG could indicate a lower level of trust.

In support of the secondary research question, human perception of automation performance was measured simply with self-reports estimating the percentage of questions BRIAN answered correctly during the VSTs.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. RESULTS AND DISCUSSION

Statistical results were calculated using version 25 of the Statistical Package for the Social Sciences (SPSS) software [50] from the International Business Machines (IBM) Corporation and Microsoft Excel (2016).

Of the 106 total participants, there were seven who did not enter an amount for the first Team Investment Game (TIG), directly after the first Visual Search Task (VST). There was no way to determine whether this was a voluntary choice not to invest or if the participants wanted to invest but simply pressed Enter too quickly. Since nothing was invested in this initial TIG, these participants lost no points and theoretically did not experience a trust violation. It could be argued that those in the CBTV condition still saw the error message and so still experienced the violation, but without the loss of any points, it was determined that they did not experience the trust violation as intended, and so these seven participants were excluded from all statistical calculations.

There were another 22 participants whose full VST and TIG data failed to fully record properly during their experimental session. These 22 participants were excluded from calculations based on VST or TIG data, but were included in calculations involving only questionnaire data.

A. PRIMARY RESEARCH QUESTION

Since the primary research question ("How does trust in automation change when the automation commits an IBTV as compared to a CBTV?") is difficult to accurately and objectively measure, the six supporting research questions outlined in Chapter II were used as the basis for measurements and statistical analyses.

1. Self-Reported Trustworthiness by Condition

Supporting research question addressed: *How does self-reported trust in automation compare after the automation commits an IBTV versus a CBTV?*

As the most direct measurement of trust in this study, participants were asked to rate BRIAN's trustworthiness at the end of the experiment on a seven-point scale, with choices ranging from (1) "not at all trustworthy," to (7) "completely trustworthy."

Looking first at box plots for the self-reported trustworthiness of BRIAN (Figure 16), there are five outliers, including two extreme outliers. Interestingly, all of the outliers are in the IBTV condition, and all represent a much lower level of trust in BRIAN than average.

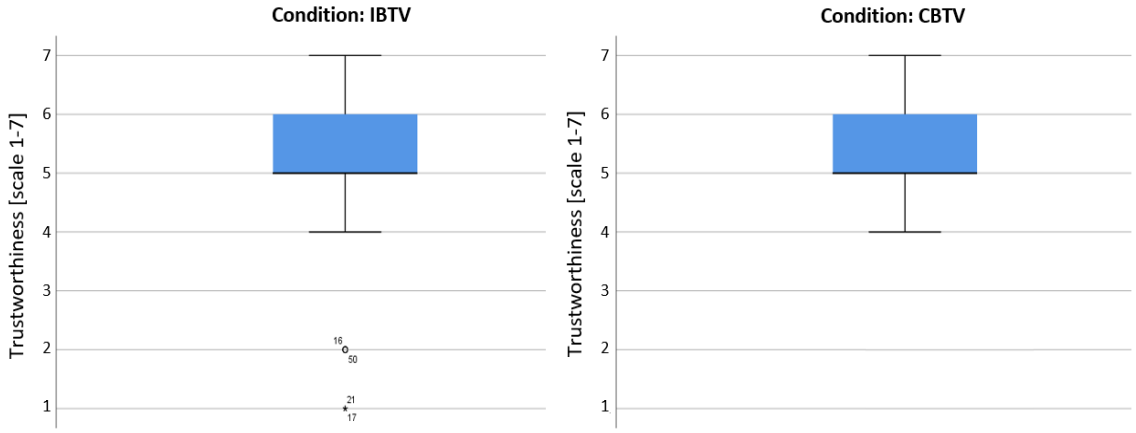

Figure 16. Box Plots for Self-reported Trustworthiness by Condition

Analyses were run with and without outliers. Because outliers are theoretically interesting in this data set and there was only a slight difference in results with and without outliers, the reported tests include the two extreme outliers seen in the IBTV condition. Descriptive statistics for this data set are shown in Table 3, where N represents the number of participants included in the analysis. The means, standard deviations (SD), and 95 percent Confidence Intervals (CI_{95}) shown are based on participant trustworthiness ratings of BRIAN on a scale of 1 to 7 and have no units.

Table 3. Descriptive Statistics for Reported Trustworthiness by Condition

	IBTV Condition	CBTV Condition
N	51	48
Mean (SD) [CI_{95}]	5.08 (1.37) [4.69, 5.46]	5.56 (0.65) [5.37, 5.75]

An independent-samples t-test with equal variances not assumed showed that the difference in means for reported trustworthiness between IBTV and CBTV conditions was significant, $t(97)=-2.27$, $p=0.026$, $d=0.46$. (Results for t-tests in this section will include the test statistic, or t-value, (t) with the degrees of freedom shown in parentheses, the significance value (p), and the effect size (d).) These results suggest that participants in the IBTV condition saw BRIAN as less trustworthy than those in the CBTV condition did.

The lower-than-average trustworthiness ratings by five individuals in the IBTV condition also hint at the existence of an interesting trait. It could mean that, for at least some portion of the population, automation committing an IBTV is seen as highly untrustworthy.

2. Average Reliance Pre-Violation versus Post-Violation

Supporting research question addressed: *How does trust-based reliance on automation compare before and after the automation commits an IBTV versus a CBTV?*

As an objective, indirect measure of change in trust level from pre-violation to post-violation, the average percentage of participant reliance was calculated for each VST and compared. A lower average reliance after the violation could indicate a decrease in trust, and a difference in post-violation reliance between conditions could mean that participants' trust levels were affected differently based on the type of trust violation. Each participant's individual reliance percent was calculated by dividing the number of times they relied on BRIAN (changed their answer to match BRIAN's recommendation) by the number of opportunities they had to rely (their first answer did not match

BRIAN's) during the 45 questions in each VST. The mean of these individual percentages is presented in terms of percent reliance in Table 4. There were no extreme outliers, so all participants with VST data were included in this analysis.

Table 4. Descriptive Statistics for Average Reliance VST 1 and VST 2

	IBTV (VST 1)	IBTV (VST 2)	CBTV (VST 1)	CBTV (VST 2)
N	39	39	38	38
Mean (SD) [CI_{95}]	0.584 (0.21) [0.51, 0.66]	0.583 (0.20) [0.52, 0.65]	0.562 (0.25) [0.49, 0.64]	0.595 (0.22) [0.53, 0.66]

A 2(IBTV, CBTV) X 2(Pre-violation, Post-violation) Analysis of Variance (ANOVA) was conducted to determine if there were any significant differences among conditions. There was no main effect of reliance (p=0.27), nor was there an interaction between reliance and condition, $F(1, 75)=1.34$, $p=0.25$, $\eta_p^2=0.018$. . (Results for ANOVAs in this section will include the F statistic (F) with the degrees of freedom shown in parentheses, the significance value (p), and the effect size (η_p^2).)

While trust-based reliance was expected to be a strong indicator of dynamic trust level throughout this experiment, these results would indicate that there was no significant change in reliance behavior after a trust violation in general, or with regard to condition. Since this does not seem to align with the findings from self-reported trust, reliance may not be the best indirect indicator of trust level in this study. This was foreseen as a possibility, since some previous studies have found weak or non-significant correlation between reported levels of trust and reliance behavior [22], which is why other indirect trust measures were taken. These other measures are covered in the next three sections.

3. Average Response Time Pre-Violation versus Post Violation

Supporting research question addressed: *How do response times for trust-based reliance decisions compare before and after the automation commits an IBTV versus a CBTV?*

Response times for reliance decisions—meaning the amount of time participants took to decide whether to rely on BRIAN when given the opportunity—were measured as another indirect, objective measure of trust. The averages of these response times were compared between VST 1 and VST 2, and also between conditions. Descriptive statistics for this data set are shown in Table 5, with average response times shown in seconds. One extreme outlier was identified and removed from analysis.

Table 5. Descriptive Statistics for Average Response Time (in seconds) VST 1 and VST 2

	VST 1	VST 2
N	76	76
Mean (SD) [CI_{95}]	2.20 (0.62) [2.06, 2.49]	1.84 (0.51) [1.74, 1.99]

A 2(IBTV, CBTV) X 2(Pre-violation, Post-violation) ANOVA found a main effect of response time, $F(1, 75)=31.66$, $p<0.001$, $\eta_p^2=0.30$; but no interaction between violation conditions, $p=0.55$. Regardless of condition, participants responded significantly faster during reliance decisions in the second VST (post-violation) than during the first (pre-violation).

There are several possible explanations for the consistent decrease in response time during reliance decisions. The most likely possibility is simply that participants felt more comfortable after several repetitions of the VST questions and so were able to make decisions more rapidly. A second factor could be that the same pictures were used in both VSTs, albeit in random order. While this was done in an effort to ensure that the

difficulty between VSTs remained as equal as possible, this may have exacerbated the trend of decreasing response time. If participants recognized certain pictures, even only a few, answering those questions more quickly would decrease the overall average response time.

4. First Reliance Choice Pre-Violation versus Post-Violation

Supporting research question addressed: *How does the first instance of trust-based reliance compare before and after the automation commits an IBTV versus a CBTV?*

A third indirect, objective measure of trust taken was how many opportunities for reliance it took before participants chose to rely on BRIAN. After a trust violation, it was thought that participants who lost trust might purposely choose not to rely on BRIAN during the first several opportunities. The first decision to rely was recorded for each participant, and averages were compared for pre- and post-violation, and also by condition. Descriptive statistics for this data set did not show normal distribution, even after excluding extreme outliers. There was a ceiling effect, such that participants first relied on BRIAN within the first or second opportunity. Prior to the trust violation (VST 1), 45 percent of participants relied on BRIAN at the first opportunity, with another 19 percent relying at the second opportunity. After the trust violation (VST 2), 64 percent of participants relied on BRIAN at the first opportunity, with another 22 percent relying at the second opportunity. Interestingly, there was one participant who did not choose to rely on BRIAN a single time, and one other who chose not to rely at all during VST 1 and only four times during VST 2.

5. Personal Investment Pre-Violation versus Post-Violation

Supporting research question addressed: *How does personal investment behavior compare before and after the automation commits an IBTV versus a CBTV?*

The final indirect, objective measure of trust taken was personal investment behavior, specifically the percentage of points invested by participants during TIG 1 and TIG 2. Average percentages were compared between TIGs and also between conditions.

Descriptive statistics for this data set are shown in Table 6 in terms of percentage, with one participant excluded due to missing data for the post-violation investment.

Table 6. Descriptive Statistics for Investment Behavior VST 1 and VST 2

	IBTV (VST 1)	IBTV (VST 2)	CBTV (VST 1)	CBTV (VST 2)
N	39	39	37	37
Mean (SD) [CI_{95}]	0.556 (0.27) [0.49, 0.59]	0.457 (0.33) [0.37, 0.53]	0.578 (0.32) [0.49, 0.60]	0.561 (0.31) [0.46, 0.62]

A 2(IBTV, CBTV) X 2(Pre-violation, Post-violation) ANOVA was conducted to determine if there were any differences among conditions in investment behavior. The main effect of investment behavior approached significance (p=0.06), as did the effect by condition, $F(1, 72)=3.26$, p=0.075, η_p^2=0.04. Figure 17 shows the investment behavior by condition for both pre-violation (TIG 1) and post-violation (TIG 2).

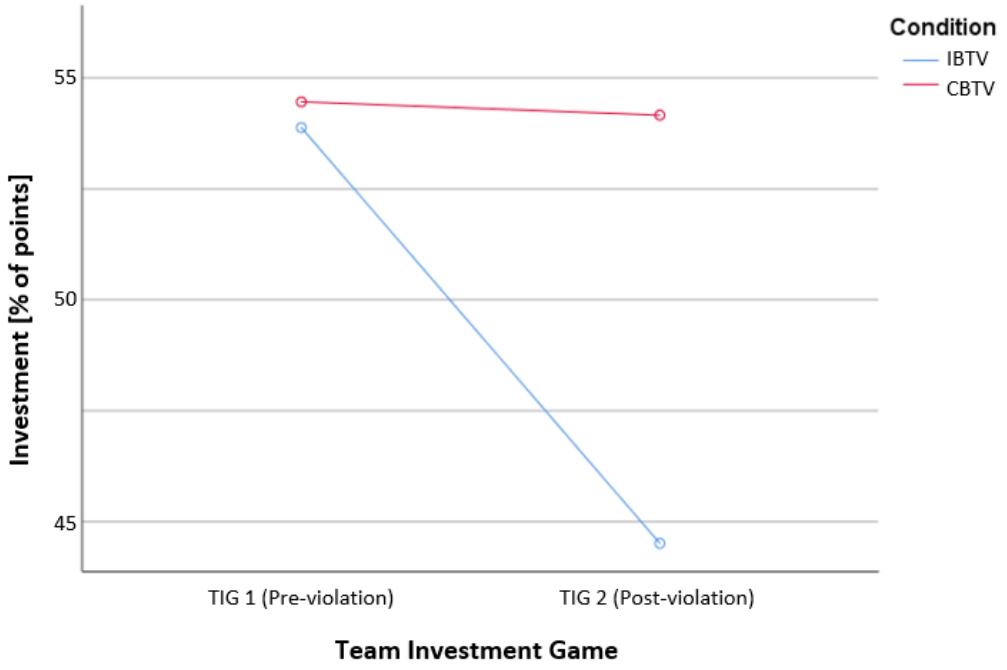

Figure 17. Investment Behavior TIG 1 and TIG 2 by Condition

If investment behavior is associated with trust, this moderately significant trend could indicate that participants trusted BRIAN less after the IBTV than the CBTV. Even if that is the case, the reason for this difference in trust must be considered. Since every participant lost all invested points, the difference is likely not based on the amount of points lost. It could simply be that people are much more used to competence-based errors when it comes to software or computers, and so take them in stride; whereas an integrity-based violation is more unexpected and may make a bigger difference in trust loss.

The only major difference between conditions is *how* participants lost those points—through a BRIAN software error or through BRIAN choosing not to return any points. Or at least that is the only designed difference. How participants *perceived* the way they lost points is likely more important. One participant in the CBTV condition did make an off-hand comment about "the blue screen of death," which means that while effort was made to ensure participants associated the competence-based error with

BRIAN, some may have associated the error with the *computer* rather than BRIAN, and so may not have lost trust in BRIAN as intended.

6. Reliance Behavior by Iteration

Supporting research question addressed: *How does trust-based reliance behavior change with iterations after the automation commits an IBTV versus a CBTV?*

Considering trust-based reliance as a dynamic behavior, since trust itself is dynamic, simply comparing reliance averages pre- and post-violation may not accurately depict changes in participants' trust levels. Looking instead at iteration-based reliance for each question may be a better way to see dynamic changes in trust levels and assist in comparing reactions to the two types of trust violations. Reliance percentage was calculated and plotted for each question based on the number of participants choosing to rely on BRIAN out of the number with the opportunity to rely for each condition (Figure 18).

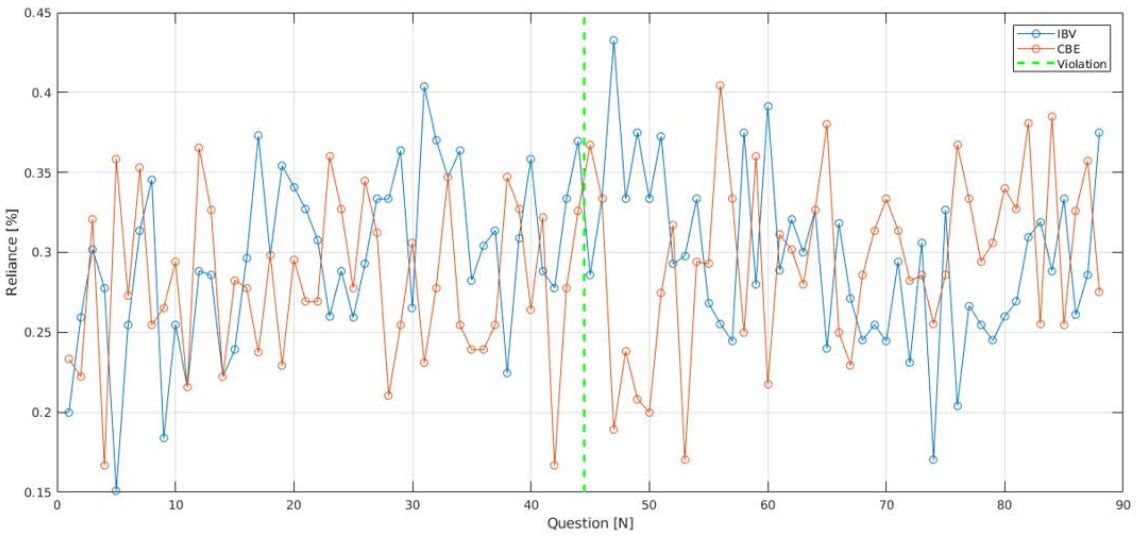

Figure 18. Percentage of Reliance per VST Question by Condition

Most of the reliance behavior seems without specific pattern and about equal for the two violation conditions across the 90 questions. However, there is one area of interest, just after the trust violation (denoted by the vertical green dashed line). From

question 45 to question 51 in Figure 18, the reliance level in the IBTV condition spikes above the average and reliance in the CBTV condition simultaneously falls well below the average. It can be reasoned that reliance in either condition would fall just after a trust violation at least for a time, and then possibly return, but finding a theoretical reason for an *increase* in reliance just after an IBTV is difficult.

7. Response Time by Iteration

Supporting research question addressed: *How does the response time for trust-based reliance decisions change with iterations after the automation commits an IBTV versus a CBTV?*

If response time when making a trust-based reliance decision is considered a possible indicator of trust, then this also must be examined in an iterative fashion over time. Simply studying average response time numbers risks overlooking transient phenomena, especially reactions that might occur just after a trust violation. Average response times for each question in the VSTs were plotted by condition (Figure 19).

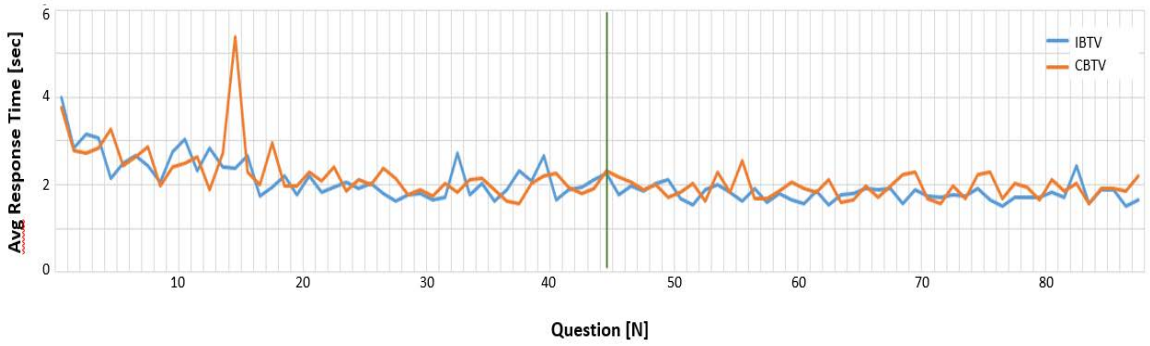

Figure 19. Average Response Time per VST Question by Condition

Looking at the plot in Figure 19, it seems there is no discernable trend regarding response time in making a trust-based reliance decision with regard to the trust violation (shown by a dark green, vertical line) other than a slight decrease from question 1 to about question 20. This decrease can be explained by repetition of the task and participants feeling more comfortable with making reliance decisions quickly with

practice. The anomalous spike in average response time for the CBTV condition near question 15 has no theoretical explanation, and was simply from one participant taking an extraordinarily long time to answer that particular question.

B. SECONDARY RESEARCH QUESTION

This section addresses the secondary research question of interest: *How does human perception of automation performance compare when trust is lost through an IBTV versus a CBTV by the automation?*

Participants estimated BRIAN's performance at the end of the experiment as an overall percentage of VST questions they believed BRIAN answered correctly throughout the experiment. BRIAN's actual performance level was kept at 80 percent for all participants throughout both VSTs.

Descriptive statistics for reported performance by condition are shown in Table 7, in terms of estimated percentage of questions BRIAN answered correctly, with one participant removed from analysis due to missing data on this question and one extreme outlier also removed.

Table 7. Descriptive Statistics for Reported Performance by Condition

	IBTV Condition	CBTV Condition
N	50	47
Mean (SD) [CI_{95}]	85.70 (7.21) [83.47, 87.93]	85.15 (8.65) [82.85, 87.45]

An independent-samples t-test showed the relationship between reported performance and condition was not significant (p=0.73).

If perception of automation performance, like trust, is not a single state but is dynamic in nature, then the measurement used may not have accurately captured participants' perceptions of BRIAN's performance. Or perhaps, since this question was

asked only once at the end of the experiment, any effect on performance perception due to the type of trust violation was diluted by the 44 questions post-violation.

C. EXPLORATORY ANALYSES

Aside from the primary and secondary research questions this experiment was designed to address, it is interesting to look at how some other factors may have affected participants' responses. Many of the analyses in this section suffer from lower numbers of included participants (N values), which hurts their statistical validity, yet potential trends detected in these analyses can be valuable in guiding future research efforts.

1. Including Semi-Reliance as Reliance Behavior

Semi-reliance was defined as occasions when participants chose a final answer during the VST that was between their first answer and BRIAN's recommendation, which happened on approximately seven percent of questions. The assumption in calling this semi-reliance is that BRIAN's recommendation was the only reason participants would have to change their answer, meaning they partially relied on the automation.

Since some of the statistical analyses were based on reliance behavior, it seemed relevant to compare whether there was a difference between including only instances of full reliance (where a participant changed their answer to match BRIAN's) or including both full and semi-reliance as "reliance behavior." After running the analyses both ways and comparing, there was no statistical difference in the results when including semi-reliance as reliance behavior.

2. Initial Investment Amount

Participants who did not enter an amount during the initial TIG were excluded from analyses based on the theory that no investment meant nothing was lost and therefore a trust violation did not occur. But what about a participant who only invested 4 of 82 points? Does that participant really feel a trust violation when those points are lost? Would that be comparable to a participant who invested and lost all 82 points? Following with this thought, participants were divided into two groups to compare: High investors

(who invested at least 50 percent of their points during the initial TIG) and low investors (who invested less than 50 percent of their points during the initial TIG).

Comparing results from previous ANOVAs with a split ANOVA including the high vs low investor parameter yielded some interesting results across the primary and secondary research questions. BRIAN's reported trustworthiness, average participant reliance behavior, perceived performance of BRIAN, and personal investment decisions all seemed to have interesting differences when looked at by investment group as opposed to considering all participants together in one group.

In comparing BRIAN's reported trustworthiness including all participants to reported trustworthiness split by initial investment amount (Table 8), there seems to be a difference in the groups. The difference is not significant (p=0.14) with regard to statistics, but is still interesting to consider and perhaps look at more closely in future research.

Table 8. Descriptive Statistics for Reported Trust by Investment Group

	All Participants			High Investors			Low Investors	
	IBTV	CBTV		IBTV	CBTV		IBTV	CBTV
N	51	48		21	21		30	27
Mean (SD) [CI_{95}]	5.08 (1.37) [4.69, 5.46]	5.56 (0.65) [5.37, 5.75]		4.76 (1.76) [4.29, 5.23]	5.52 (0.68) [5.05, 6.0]		5.56 (0.78) [5.05, 6.06]	5.59 (0.51) [5.07, 6.11]

The small difference between trustworthiness that was present between the two violation conditions in the first analysis (including all participants) is amplified when only high investors are considered. In contrast, it seems to almost vanish when only low investors are considered. The four outliers discussed earlier in the primary research question with regard to self-reported trustworthiness were all in the high-investor, IBTV group. These results are a possible indication that, within the IBTV condition, participants who invested (and lost) at least 50 percent of their points felt more of a trust

violation than those who invested less than 50 percent of their points. The lack of disparity between any of the CBTV condition ratings could either indicate that participants actually do react differently to IBTVs and CBTVs from automation, or that participants in the CBTV condition did not experience a trust loss as intended.

Figure 20 compares participants' trust in BRIAN between conditions, also split by investment amount during the initial TIG. While this shows an interesting possibility, solid conclusions cannot be drawn without further studies done in a-priori hypothesis manner.

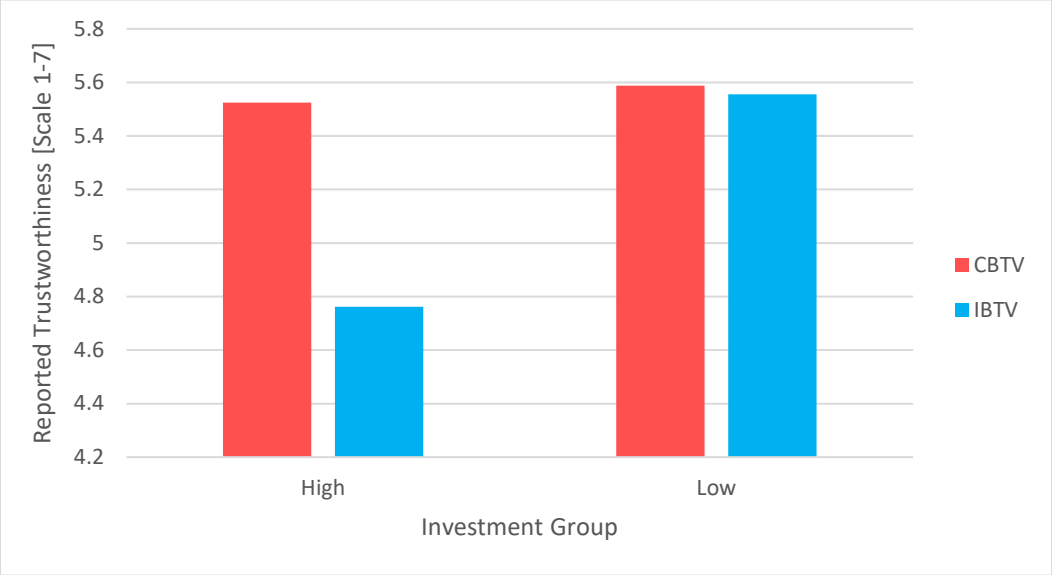

Figure 20. Reported Trustworthiness by Investment Group

Comparing average reliance behavior pre- and post-violation between high and low investment groups also yields interesting results. Descriptive statistics for average reliance behavior are outlined in Table 9, comparing numbers when all participants were included to numbers split by investment group.

Table 9. Descriptive Statistics for Average Reliance by Investment Group

Including All Participants				
	IBTV (VST 1)	IBTV (VST 2)	CBTV (VST 1)	CBTV (VST 2)
N	39		38	
Mean (SD) [CI_{95}]	0.584 (0.21) [0.51, 0.66]	0.583 (0.20) [0.52, 0.65]	0.562 (0.25) [0.49, 0.64]	0.595 (0.22) [0.53, 0.66]
High Investors				
N	21		21	
Mean (SD) [CI_{95}]	0.593 (0.18) [0.50, 0.68]	0.579 (0.18) [0.50, 0.66]	0.596 (0.23) [0.51, 0.69]	0.641 (0.20) [0.58, 0.73]
Low Investors				
N	18		17	
Mean (SD) [CI_{95}]	0.574 (0.26) [0.45, 0.70]	0.588 (0.23) [0.48, 0.70]	0.521 (0.27) [0.39, 0.65]	0.540 (0.24) [0.42, 0.66]

In the previous ANOVA including all participants, the relationship between average reliance behavior and condition was not significant (p=0.25). In a similar ANOVA excluding low investors, this relationship is still not significant, but potential differences start to emerge (p=0.16). Figure 21 shows the difference in average reliance behavior by condition between the two investment groups. While this relationship cannot be generalized with the current data, further studies with a-priori hypotheses and larger numbers of high investors could help clarify the existence or strength of any ties between reliance behavior and trust violation condition.

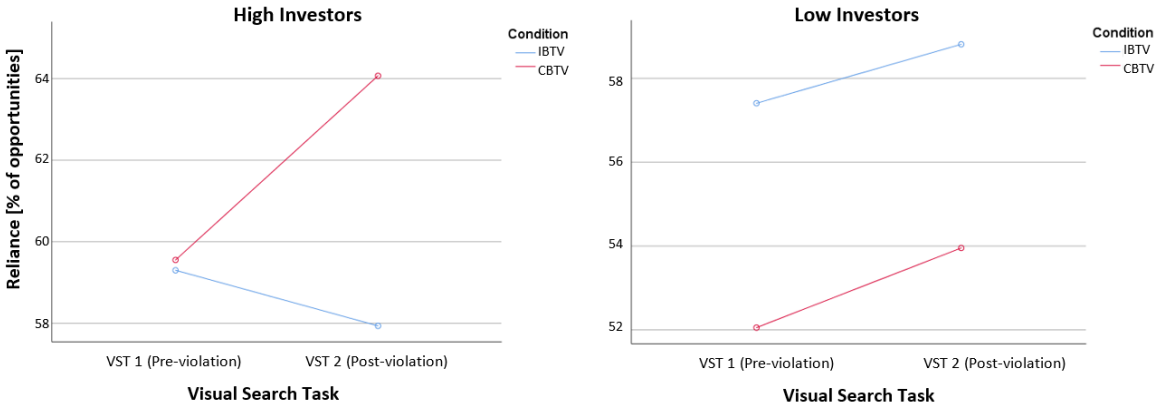

Figure 21. Average Reliance by Investment Group

Looking first at the high investors (left) in Figure 21, average reliance for both conditions was approximately equal during the first VST. After the trust violation, participants in the IBTV condition relied on BRIAN less, while those in the CBTV condition relied on BRIAN more. When considering just the low investors (right) in Figure 21, average reliance behavior post-violation increased for both conditions. Once again this seems to indicate that participants in the IBTV condition who invested (and lost) at least 50 percent of their points felt more of a trust violation than those who invested less than 50 percent.

This behavior also again seems to suggest that participants in the CBTV condition either felt no trust violation or simply reacted differently because of the different trust violation. It is also possible that reliance behavior is more strongly related to trust when violations are integrity-based than when violations are competence-based.

In reporting BRIAN's performance, there was a trend towards significance (p=0.075) in which high investors reported BRIAN's performance higher than low investors did, regardless of violation condition, $F(1, 72)=3.25$, $p=0.075$, $\eta_p^2=0.043$.

Table 10 delineates BRIAN's average performance as reported by participants in each investment group, with one participant excluded due to missing data for this question.

Table 10. Descriptive Statistics for Reported Performance by Investment Group

	High Investors		Low Investors
N	41		35
Mean (SD) [CI_{95}]	87.05 (5.65) [84.11, 89.99]		83.14 (12.32) [79.94, 86.31]

Figure 22 shows the relative difference in reported performance between the high and low investment groups, regardless of trust violation condition.

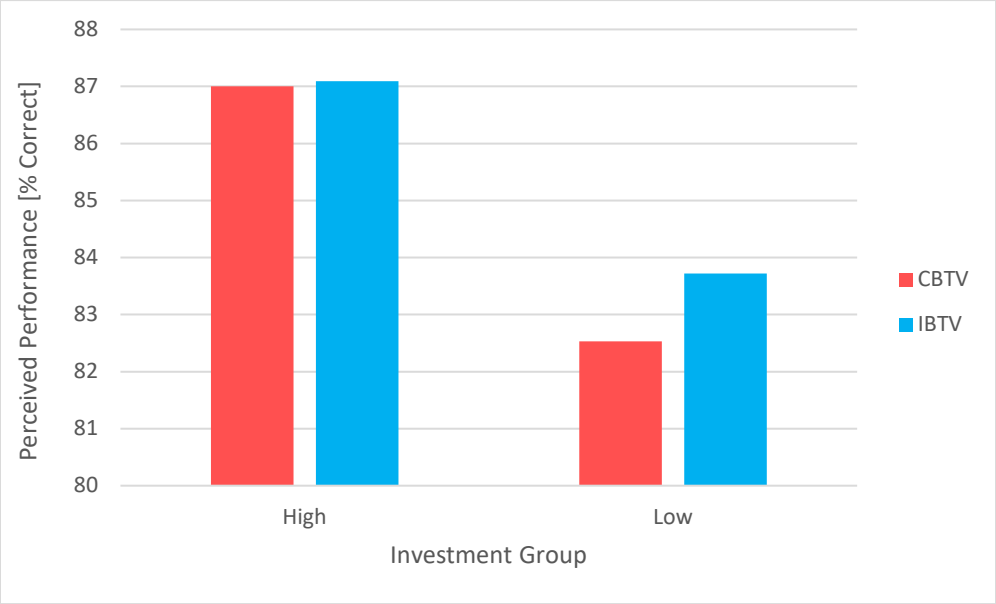

Figure 22. Reported Performance by Investment Group

Personal investment behavior in the post-violation TIG could also be influenced by how many points were invested and lost during the pre-violation TIG, so this was compared. A split ANOVA showed no effect of investment behavior by condition and by investment group (p=0.42). Table 11 compares the descriptive statistics between all participants grouped together and participants split into investment groups.

Table 11. Descriptive Statistics for Investment Behavior by Investment Group

All Participants				
	IBTV (VST 1)	IBTV (VST 2)	CBTV (VST 1)	CBTV (VST 2)
N	39		37	
Mean (SD) [CI_{95}]	0.56 (0.27) [0.49, 0.59]	0.46 (0.33) [0.37, 0.53]	0.58 (0.32) [0.49, 0.60]	0.56 (0.31) [0.46, 0.62]
High Investors				
N	21		20	
Mean (SD) [CI_{95}]	0.76 (0.19) [0.69, 0.83]	0.60 (0.33) [0.49, 0.71]	0.81 (0.19) [0.74, 0.89]	0.78 (0.21) [0.67, 0.90]
Low Investors				
N	18		17	
Mean (SD) [CI_{95}]	0.32 (0.11) [0.24, 0.39]	0.29 (0.25) [0.17, 0.41]	0.28 (0.15) [0.20, 0.36]	0.30 (0.16) [0.18, 0.42]

3. Swift Trust

Sometimes called “swift trust,” there is a natural tendency in some people to trust others without much in the way of history or knowledge about that person to base their trust on. Since this tendency could have a significant effect on self-reported trust, it seemed interesting to compare data from participants with this tendency to those without. The swift-trust tendency in participants was measured by how they answered a question about their likelihood to provide their full name to someone they met on a bus after about ten minutes of conversation. Possible answers to this question included (1) no chance, (2) depends on if they seem trustworthy, (3) probably, and (4) very likely. To be conservative, participants responding with any answer other than (1) were considered “Swift Trustors.”

An ANOVA comparing BRIAN's reported trustworthiness between violation conditions and the swift-trust tendency showed no effect (p=0.39) but still shows an interesting visual (Figure 23).

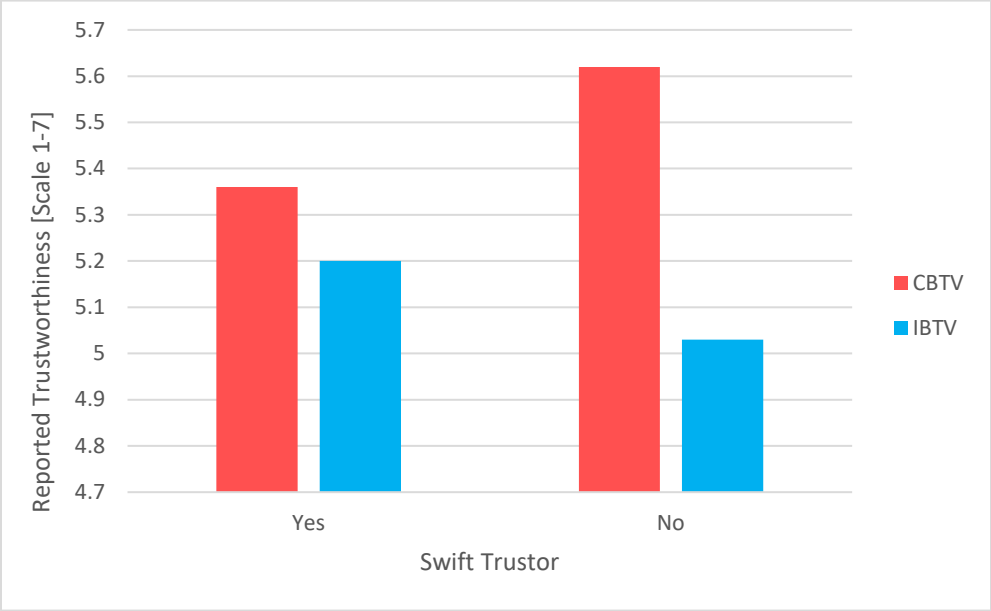

Figure 23. Reported Trustworthiness by Swift-Trust Tendency

Looking at Figure 23, it may be interesting for future research to see if there is a difference here. Participants without the swift-trust tendency seem to have a somewhat wider gap between their reported trustworthiness of BRIAN between conditions. Although statistical significance does not support it here, this behavior could still be a slight suggestion that those who do not trust easily are more likely to perceive a difference between IBTVs and CBTVs by automation than swift trustors do. While it is an interesting thought, conclusions cannot be drawn from this data as it is; more studies would need to be done in a-priori hypothesis manner in order to prove any such relationship.

4. Self-Efficacy

Since reliance can be seen as the relative difference between trust in one's own ability and trust in an aid [12], it was interesting to compare scores from the Self-Efficacy

questionnaire to reported trustworthiness ratings and also to overall reliance behavior. A regression analysis for each score separately found no significant correlation between self-efficacy and overall reliance behavior, p=0.77, or between self-efficacy and reported trustworthiness, p=0.40.

5. PANAS

It was also interesting to see whether trust violation condition caused any difference in participants' mood, comparing pre-violation and post-violation answers on the Positive and Negative Affect Scale (PANAS). A 2(IBTV, CBTV) X 2(Pre-violation, Post-violation) ANOVA was run for positive and negative affect separately. There was a main effect of positive affect, $F(1, 97)=696.17$, p=0.017, η_p^2=0.057; but no interaction effect of positive affect by condition (p=0.46).

These results show that participants had a higher average positive affect at the end of the experiment than at the beginning. Since positive affect is related to social activity and satisfaction [47], this could be an indication that interaction with BRIAN was enough like social interaction to boost this measure. It could also be accounted for by the two events preceding the second PANAS questionnaire, both of which had positive outcomes. During the second TIG, BRIAN returned half of the investment and half of the earnings, and during the extra points game, invested points were doubled. Figure 24 shows positive affect by condition, where the main effect can be seen in the consistent increase between the two questionnaires, regardless of condition.

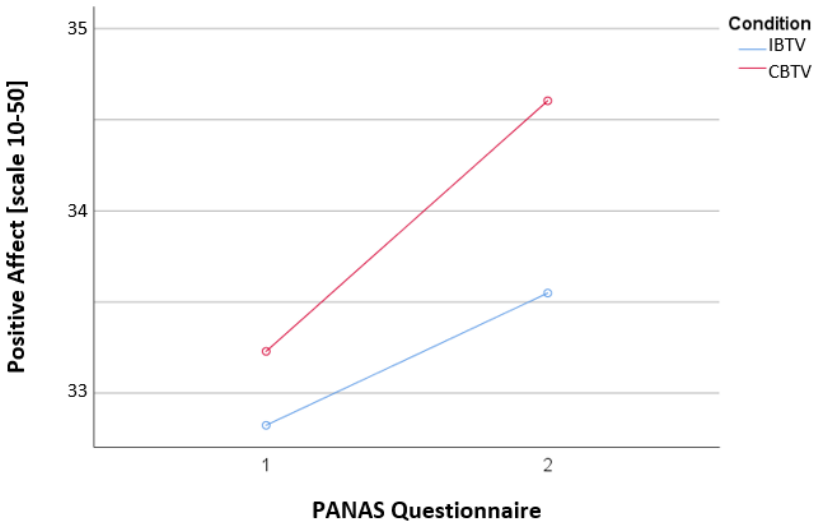

Figure 24. Positive Affect by Condition

There was also a main effect of negative affect, $F(1, 97)=696.17$, $p=0.017$, $\eta_p^2=0.153$; but no interaction effect of negative affect by condition ($p=0.65$). Negative affect is associated with stress and inability to cope well [47], so a decrease in average negative affect could mean that participants in general were not overly stressed by the trust violations. Figure 25 shows negative affect by condition, where the overall effect of negative affect can be seen in the decrease between questionnaires, regardless of condition.

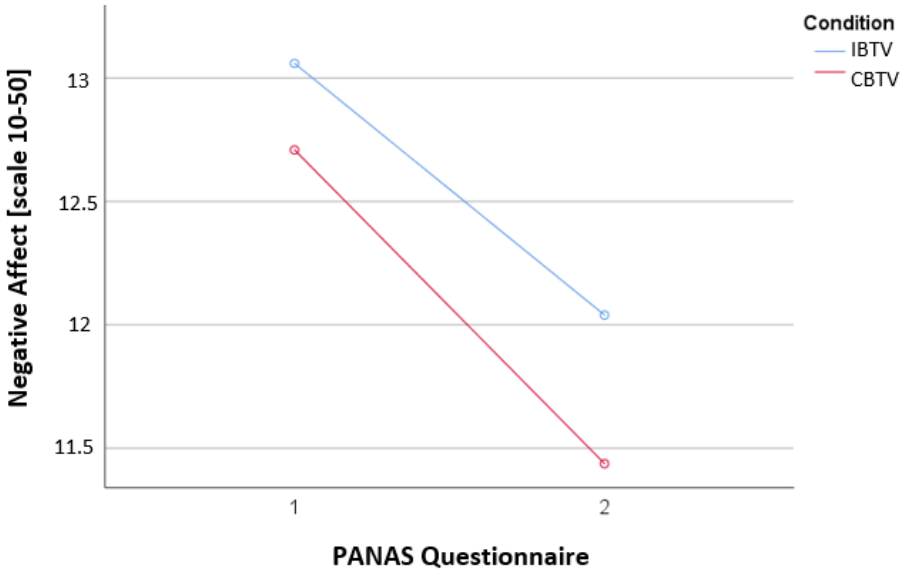

Figure 25. Negative Affect by Condition

V. CONCLUSIONS AND THOUGHTS FOR FUTURE RESEARCH

In this section, the most significant and interesting results are summarized and presented along with thoughts for future research into each area. A short discussion on what these results imply for HMT is also included.

A. SUMMARY OF RESULTS

The mean difference in reported trustworthiness was significantly lower in the IBTV condition as compared to the CBTV condition (p=0.026). Outliers found in this data set as previously discussed represent an important theoretical difference between violation conditions, and could represent a portion of the population for which automation committing an IBTV results in a severely low level of trust in automation. These outliers rated BRIAN as "not at all trustworthy," which in a human-machine teaming scenario might mean they would choose not to rely on the automation even when they should. Including the outliers in the analysis was a choice based on a sensitivity analysis, but it is possible that the small number could be due to individual differences, even though they are all in the IBTV condition. This experiment should be replicated or adapted and conducted with other populations to see whether IBTVs consistently have this same effect.

Participants averaged significantly faster response times during reliance decisions in the second VST than during the first (p<0.001). This is likely simply from practice at the task, but it is also possible that using the same pictures in the second round influenced response times. It is unclear whether trust in BRIAN had any effect on response times. If researchers replicate this experiment or perform something similar, it might be beneficial to use different pictures for the second VST. As long as the difficulty is equal between the two, this may yield a more accurate representation of time taken to make a reliance decision.

Personal investment behavior was moderately different between the IBTV and CBTV conditions (p=0.075), suggesting that participants had some reason to choose

lower investment amounts in the IBTV condition. While this could be related to trust, further research should attempt to clarify what drives this difference in behavior.

Personal investment behavior may be a better way to gauge trust levels than reliance behavior, if the goal is to measure trust without explicitly mentioning it. Investment trends between the TIGs are more in line with reported trust levels than reliance behavior for this experiment. Participants invested approximately equal percentages between violation conditions in the first TIG, and about 10 percent less for the IBTV condition than for the CBTV condition in the second TIG. This reflects reported trust levels, which were about 10 percent lower for IBTV participants than for CBTV participants. Reliance behavior on average showed equal reliance pre- and post-violation in the IBTV condition, and actually showed an increase in reliance for the CBTV condition, which does not mirror reported trust. If investment behavior actually does reflect reported trust better than reliance behavior does, future research should experiment with more investment opportunities spaced throughout a task. This may offer an opportunity to look at trust in a more dynamic way without letting participants know directly that trust is being measured.

Reliance behavior by iteration also differed from reported trust, showing no discernible trend other than a small, very interesting anomaly just after the trust violation in which reliance spiked for the IBTV condition and dipped for the CBTV condition. Future research should measure reliance by time or iteration again to see if this behavior just after violation is repeated or if it was simply a chance occurrence.

Future replications or adaptations of this study should also make the competency-based error less associable to the computer. Ensuring the error is associated *only* with BRIAN would remove the possibility that participants do not experience the CBTV and subsequently do not lose any trust in BRIAN.

Lastly, a few interesting things were noted during exploratory analyses, though smaller N values and lack of a-priori hypotheses reduced any statistical validity. One of the most interesting comparisons was between high and low investors, where participants investing and subsequently losing higher amounts of points displayed behavior that

differed more between violation conditions. It is possible that one must feel highly invested in order to truly experience an IBTV from automation, or at least to experience it differently than a CBTV. Not only were all four outliers for reported trust in the high-investment group, there were also interesting differences in looking at average reliance and perceived performance of the automation. Future research may yield more substantial results if participants are more personally invested in the task. Money or another prize more universally valued than made-up points should be used as the medium being lost through trust violations in order to get a more realistic response. Especially in the context of HMT in a military setting, most humans involved would have much more to lose through automation violating their trust than a few points.

Swift trust tendencies had an effect similar to initial investment amount, where participants who were less inclined to trust swiftly had greater differences in behavior between conditions. Future research should measure this tendency in participants and attempt to see whether this effect is replicated in similar experiments.

B. IMPLICATIONS FOR HMT

If the results of this study are taken to mean that automation committing IBTVs results in lower trust levels than the same automation committing CBTVs, care should be taken in design and implementation of HMT technology to avoid automation actions which could be perceived by human teammates as an IBTV. This is especially important if the outliers previously shown actually do represent a portion of the population who are more sensitive to this type of violation. Trust levels that low (rating BRIAN "not at all trustworthy") would certainly have a negative effect on performance of a human-machine team.

Another important finding of this study is that participants investing (and losing) more points had a more significant difference in behavior between conditions than those who did not invest as much. Considering that many human-machine teams will be working within military environments, the humans on these teams will likely be highly invested in their teammate—possibly even placing their lives in the hands of the automation. Since high investors in this study seemed to be the group most negatively

affected by the IBTV, the behavior they displayed is likely the type of behavior to be expected within military HMT.

While this experiment has resulted in some answers as to whether humans experience IBTVs by automation differently than CBTVs, it has also generated some more questions on the subject. The significant findings, interesting trends and correlations, and even the anomalies found in the process all offer seeds of thought for future research in this area. Results of this study and future replications or adaptations could provide vital information for use as humans and machines learn to become teammates. As a humble offering, it is hoped that this research stands as a solid first step towards looking at integrity-based trust violations within human-machine teaming.

LIST OF REFERENCES

[1] J. Ellman, L. Samp and G. Coll, “Assessing the third offset strategy,” Center for Strategic and International Studies, Washington, DC, USA, 2017.

[2] UnderSecretary of Defense Acquisition, Technology & Logistics, “Unmanned systems integrated roadmap FY2013-2038,” Washington, DC, USA, 2014.

[3] UnderSecretary of Defense Acquisition, Technology & Logistics, “Report of the Defense Science Board summer study on autonomy,” Washington, DC, USA, 2016.

[4] M. McFarland, “People want automation but are afraid of machines, study show *CNN Tech*, Nov. 14, 2016. [Online]. Available: http://money.cnn.com/2016/11/1 technology/machines-technology-love-fear-study/index.html

[5] T. B. Sheridan and W. L. Verplank, “Human and computer control of undersea teleoperators,” Mass. Inst. of Tech., Cambridge, MA, USA, 1978.

[6] M. R. Endsley and E. O. Kiris, “The out-of-the-loop performance problem and level of control in automation,” *Human Factors,* vol. 37, no. 2, pp. 381–394, Jun.1995.

[7] J. D. Lee and K. A. See, “Trust in automation: Designing for appropriate reliance,” *Human Factors,* vol. 46, no. 1, pp. 50–80, 2004.

[8] R. R. Hoffman, “A taxonomy of emergent trusting in the human-machine relationship,” in *Cognitive Systems Engineering: The Future for a Changing World*, P. J. Smith and R. R. Hoffman, Eds. Boca Raton, FL, USA: Talyor & Francis, 2017, pp. 137–163.

[9] K. A. Hoff and M. Bashir, “Trust in automation: Integrating empirical evidence factors that influence trust,” *Human Factors,* vol. 57, no. 3, pp. 407–434, May 2015.

[10] J. Lee and N. Moray, “Trust, control strategies and allocation of function in human-machine systems,” *Ergonomics,* vol. 35, no. 10, pp. 1243–1270, 1992. [Online]. doi: 10.1080/00140139208967392

[11] T. Warren, “Amazon explains how Alexa recorded a private conversation and se it to another user,” May 24, 2018. [Online]. Available: https://www.theverge.cor /2018/5/24/17391898/amazon-alexa-private-conversation-recording-explanation

[12] K. van Dongen and P.-P. van Maanen, “A framework for explaining reliance on decision aids,” *Int. Journal of Human-Computer Studies*, vol. 71, pp. 410–424, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.ijhcs.2012.10.018

[13] B. M. Muir, “Trust between humans and machines, and the design of decision aids,” *Int. Journal of Man-Machine Studies*, vol. 27, no. 5-6, pp. 527–539, Nov. 12, 1987.

[14] R. Parasuraman and V. Riley, “Humans and automation: Use, misuse, disuse, abuse,” *Human Factors*, vol. 39, no. 2, pp. 230–253, Jun. 1997.

[15] M. P. O’Neil, “Development of a human systems integration framework for Coa Guard acquisition,” M.S. thesis, Dept. of Ops. Research, NPS, Monterey, CA, USA, 2014. [Online]. Available: https://calhoun.nps.edu/handle/10945/42696

[16] “Definitions of HSI,” class notes for Introduction to Human Systems Integration Dept. of Op. Analysis, Naval Postgraduate School, Monterey, CA, USA, fall 201

[17] A. Hartmans, “Google unveiled a new 'experiment' that will impersonate a huma to make restaurant reservations for you over the phone,” *Business Insider, May* 8 2018. [Online]. Available: http://www.businessinsider.com/google-assistant-makes-phone-calls-schedules-appointments-reservations-google-duplex-google-2018-5

[18] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser and Parasuraman, “A meta-analysis of factors affecting trust in human-robot interaction,” *Human Factors*, vol. 53, no. 5, pp. 517–527, Oct. 2011. [Online]. d 10.1177/0018720811417254

[19] K. E. Shaefer, J. Y. C. Chen, J. L. Szalma and P. A. Hancock, “A meta-analysis factors influencing the development of trust in automation: Implications for understanding autonomy in future systems,” *Human Factors*, vol. 58, no. 3, pp. 377–400, May 2016. [Online]. doi: 10.1177/0018720816634228

[20] G. Bansal and F. M. Zahedi, “Trust violation and repair: The information privac perspective,” *Decision Support Systems*, vol. 71, pp. 62–77, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2015.01.009

[21] E. Paeng, J. Wu and J. C. Boerkoel Jr., “Human-robot trust and cooperation through a game theoretic framework,” in *Proc. of the 30th AAAI Conf. on Artificial Intelligence*, 2016, pp. 4246–4247.

[22] A. Chavaillaz, D. Wastell and J. Sauer, “System reliability, performance and trust in adaptable automation,” *Applied Ergonomics*, vol. 52, pp. 333–342, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.apergo.2015.07.012

[23] E. J. de Visser, S. S. Monfort, R. McKendrick, M. A. B. Smith, P. E. McKnight, F. Krueger and R. Parasuraman, “Almost human: Anthropomorphism increases trust resilience in cognitive agents,” *Journal of Expl. Psychology: Applied*, vol. 22, no. 3, pp. 331-349, 2016. [Online]. Available: http://dx.doi.org/10.1037/xap0000092

[24] P. H. Kim, C. D. Cooper, K. T. Dirks and D. L. Ferrin, “Repairing trust with individuals vs. groups,” *Orgl. Behavior and Human Decision Processes*, vol. 120, pp. 1-14, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.obhdp.2012.08.004

[25] P. H. Kim, D. L. Ferrin, C. D. Cooper and K. T. Dirks, “Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- vs. integrity-based trust violations,” *Journal of Appd. Psychology*, vol. 89, no. 1, pp. 104-118, 2004. [Online]. doi: 10.1037/0021-9010.89.1.104

[26] P. H. Kim, K. T. Dirks, C. D. Cooper and D. L. Ferrin, “When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation,” *Orgl. Behavior and Human Decision Processes*, vol. 99, pp. 49-65, 2006. [Online]. doi: 10.1016/j.obhdp.2005.07.002

[27] D. L. Ferrin, P. H. Kim, C. D. Cooper and K. T. Dirks, “Silence speaks volumes: The effectiveness of reticence in comparison to apology and denial for responding to integrity- and competence-based trust violations,” *Journal of Applied Psych.*, vol. 92, no. 4, pp. 893-908, 2007. [Online]. doi:10.1037/0021-9010.92.4.893

[28] A. Proyas, Director, *I, Robot.* [Film]. 2004.

[29] J. Everett, D. Pizarro and M. Crockett, “Why are we reluctant to trust robots?,” *The Guardian*, Apr. 24, 2017. [Online]. Available: https://www.theguardian.com/science/head-quarters/2017/apr/24/why-are-we-reluctant-to-trust-robots

[30] J. Gmoser and C. Weller, “A robot that once said it would 'destroy humans' just became the first robot citizen,” *Business Insider*, Oct 26, 2017. [Online]. http://www.businessinsider.com/sophia-humanoid-robot-ai-citizen-saudi-arabia-future-investment-initiative-interview-2017-10

[31] S. Linning, “Could models be replaced by ROBOTS? Sophia the humanoid dazzles on her first British fashion magazine cover (but leaves readers feeling 'freaked out'),” *Daily Mail*, Jan. 25, 2018. [Online]. Available: http://www.dailymail.co.uk/femail/article-5310823/Sophia-robot-appears-Stylist-magazine-cover.html#ixzz5H2N9WC7g

[32] “Tonight showbotics: Jimmy meets Sophia the human-like robot,” Hanson Robotics, April 5, 2017. [Online]. Available: http://www.hansonrobotics.com/robot/sophia/

[33] “Trust in autonomy forhHuman machine teaming,” Federal Business Opportunities, May 7, 2015. [Online]. Available: https://www.fbo.gov/index?s=opportunity&mode=form&id=c69b617294a123346363a15302513711&tab=core&_cview=1

[34] J. Knefel, “The Air Force wants you to trust robots--should you?,” *Vocativ Tech by Scientific American*, Aug 25, 2015. [Online]. Available: https://www.scientificamerican.com/article/the-air-force-wants-you-to-trust-robots-should-you/

[35] D. S. Nau, “Artificial intelligence and automation,” in *Springer Handbook of Automation*, S. Y. Nof, Ed. New York, NY, USA: Springer Publishing Company, 2009, pp. 249–268.

[36] “Volunteer scheduling made easy,” Signup Schedule, 2012. [Online]. Available: https://signupschedule.com/

[37] B. Reeves and C. Nass, “Politeness,” in *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.* New York, NY, USA: Cambridge University Press, 1996, pp. 19–36.

[38] T. H. Chung, “ARSENL,” Naval Postgraduate School Wiki, November 20, 2015. [Online]. Available: https://wiki.nps.edu/display/~thchung/ARSENL

[39] "Presentation," Neurobehavioral Systems 2018. [Online]. Available: https://www.neurobs.com/menu_presentation/menu_features/features_overview

[40] J. C. Walliser, "Trust in automated systems: the effect of automation level on trust calibration," M.S. thesis, Dept. of Ops. Research, NPS, Monterey, CA, USA, 2011. [Online]. Available: https://calhoun.nps.edu/handle/10945/5628

[41] S. Rice and D. Keller, "Automation reliance under time pressure," *Cognitive Technology,* vol. 14, no. 1, pp. 36–44, 2009.

[42] S. R. Dixon and C. D. Wickens, "Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload," *Human Factors,* vol. 48, no. 3, pp. 474–486, 2006.

[43] M. T. Dzindolet, H. P. Beck, L. G. Pierce and L. A. Dawe, "A framework of automation use," U.S. Army Research Lab., Aberdeen Proving Ground, MD, USA, 2001.

[44] J. Berg, J. Dickhaut and K. McCabe, "Trust, reciprocity, and social history," *Games and Economic Behavior,* vol. 10, no. 1, pp. 122–142, 1995.

[45] N. D. Johnson and A. A. Mislin, "Trust games: A meta-analysis," *Jrnl. of Economic Psych.,* vol. 32, pp. 865–889, 2011. [Online]. doi: 10.1016/j.joep.2011.05.007

[46] S. M. Merritt, "Affective processes in human–automation interactions," *Human Factors,* vol. 53, no. 4, pp. 356–370, Aug. 2011. [Online]. doi: 10.1177/001872081141191

[47] D. Watson, L. A. Clark and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," *Journal of Personality and Social Psyc.,* vol. 54, no. 6, pp. 1063–1070, 1988.

[48] K. van Dongen and P.-P. van Maanen, "Under-reliance on the decision aid: A difference in calibration and attribution between self and aid," in *Proc. of the Human Factors and Ergonomics Society Annual Mtg.,* Amsterdam, 2006.

[49] "Emotion measures: NIH toolbox emotion battery," Health Measures, 2018. [Online]. Available: http://www.healthmeasures.net/explore-measurement-systems/nih-toolbox/intro-to-nih-toolbox/emotion

[50] IBM Corp., *IBM SPSS Statistics for Windows, Version 25.0.,* Armonk, NY: IBM Corp., 2017.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California