MODERATORS OF TRUST AND RELIANCE ACROSS MULTIPLE DECISION AIDS

by

JENNIFER MARIE ROSS
B.A. University of North Carolina in Asheville, 2002
M.S. University of Central Florida, 2004

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Psychology
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Spring Term
2008

Major Professors: P.A. Hancock
James L. Szalma

ABSTRACT

The present work examines whether user's trust of and reliance on automation, were affected by the manipulations of user's perception of the responding agent. These manipulations included agent reliability, agent type, and failure salience. Previous work has shown that automation is not uniformly beneficial; problems can occur because operators fail to rely upon automation appropriately, by either misuse (overreliance) or disuse (underreliance). This is because operators often face difficulties in understanding how to combine their judgment with that of an automated aid. This difficulty is especially prevalent in complex tasks in which users rely heavily on automation to reduce their workload and improve task performance. However, when users rely on automation heavily they often fail to monitor the system effectively (i.e., they lose situation awareness – a form of misuse). However, if an operator realizes a system is imperfect and fails, they may subsequently lose trust in the system leading to underreliance. In the present studies, it was hypothesized that in a dual-aid environment *poor* reliability in one aid would impact trust and reliance levels in a companion *better* aid, but that this relationship is dependent upon the perceived aid type and the noticeability of the errors made. Simulations of a computer-based search-and-rescue scenario, employing uninhabited/unmanned ground vehicles (UGVs) searching a commercial office building for critical signals, were used to investigate these hypotheses. Results demonstrated that participants were able to adjust their reliance and trust on automated teammates depending on the teammate's actual reliability levels. However, as hypothesized there was a biasing effect among mixed-reliability aids for trust and reliance. That is, when operators worked with two agents of mixed-reliability, their perception of how reliable and to what degree they relied on the aid was effected by the reliability of a current aid.

Additionally, the magnitude and direction of how trust and reliance were biased was contingent upon agent type (i.e., 'what' the agents were: two humans, two similar robotic agents, or two dissimilar robot agents). Finally, the type of agent an operator believed they were operating with significantly impacted their temporal reliance (i.e., reliance following an automation failure). Such that, operators were less likely to agree with a recommendation from a human teammate, after that teammate had made an obvious error, than with a robotic agent that had made the *same* obvious error. These results demonstrate that people are able to distinguish when an agent is performing well but that there are genuine differences in how operators respond to agents of mixed or same abilities and to errors by fellow human observers or robotic teammates. The overall goal of this research was to develop a better understanding how the aforementioned factors affect users' trust in automation so that system interfaces can be designed to facilitate users' calibration of their trust in automated aids, thus leading to improved coordination of human-automation performance. These findings have significant implications to many real-world systems in which human operators monitor the recommendations of multiple other human and/or machine systems.

To
my parents,
without whom none of this would be possible.

Patricia Brake-Ludwig

And

Aubrey Clement Ross

ACKNOWLEDGMENTS

I wish to thank my committee members who were more than generous with their expertise and precious time. A special thanks to Drs. Peter Hancock and James Szalma, my committee chairmen for their countless hours of reflecting, reading, encouraging, and most of all patience throughout the entire process. Thank you Dr. John Barnett, Dr. Moustapha Mouloua, and Dr. Valerie Sims for agreeing to serve on my committee, your advice has been invaluable at every step of this process.

I would also like to acknowledge and thank the Army Research Laboratory, the Institute for Simulation and Training, and the Department of Psychology here at UCF for allowing me to conduct my research and providing a great deal of assistance. Special thanks go to my coworkers at ARI and the administrative staff in the department of psychology, particularly Lisa Mindak, for their continued support and advice in completing this sizeable project.

Finally I would like to thank the professors, fellow graduate students, and my undergraduate research assistants who assisted me with this project. Their excitement and willingness to provide feedback made the completion of this research an enjoyable experience.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

ATS	Anthropomorphic Tendencies Scale
CPRS	Complacency Potential Rating Scale
ES	Effect Size
IED	Improvised Explosive Device
IRT	Item Response Theory
ISI	Inter-Stimulus Interval
ITS	Interpersonal Trust Scale
NASA-TLX	NASA Task Load Index
RT	Response/Reaction Time
SDT	Signal Detection Theory
UAV	Uninhabited/Unmanned Arial Vehicle
UGV	Uninhabited/Unmanned Ground Vehicle
UV	Uninhabited/Unmanned Vehicle

INTRODUCTION

Advances in modern technology are increasing the ability of human beings to travel and communicate, as well as automate their work. The development of complex robotics and mathematical algorithms to guide artificial intelligence allows for the technology to permit non-human agents to simulate and hence automate many human intellectual functions. The capacity of these electronic avatars is growing as a function of increasing computational capacity (see Moore, 1965), granting automation functions that include actively selecting data, transforming information, making decisions, and associated output processes (Beck, Dzindolet, & Pierce, 2002; Lee & See, 2004). Such advances have revolutionized the role of semi-autonomous and autonomous agents in military, transportation, medical environments, and a spectrum of other applied realms.

The use of robotic-agents offers a wide range of advantages, including increased safety for human operators. With the application of a non-human agent with a remote operator, the human becomes one-step removed from the dangerous situation (e.g., gathering reconnaissance information in a combat environment). This allows Unmanned Vehicles (UVs) to act "fearlessly" in battle, operate in areas contaminated with biotoxins or radiation, and removes the need for expensive on-board environmental systems (Mouloua, Gilson, & Hancock, 2003). Further, a large potential benefit for the military and industry is that employing autonomous and semi-autonomous agents reduces personnel requirements. A hypothetical example of this benefit would be a single operator controlling multiple UVs, perhaps hundreds of Unmanned Arial Vehicles (UAVs) to a single operator (Hancock, Mouloua, Gilson, Szalma, & Oron-Gilad, 2007; Squire, Trafton, & Parasuraman, 2006). This can be compared to traditional manned vehicles

which may each require a separate individual operator or in some cases multiple operators (e.g., M2 Bradley Fighting Vehicle System requires a crew of 3: the commander, gunner, and driver; Global Security, 2007). Additionally, one of the primary uses of automation is to make repetitive or detailed tasks easier (e.g., using automated speed dial rather than dialing a phone number one digit at a time; Wiener, 1988).

However, automation is often applied haphazardly without regard to the intricacies of the human-automation interaction. This can often lead to negative consequences, such as, operator complacency (Chappell, 1997; Morgan, Herschler, Wiener, & Salas, 1993), increased user monitoring requirements (Kantowitz & Campbell, 1996), and degeneration of operator manual skills (McClumpha, James, Green, & Belyavin, 1991). One factor that has been shown to strongly affect how an operator will interact with a system is operator system trust (i.e., one's confidence in an automated system). If an operator has too little trust in a system they may fail to use the automated system, which in effect negates the potential of the automated system to benefit operator performance (Parasuraman & Riley, 1997). Automated systems are often developed at great cost to the organization, but operator trust is essential to ensure that they are utilized. On the other hand, if an operator overtrusts a system this may lead to complacency and automation bias (Barnett, 2000).

As the goal of automation is to extend human capabilities, often by using multiple machine systems, it becomes imperative to examine whether individuals are able to compartmentalize their trust of individual automated systems or if there is a blending of trust levels across systems. That is, could a soldier working with a network of UVs observe an error on one of the robotic systems and still respond in an unbiased manner to the other vehicles, or would this error then predispose the soldier to lose trust in the other systems (i.e., trigger disuse

across all systems)? To ensure the future of successful collaboration between humans and machines, it is imperative that designers know in what way operators are able to calibrate their actions with those of 'intelligent' machines (Beck, Dzindolet, & Pierce, 2002) and in what ways their calibrations are influenced by defective agents.

Organization of the Thesis

This thesis proposes that complex (i.e., dual-aid) environments will encounter carry-over bias between mixed reliability aids. However, it is believed that this effect will be influenced by the type of the agent (i.e., whether operators believe they are working with other humans or robotic aids). This problem is examined first by setting the theoretical and empirical grounding for the following studies and subsequently explicating the methodology ad results of a series of four experiments.

The first and second experiments focus on validation and construction of the experimental test bed. Drawing from methods from psychometrics, these studies sought to minimize potential sources of error and variance associated with the task itself. The results from study one determined the pace of the task while study two was critical for determining stimuli error salience (i.e., the difficulty of the trials). Thus the goal of the first two experiments was to employ a simulation of a computerized search-and-rescue scenario without a decision-aid to determine required trial duration, trial inter-stimulus interval (ISI), and analyze stimuli difficulty.

The third experiment was designed to extend experiment one and two, by applying a single automated-decision aid to the created search-and-rescue test bed and varying the reliability

of the aid. The goal of the third experiment was to examine trust and reliance levels on the automated aid and to determine appropriate *low* and *high* levels of reliability.

The fourth experiment extended this information one step further by adding a second decision-aid. This final experiment examined operator trust and reliance on automatic decision aids when working with multiple agents. This experiment provides empirical evidence on the influence and possible biasing effects of monitoring multiple decision-aid agents of varying reliability, agent type, and error salience. The overall goal of this research was to develop a better understanding how the aforementioned factors affect users' trust in automation so that system interfaces can be designed to facilitate users' calibration of their trust in automated aids, thus leading to improved coordination of human-automation performance.

REVIEW OF LITERATURE

Automation has often been touted as a panacea for improving how human beings interact with their environment. Indeed, automation has given us modern day assembly lines, automobile cruise control, aircraft autopilot features, and even semi-autonomous vacuum cleaners. As almost any task can be, and often is, automated, we find that automation is becoming increasingly prevalent in modern day society. With this progress the shift from operators serving as active controllers (i.e., directly involved with the system) to supervisory controllers (i.e., indirect management of a system) has become more common (Lee & Moray, 1994). Accompanying this evolution of the operator from their original role, there is a need to explore the components that influence effective cooperation between operators and semi-autonomous agents. One particular area of study is that of when the use of automation backfires (Parasuraman & Riley, 1997).

Problems with Automation

There can be potentially harmful consequences of automation when users fail to rely upon automation appropriately, through either misuse (overreliance) or disuse (underreliance; Parasuraman & Riley, 1997). That is, human judges may face difficulty in understanding how to calibrate their judgment with that of an automated aid (Bass & Pritchett, 2006).

Misuse

Individuals may misuse automation by over relying on automation when a manual alternative would have achieved a better end (Mouloua, Gilson, & Koonce, 1997). Operators

who have high levels of trust in an automated system may assume, often incorrectly, that it is highly reliable and requires little to no monitoring (Parasuraman, Molloy, & Singh, 1993). Overdependence on automated systems has been related to skill degeneration or inattention in the lab; which may result in more serious consequences in the real world (Young & Stanton, 2001). For example, pilots trusting the ability of their autopilot, failed to intervene and take manual control even as the autopilot crashed the Airbus A320 they were flying (Lee & See, 2004). In another instance, an automated navigation system malfunctioned and the crew failed to intervene, allowing the *Royal Majesty* cruise ship to drift off course for 24 hours before it ran aground (National Transportation Safety Board, 1997). Misuse of automation often occurs in cases where people have attributed greater intelligence to the automation than it actually possesses. Bergeron and Hinton (1985) pointed out that "*the pilot thinks of the autopilot as a copilot and expects it to think for itself. He allows himself to become completely engrossed in other tasks once the autopilot is set. Hence, he is frequently late in resetting new functions, or he may become confused as to exactly where he is in the approach*" (p. 145). Trusting automation to function on its own without supervision is a flawed approach. Automation is inherently limited to what it was programmed to do (i.e., dumb and dutiful) which may not always be desirable or even expected by the operator (Wiener, 1988; Sheridan, 2002). These 'automation surprises' occur when the system is behaving according to its programmed specifications, yet in a way that is contrary to what the operator expects or desires (Young & Stanton, 2001).

Disuse

On the other hand, disuse occurs when users under-utilize automation by manually performing a task that could best be done by automation. For instance, some operators rejected automated controllers in paper mills, undermining the potential benefits of competent and reliable automation (Zuboff, 1988). In one form of disuse automation may hinder performance by raising workload levels. This can occur when operators perform the task manually but then check the automation anyway thereby adding to their workload (Bainbridge, 1983). Indeed, unwillingness of workers to accept effective technology is frequently cited as an impediment to improving worker productivity (DiBello, 2001).

In the case of automated internet commerce technology trust becomes a critical factor in determining if potential customers are willing to submit personal information (e.g., credit card numbers) to a commercial website. Research by Karvonen and Parkkinen (2001), found that trust was a necessary factor in order to indulge in the risk to personal privacy (i.e., identity theft). This research points out how the use of automation entails a certain amount of accepted vulnerability by the user. In a separate study by de Vries and colleagues (2003), in which participants wagered study credits on the likelihood of accurate automation performance, it was found that higher risk was correlated with higher ratings of system trust. With distrust users are less willing to take risk and in the case of internet commerce they withdraw from the website and the company loses business.

One of the most dangerous forms of disuse is that of the 'cry wolf effect' (Bliss, 1993), in which case a user ignores warning signals that have previously signaled a false alarm (e.g., a fire alarm that has previously only been yearly tests). In his book *Set Phasers on Stun* (1993), Casey

points out a particularly ingenious use of automation disuse, in which a prisoner deliberately sets off a motion detector alarm during his escape. While this might seem counter-intuitive the prisoner (a very astute fellow) was well aware of the previous high false alarm rates and the subsequent distrust of the alarm by the guards. Thus, even though the alarm correctly signaled a prisoner's break out, the guards responded slowly to the alarm believing it was merely another automation error, allowing the prisoner to escape!

Calibrated Reliance

Misuse and disuse are two examples of inappropriate reliance on automation that can compromise safety, profitability, and performance. Ideal reliance in an automated system requires discriminating operators who can determine a proper calibration between their own and system performance; that is, they know when to and when not to depend upon automation (see Figure 1).

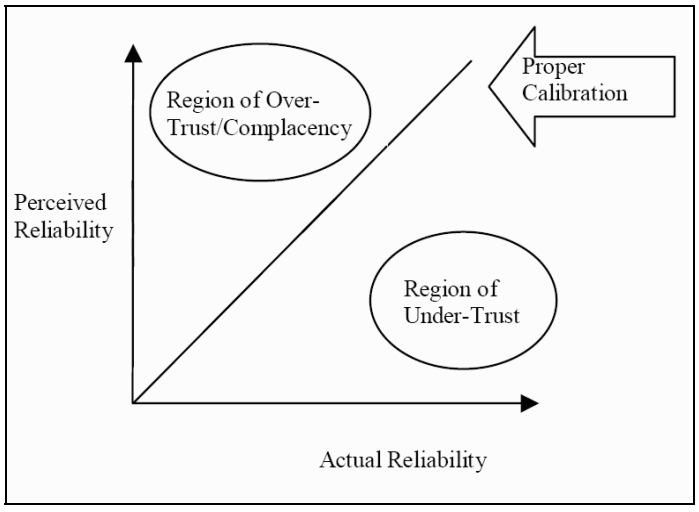


Figure 1. Reliability calibration (Gempler & Wickens, 1998).

When operators place unquestioned trust in the perceived reliability of the automation, that is not appropriate given the actual reliability of the automation, they fall into the region of over-trust/complacency. In this case operators often fail to monitor the automation adequately because they exhibit excessively high confidence in the system. A lower level of trust in this case would be more appropriate. On the other hand when operators fall into the region of under-trust, they perceive the reliability of the automation as lower than the actual reliability of the system. In this case their lack of trust in the automation leads to disuse. Between these two extremes is the region of proper trust calibration, in which the operator trusts the automation enough to use it when it is helpful but distrusts it enough to monitor it for proper operation (Barnett, 2000).

Human-Automation Interaction

To appreciate the impact of trust on properly calibrating user reliance, an understanding of how humans and machines work together is needed. While, neither humans nor machines are infallible, exploiting the strengths of each can lead to a joint performance that is higher than either's individual performance alone (Young & Stanton, 2001). That is, the hybrid human-automated system should exhibit superior performance compared to the human alone (Hancock & Parasuraman, 1992; Hancock, Parasuraman, & Byrne, 1996). An optimally calibrated interaction involves a human user who knows when to heed or ignore an aid's suggestion (Bass & Pritchett, 2006). The question then becomes what processes do people use to determine when to rely on themselves or when to rely on an automated aid? Several studies have established that humans actually respond socially to technology, and reactions to computers can be similar to reactions to human collaborators (Muir & Moray, 1996; Reeves & Nass, 1996).

Social Response to Technology

Research suggests that misuse and disuse of automation may depend on certain feelings and attitudes that operators hold. These feelings and attitudes may be miscalibrated and distort one's perception of the automation. One particular factor that past research has shown to guide reliance is trust (Halpin, Johnson, & Thornberr, 1973; Muir, 1988; Sheridan & Hennessy, 1984).

Trust

Trust is a basic feature of all social situations that demand cooperation and interdependence (Corritore, Kracher, & Wiedenbeck, 2001). This social psychological concept is particularly important for understanding human-automation partnerships, and can be defined as the belief that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability (De Vries, Midden, & Bouwhuis, 2003). In this definition, an agent can be any entity that actively interacts with the environment on behalf of the individual (e.g., another human being, an automated aid, etc.). Research has shown that just as trust mediates interactions between people (Deutsch, 1958, 1960; Rempel, Holmes, & Zanna, 1985; Ross & LaCroix, 1996; Rotter, 1967), it has also been established that trust mediates the relationship between people and automation (Lee & Moray, 1992; Lee & See, 2004; Lewandowsky, Munday, & Tan, 2000; Muir, 1994; Seong, Bisantz, & Gattie, 2006; Sheridan & Hennessy, 1984). Indeed, in a series of empirical studies by Jian, Bisantz, and Drury (2000) it was demonstrated that people do not perceive concepts of trust differently across general trust, human-human trust, and human-automation trust.

One model of trust is Barber's (1983) taxonomy of trust which divides trust into three specific expectations: persistence, technical competence, and fiduciary responsibility (See Table 1). Barber defines persistence as the foundation for trust. Persistence allows for trustors to form the expectation that something will work in a predictable way; this reduces the complexity of a task by limiting the possible outcomes. Without persistence an operator would have to consider every possible positive and negative outcome at each step of the interaction. Of equal importance is the notion of technical competence. Technical competence reflects the ability of the teammate in regards to technical facility and expert knowledge. Indeed, an individual may increase or decrease vigilance depending upon the perceived competency of a teammate (Mosier & Skitka, 1998). Perceived technical competence may vary depending on whether a task is routine or unusual. For instance, an operator may trust automation to be technically competent to perform a routine task, but switch to manual control for more difficult or unusual tasks. The third dimension of trust in Barber's model is fiduciary responsibility. Fiduciary responsibility refers to moral and social obligations that people have to hold the interest of others above their own, and has been contended to be irrelevant to the human-automation interaction (Uggirala, Gramopadhye, Melloy, & Toler, 2004).

Table 1. Barber's Taxonomy of Trust (recreated from Uggirala, Gramopadhye, Melloy, & Toler, 2004).

Expectation	Impact	Description
Persistence	Provides basis for all other forms of trust.	The foundation of trust that establishes a constancy in the fundamental moral and natural laws.
Technical Competence	Supports expectations of future performance based on capabilities, knowledge or expertise.	The ability of the other partner to produce consistent and desirable routine performance, technical facility, and expert knowledge.
Fiduciary Responsibility	Extends the idea of trust beyond that based on performance to one based on moral obligations and intentions	The expectation that people have moral and social obligations to hold the interest of others above their own.

Another three-stage model of trust was proposed by Rempel, Holmes, and Zanna (1985); this model is based on a hierarchical model of trust, and contends that certain factors of trust may change with time and increasing emotional investment (See Table 2). In this model the first stage of trust is predictability, which is judged by the operator as the consistency and desirability of the machines recurrent behavior (i.e., the confidence they have in their ability to predict future behaviors). Predictability is drawn from the actual predictability of the machine's behavior, the operator's ability to estimate the predictability of the machine's behavior, and the stability of the environment in which the system operates (Uggirala et al., 2004). The more variable a machine's performance the lower its predictability. As the relationship progresses an operator may enter the second stage of trust: dependability. Dependability is an understanding of the stable dispositions that guide a partner's behavior. In terms of monitoring machine systems, or human systems for that matter, this factor is dependent on positive assessments of predictability in the realm of personal vulnerability and conflicts of interest. The final stage is that of faith, in faith an operator summarizes past predictability and dependability experiences to summarize them into a belief in

how the machine will operate in unknown future situations. In order to develop faith in any particular machine, a human operator must have extensive experience with the system to let faith develop.

Table 2. Rempel, Holmes, and Zanna's (1985) model of trust.

Stage of Trust	Description
Stage 1: Predictability	Judged by actual predictability (variance) of the system, operator's ability to estimate that predictability, and environmental factors.
Stage 2: Dependability	Related to the reliability of the system over time.
Stage 3: Faith	Based on extensive past experiences with the system. Summarize past experiences based on predictability and dependability.

Both Barber (1983) and Rempel et al. (1985) have major benefits. Barber's model provides a broader context and richness of meaning needed to characterize many interactions in automated systems. On the other hand Rempel and colleagues provide the dynamic factor needed to predict how trust may change as a result of experience with the system. Muir (1994) combined these two models to develop a more comprehensive model of trust in automation that contains six components: predictability, dependability, faith, competence, responsibility, and reliability (See Table 3). Muir and Moray (1996) were able to empirically prove that subjective trust ratings, along these constructs, from an operator could be used to measure user trust in a system.

Table 3. Muir's model of trust. Replicated from Uggirala and colleagues (2004).

Expectation	Basis of expectation at different levels of experience		
	Predictability (of acts)	**Dependability (of disposition)**	**Faith (in motives)**
Persistence			
Natural Physical	Events conform to natural laws	Nature is lawful	Natural laws are constant
Natural Biological	Human life has survived	Human survival is lawful	Human life will survive
Moral Social	Humans and computers act decently	Humans and computer are inherently good and decent.	Humans and computers will continue to be good and decent in the future
Technical Competence	One's behavior is predictable	One has a dependable nature	One will continue to be dependable in the future
Fiduciary Responsibility	One's behavior is consistently responsible	One has a responsible nature	One will continue to be responsible in the future

Self-Confidence

The benefit of using trust to guide one's attitude towards automation is that it serves as a heuristic to quickly and easily compare one's self-confidence in doing the task themselves (i.e., one's own perceived reliability) to the perceived reliability of the automation doing the task correctly. While perceived reliability in the automation is strongly determined by the actual reliability of the system, self-confidence in one's ability to manually perform a task is related to Bandura's (1986) concept of self-efficacy. Self-efficacy "*refers to beliefs in one's capacities to organize and execute the courses of action required to produce given attainments*" (Bandura, 1997, p. 3). However, self-efficacy is situation specific and while an individual may have high self-efficacy in general or in one area (e.g., academics) they may have lack self-efficacy concerning another area (e.g., athletics).

In this vein, if one has worked with a system that has consistently helped them to achieve their goals, and they have low perceived self-confidence in accomplishing the task themselves, then most likely their trust and reliance on that system should be high. On the other hand, if the system consistently fails in helping the individual achieve their goals, and they have high self-confidence in their own ability to perform the task, the individual's trust and reliance on the system should be low. Indeed research supports that when trust in an automated agent exceeds operators' self-confidence, automation is likely to be used; while, if self-confidence exceeds automation trust, then manual control is more likely to be maintained (Lee & Moray, 1994). As one's feelings of trust in a system vary, according to how they view the reliability of both themselves and the automation, their corresponding use (i.e., reliance) of that system should change as well.

Trust and Reliance

Automation reliance relates to the use of automatic rather than manual control (Wiegmann, Rich, & Zhang, 2001). Research has shown that perceived trust in an automated system is tightly coupled with reliance upon that system (Muir, 1989). These findings typically indicate that ratings of trust tend to be slightly more conservative than users' reliance (i.e., actual agreements with the aids; Muir & Moray, 1996; Wiegmann; 2001). It is also important to mention that empirical findings by Jian et al. (2000) indicate that ratings of trust and distrust are opposites lying along a single dimension of trust, so that low measures of trust actually reflect distrust of a system.

Before moving on it is important to emphasize that there is a distinction between measures of reliance and those of performance. *Reliance* is the tendency to employ automation to replace manual control. For instance, selecting the automated option 80% of the time exhibits greater reliance than selecting the automated option 50% of the time. On the other hand, *performance* is directly related to the number of correct and incorrect responses, which may or may not be related to reliance. In this vein trust may lead to more or less reliance (i.e., cooperation) with the aid, which may be desirable or undesirable (i.e., calibrated or miscalibrated) in regards to performance (Corritore, Kracher, & Wiedenbeck, 2001). Indeed, Gempler and Wickens (1998) found that individuals became complacent when observing highly reliable traffic-information displays. In their study, observers relied heavily on the automation even though several automation failures reduced overall performance dramatically. Surprisingly, no changes occurred to user reports of trust in the automation. Alternatively, Lee and Moray (1992) found that operators, performing a simulated processing control task, demonstrated drops in automation trust and reliance after an automation failure even though performance did not change significantly. In a hypothetical example, you can imagine two users may have the same level of system performance and yet their subjective interaction may be quite different. One operator may trust the automation and use it while concurrently performing other tasks; meanwhile, the other operator may distrust the automation, monitor it intensively or even do the task manually, experiencing greater stress, time pressure, and mental workload. Thus, achieving ideal performance requires that the operator properly calibrate their level of trust, and hence reliance, in the automation to maximize performance (i.e., minimize both misuse and disuse) and optimize their subjective interaction.

Trust as a Function of Experience

Research by Rotter (1967) established that an individual's general level of trust has a temporal factor, in that it is based on past experiences with others (e.g., parents, teachers, peers, etc.), that leads an individual to develop their generalized attitude of trust. That is, the way one reacts in a particular situation is not only determined by that situation but by previous experiences that individual has had. This relates to social learning theory in that "*expectancies in each situation are determined not only by specific expectancies in that situation but also, to some varying degree, by experiences in other situations that the individual perceives as similar*" (Rotter, 1980, p. 2). Thus, children who have experienced a higher proportion of promises kept, including threats of punishment, by parents and authority figures in the past have a higher generalized expectancy for interpersonal trust from other authority figures (Rotter, 1971). Research has carried these finding over to the human-automation literature as well. In this vein if the trustee, whether human or automation, performs according to the trustor's expectations, trust may be maintained or increased based on these experiences. On the other hand, not living up to expectations will lower trust (de Vries et al., 2003). Pritchett and Bisantz (2006) found that when an alerting systems acts contrary to an operator's expectations or produces alerts that are interpreted as false alarms, user trust and acceptance of the automated alerting system decreases. That is, as a user observes or believes that an automated aid has made an error they develop an expectancy that the aid is unreliable (Lee & See, 2004).

It is also commonly accepted that individuals generally differ in their trait generalized expectancy of trust in others (Rotter, 1967). Research has shown that individual differences in generalized trait expectancy for automation also exist. In a national survey by Halpin, Johnson,

and Thornberry (1973) evidence of a generalized technology trait trust expectancy was found. Their findings indicating that while most people believed computers and other forms of technology would improve their lives, others viewed these as dehumanizing and prone to errors. In a similar study by McClumpha and James (1994), aircraft pilots were shown to demonstrate previously established favorable or negative views of cockpit automation. These results were further supported by Lee and Moray (1994), who found that individual differences in the preference of using automation heavily influenced reliance upon automation in a laboratory based study. That is, some operators were consistently prone to using, or not using, automation regardless of their ratings of trust and self-confidence (Beck, Dzindolet, & Pierce, 2002). On a short-term scale, Lee and Moray (2004) conducting a time series analysis, found that future reliance upon automation was also influenced by past use of the automation. In their discussion of these results, the researchers took this information to mean that human beings are reluctant to change, and that includes the use (or alternatively the disuse) of automation.

In research by Riley (1994), a definite difference in allocation strategy was found between students and pilots using faulty automation. While nearly all the students turned off the faulty automation, almost half of the pilots used the automation when it failed. This difference in allocation strategy may be due to pilots employing automation more often in their work environment; hence they were more influenced by using automation in the past. Indeed, experience with automation has been shown to mediate generalized trust expectancies in technology. For instance, those with experience with automation and/or computers tend to have more favorable attitudes toward automation then those without such experience (Lee, 1991; Lerch & Prietula, 1989; McQuarrie & Iwamoto, 1990). However, the reverse has been found with extensive experience with the task being automated. Thus, individuals who are experts in

the task to be automated tend to have more negative opinions of the automation. However, this may be because they have greater self-confidence in performing the task and thus less need for the information the aid is providing (Sheridan, Vamos, & Aida, 1983).

Task Complexity

A second related factor influencing the human-agent team interaction is that of complexity of the task (Parasuraman & Riley, 1997). Task complexity can be defined as increasing the cognitive and/or physical characteristics of a task, which correspondingly increase demand on operator resources. It has been found that as task complexity increases it negatively impacts operator self-confidence (Lee & Moray, 1994). Complexity makes a complete understanding of the automation impractical, thus resulting in greater reliance upon the automation. By guiding reliance, trust helps to overcome the cognitive complexity people face in managing increasingly sophisticated automation. Therefore, it is theorized that task complexity has a moderating impact on trust in automation, that is, increased trust in automation serves as a heuristic replacement for vigilant information seeking and processing, thus simplifying the complexity of the task at hand (Moiser & Skitka, 1996; Parasuraman & Riley, 1997).

In the following studies complexity has been imposed upon the operator by having them monitor multiple agents. Automated decision recommendations help reduce the complexity of the task if the operator trust and relies upon them, thus reducing their own processing requirements. The concern then becomes how does trust/distrust in automation spread in a system with multiple decision aids? If operators come to distrust one component of a system, will their distrust spread to other components of the system? A study by Muir and Moray (1996)

found that distrust in one automated system did spread to reduce the trust of structurally, functionally, or causally related components. The impact of this is that distrust of a poor system often lead to unwarranted distrust of a concurrently running correctly functioning automated system. However, Muir and Moray also found that distrust of one component did not spread indiscriminately over the entire system. That is, trust levels of two subsystems that are structurally and functionally independent may not be contingent upon each other. It is important to mention that both aids in the Muir and Moray (1996) study were not cognitive aids but physical aids which guided several processes in a process control simulation. Additionally, Lee and Moray (1994) were able using discriminate validity measures, to prove that operators were able to partition their trust and self-confidence independently among several subsystems in a pasteurization process control simulation similar to that used by Muir and Moray (1996). An extension of this work would be to look at the effect of complexity on trust and reliance in several different types of automated aids (see Source Characteristics below), and how this relationship depends on agent reliability and error salience.

Agent Reliability.

Experiments assessing the association between machine reliability to performance have yielded a collection of myriad findings (Wiegmann, Rich, & Zhang, 2001). While the dominant viewpoint has been that as automation reliability increases so does reliance upon that automation (de Vries et al., 2003; Liu & Hwang, 2000; Moray, Inagaki, & Itoh, 2000; Muir, 1987; Muir, 1994; Riley, 1994), other lines of research found that overall reliability was not related to reliance upon the automation (Parasuraman, Molloy, & Singh, 1993; Singh, Molloy, &

Parasuraman, 1997). As with human teams increasing the reliability of one team member's performance may not necessarily affect the overall team's performance (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003).

In one of the first studies on trust upon automation usage, Muir (1989) found that automation accuracy had a strong correspondence with automation use. However, different results were found in a study by Dzindolet, Pierce, Beck, Dawe, & Anderson (2001). In this study individuals were told that the automation would be correct on 60%, 75%, or 90% of the trials; additionally, there was a control condition in which no automation was presented. Analysis indicated that there were no significant differences in the probabilities of errors associated with the four reliability conditions; that is, reliability of the aid did not affect accuracy. One explanation for this null effect could be that participants were unaffected by the detector; they ignored the aid and continued with manual performance. If this is true than participant accuracy should be independent of aid accuracy; that is, the probability of an operator error when the aid was correct (p(error | aid correct)) should equal the probability of an error when the aid was incorrect (p(error | aid incorrect)). A reliable difference between the probabilities of an operator error associated with the agents correct and incorrect recommendations would suggest that the detector's responses influenced the operators' decisions. This is exactly what they found, an incorrect recommendation by the machine caused significantly greater probability of an operator error, than a correct recommendation by the machine (0.27 vs. 0.13 respectively). Thus, operators' decisions were related to the detectors recommendations but not to the accuracy of the machine in general. Other research has shown that trust is greatly reduced by small automation errors (compared to perfect automation performance), and increasingly less sensitive to larger automation errors (Muir & Moray, 1996).

System designers should not assume that more reliable decision aids will always produce better performance by human-machine teams (Beck, Dzindolet, & Pierce, 2002). In fact the literature has shown that with perfect reliability, individuals tire of monitoring it (i.e., become complacent), and are less able to deal with errors when they occur then if they were responding autonomously or with a less reliable aid (Sheridan, 2002). Indeed, in a study by May, Molloy, & Parasuraman (1993) it was found that the detection rate of automation failures varied inversely with automation reliability. That is, the more reliable the automation the more complacent the operator.

Object of Trust

The object of trust may simply be defined as *what* the trustor is trusting. In this definition the object of trust may be another individual or even an entity (e.g., robot; Corritore, Kracher, & Wiedenbeck, 2001). Based on past research there are some cases in which trust differs between machines and fellow humans. One well documented case of these differences is polarization bias.

Polarization bias refers to the unrealistic extremely favorable (perfection bias) or unfavorable views (rejection bias) of automated decision aids. Due to this effect individuals tend to be unforgiving of automation that deviates from perfection (Wiegmann, Rich, & Zhang, 2001). On the other hand, human beings are imperfect; no one is immune to occasional mistakes. Thus, a human operator could be expected to make a mistake on one problem and then be correct on the next. However, automated devices are generally considered to work perfectly or not at all. For instance, if the numbers are entered correctly a calculator it will give one the correct or

incorrect answer to every problem (Beck, Dzindolet, & Pierce, 2002). Machines, like calculators, tend to be either functional or dysfunctional.

In a study by Dzindolet, Pierce, Beck, and Dawe (2001) individuals were asked to rate the expected performance of either a human or a machine partner in a detection study. The detection study asked participants to view slides that displayed only terrain or terrain plus a camouflaged soldier (in various levels of camouflage). The users were also presented with their partner's (who they were lead to believe was either human or machine) decision on whether the photo contained a human form. Participants consistently rated the machine as being more accurate, "perfection bias", prior to experience using the automation. However, after practice with the soldier detection task there were no significant differences in user expectations between human and machine partners. In another study by Wiegmann and colleagues (2001), "rejection bias" was found in which user's underestimated a system's true reliability because the automated diagnostic aids were not perfectly reliable. While this bias has been found it has not been investigated among multiple agents of varying reliability levels.

Failure Salience

Failure salience is defined as how visible an automation failure is to users. Failure salience may significantly impact trust in automation (Barnett, 2000). In a study by Beck and colleagues (2001), participants were briefly shown pictures that did or did not contain a camouflaged soldier. For each trial, participants first reported whether or not they had detected a soldier; after their response they received a recommendation from an automated contrast detector as to whether or not it had detected a soldier in that trial. When a user did detect a signal, but the

aid did not, they could be certain that the detector had indeed missed the signal. On the other hand, when an operator did not detect a signal, but the aid did, they were unsure whether they had missed the signal or the automation was in error. That is, if the human monitoring the display detected nothing and the decision aid reported the presence of a signal, one would not know if there really was nothing present (a false alarm on part of aid) or that the human observer actually did miss a signal (a miss on part of the participant). One study by Mosier and Skitka (1998) observed the effect of faulty automation cues on aircrews during a flight simulation. One such faulty automation cue was a false alarm indicated an engine fire. Rather than lose trust in the automation, 74% of the aircrews erroneously recalled diagnostic cues to support the engine fire alert. Thus, it appears that operators may interpret automation false alarms in varying ways, even to the point of misremembering information to support the automation false alarm. On the other hand operators tend to lose automation trust and be more confident in their own responses when confronted with highly salient misses by the automation. Operators may thus establish a false belief in their superiority because the detector's misses may be more noticeable than their own misses. Another potential explanation is that initially users have a "perfection bias" in automation, observing errors made by the automation are inconsistent with the expectation (i.e., schema) of perfect automation performance and are thus going to be more vivid in memory and play a larger role in information processing.

To examine this issue of failure salience Beck and colleagues (2001) paired students with machines that performed at an inferior or superior level to their own manual performance. They then instructed students that their extra credit for participation in research would be dependent on the number of correct trials among 10 random trials drawn from their or the machine's performance (200 total trials). No misuse occurred, students working with inferior machines

made no biased decisions. However, despite being given feedback that the machine had superior overall performance, it was found that 31 of 36 students made extra credit contingent on their own inferior performance (disuse). There are several potential explanations for this finding, students may have felt a desire to be mentally engaged in the activity (i.e., avoid boredom), a moral obligation to contribute to the task, a need to be in control of the process, or a false and distorted belief perseverance (Lee & Moray, 1994; Wiegmann, Rich, & Zhang, 2001). In regards to belief perseverance this is when false ideas (e.g., more salient machine error rate) can continue to influence attitudes after they have been discredited (i.e., when the students were informed of the machines superior overall performance). In this vein memory is highly selective and not all mistakes are going to have an equal influence on future judgments (Beck, Dzindolet, & Pierce, 2002); that is, an operator who may have vivid memories of the detector's errors may have less prominent recollections of their own errors. Operators are attuned to the worst observed machine behaviors, so to encourage trust automation must be desirable and consistent. Thus, even if automation only degrades system performance momentarily it may still highly degrade trust.

Another potential avenue related to failure salience is the difficulty of the trial in which the automation makes the error. That is, in easier trials in which a signal is highly apparent an error in the automation would be quite obvious. However, an error in a more difficult trial would be less salient. A hypothetical example would be to use a automated weapon shape contrast detector with a baggage screening task; in this case a large assault rifle would be a highly salient signal (an error in the detector would be quite blatant), compared to a less salient partially occluded handgun (an error would be less obvious). It would stand to reason that the more apparent and vivid an error the greater the decrement to user trust. This has been found in the case of Lee and Moray's work (1992) examining the effect of automation with different levels of

error (i.e., large or small error). They found that trust was found to decline with increasing magnitude of the faults. However, the nature of the aid and the task was quite different from the current research. It is of theoretical and practical important to see if these findings, with a physical aid in a simulated pasteurization task, hold true with a decision aid in a simulated search-and-rescue task.

Additional Moderating Factors.

Of course, these variables are not the only influencing factors upon trust and reliance of automated aids. Other factors include workload, situational awareness (SA), validity, transparency, utility, etc (Liu & Hwang, 2000; De Vries et al., 2003). Further there are other constraints that may interfere with reliance. For example, the operator may not have enough time to engage the automation even if they trust it and intend to use it, the effort to engage the automation may outweigh its benefits, or they may simply use automation they don't trust because they are unable to do the task themselves (Corritore, Kracher, & Wiedenbeck, 2001). For example, in one study by Desmond, Hancock, and Monette (1998) it was found that monitoring an automated system that drove an automobile was just as fatiguing as actually driving the automobile. Another important issue is that of social loafing which can occur when individuals operate in groups in which their individual performance is masked by the efforts of others (Burdick, Skitka, & Mosier, 1997; Dzindolet, Beck, Pierce, & Dawe, 1998). Indeed, research has shown that merely providing operators with the opportunity to rely on an automated aid actually decreases their motivation to perform the task, though conflicting research has found that social loafing does not occur in human-automated-interaction as there is still only one person

who bears responsibility for the system. In this light another human must be present for the responsibility to be shared between the individuals (Lewandowsky, Mundy, & Tan, 2000). Another factor influences trust is the consistency of error within the automation. If an automated system has constant error users learn to compensate for the constant error and trust in the automated system increases, on the other hand with variable error systems trust stays low even with practice with a system. Indeed, Muir and Moray (1996) found that a small variable error is just as damaging as a large constant error on trust. All of these factors may influence operator trust and reliance in an automated system, but are beyond the scope of the current dissertation, it would be recommended that future research examine these factors in combination with the variables examined in the current research.

Purpose of the Current Study

Although several researchers have examined the differences in trust in relating to humans or machines agents separately, the literature is severely lacking in examining how operator trust is impacted by interacting with multiple human or machine agents who vary in their actual reliability levels. Therefore, the present study was undertaken to determine if operator trust in an agent is affected by a concurrent agent, and to what degree this relationship is moderated by agent type, mixed reliability levels, and salience of the automation failures.

Mixed Reliability Levels

Several researcher mentioned previously have examined the effect of various automation reliability levels on user acceptance of the automation. Muir (1994) has suggested that

developing trust in an automated system requires being able to predict its operation. This stands to reason that increased experience with a transparent system will increase users' ability to predict an automated systems response. However, when users are monitoring multiple automated aids of mixed reliability (i.e., low and high reliability), increased experience with a low reliability aid may negatively impact user reliance on a concurrent high reliability aid. However, the degree of impact of this automation bias crossover is currently unknown. It is also, for theoretical and practical interest, important to examine whether this bias works in the opposite direction; that is, whether experience with a high reliability aid increases reliance in a concurrent low reliability aid.

For the present research, it was decided to evaluate a human operator monitoring two automated agents. To accomplish this goal four experiments were conducted employing a search-and-rescue task. The testbed was created and adjusted in experiments 1 and 2. Based on the results of experiment 3, the reliability of the low- and high-reliability aids were determined. In Experiment 4, effects of reliability conditions (i.e., both low, mixed, both high) and agent characteristics (i.e., human agents, same-type robotic agents, different-type robotic agents) were tested. Bias between the mixed reliability levels were examined by comparing them to the uniform reliability levels.

Agent Type

Differences, such as polarization bias, have been found in the way humans trust other humans versus how they trust machines. In general, it has been found that operators are less forgiving of machine failures. In this study I examined how agent type influenced operator trust

and reliance on multiple agents. It's predicted that robotic-agents will suffer greater drops in trust following errors compared to human agents, due to polarization bias. Additionally, I believe that agent type will influence the biasing effect that I expect to occur between mixed reliability levels. That is, two human agents will be perceived as independent so the mixed reliability bias should not occur. However, two similar machine agents will be perceived as very similar so the mixed reliability bias should occur. An intermediate level was chosen for comparison in which two unique machine agents performed the task. In this last case it is believed that some bias would occur but to a lesser degree than it would with two similar machine agents. This effect was examined, in Experiment 4, by having participants monitoring what they believe is two human agents, two same-type robotic agents, or two different-type robotic agents.

Salience of Automation of Failures

It is known that number of errors, in the form of overall reliability, often affects user's trust and reliance in an automated system (Lee & Moray, 1992). Given this it would be interesting to examine if type of automation error impacted user trust and reliance in different ways. That is, will automation errors on easier difficulty trials cause greater drops in operator trust and reliance? It stands to reason that how visible a failure is to users may have a significant effect on their confidence in the automation they use. Further, it is believed that this effect will be moderated by the object of trust. The literature demonstrates that automation bias indicates that any error on the part of the machine is detrimental to operator trust and reliance. However, it is believed that if the error comes from a human agent that users will be more forgiving of an error, especially if the error occurred on a particularly difficult trial.

Testing the Theories

The following studies tests the above theories by having participants perform a task in which they are aided by two automated systems. The experimental groups were divided into the cells of a 3 (reliability condition) x 3 (source characteristics) x 3 (failure salience) mixed design with within subjects on the latter factor. Participant self-confidence, trust, reliance, and performance in the automated systems were measured.

During the task the automated systems would occasionally fail, the number of times would be dependent upon the reliability condition. The participants' trust and reliance in the two automated agents would be compared between the mixed vs. uniform reliability groups for each reliability level (i.e., low and high). Any differences between group scores would support the theory that mixed reliability levels experience a biasing effect.

The interface of the design will be manipulated so that the automated agents are either represented as distributed human agents, machine agents of similar types, or machine agents of dissimilar types. The participant responses across these groups and the interaction between reliability levels and agent type will be examined. Difference in a main effect would support the theory that operators respond differently to other humans compared to machines. An interaction would indicate that not only do operators respond differently to humans compared to machines but that this is impacted by the reliability levels of the agents.

Finally the failure salience will be manipulated so that trials in which automation errors occur will vary in their difficulty. This examines whether aids are viewed as more reliable if their errors occur on more difficult stimuli. One way to examine this would be to look at participant reliance following automation errors on easy trials compared to reliance following automation

errors on more difficult trials. Here an effect would indicate that the salience of the error impacts user reliance. It would be most advisable to include multiple levels of automation error difficulty levels so as to examine interaction effects. That is, to examine whether the salience of an automation failure would differ depending on what reliability level a user was experiencing (i.e., a difficult error may go unnoticed in a high reliability condition due to operator complacency) or particular agent characteristics (i.e., humans may differ in how they treat difficult errors by other humans and automation, but not to how they treat easy errors by either group). In regards to agent type, Dzindolet et al. (2003) mention determining agent competence based on item difficulty is a somewhat flawed strategy, due to the fact that humans and automated aids often process information differently. What might be considered an easy unambiguous stimulus for a human decision-maker may be considered an ambiguous and difficult stimulus for an automated decision aid. The greater this difference, the less trustworthy an automated aid may be perceived to be.

Research Hypotheses

Therefore, the present studies were designed to test these theories. Regarding the role of reliability, source characteristics, and failure salience with an automated system, a number of hypotheses emerge.

1. In a complex, dual-aid, condition there will be bias between two agents of mixed reliability compared to two uniform agents.
    a. Trust and reliance of a high-reliability agent will be negatively influenced by a concurrent low-reliability agent.

b. Trust and reliance of a low-reliability agent will be positively influenced by a concurrent high-reliability agent.

2. Operators experiencing high automation reliability will have significantly more subjective trust in the automation than those experiencing both low or the mixed reliability conditions. Additionally those with low automation reliability will experience significantly less subjective trust of the automation than those in the mixed reliability condition (H_O = There is no significant difference between reliability group trust scores).
    a. Increased levels of automation trust will be accompanied by increased levels of reliance on the aid and lower levels of reported workload.
    b. Decreased levels of automation trust will be accompanied by decreased levels of reliance on the aid and higher levels of reported workload.
3. Subjective levels of trust, automation reliance, and workload are expected to differ across agent type (i.e., human, similar computer agents, dissimilar computer agents). Such that human agents have increased trust, increased reliance, and decreased workload, compared to the computer agents. The computer agents are not expected to differ in overall trust, reliance, or workload (H_O = There is no significant difference between agent type group trust ratings, reliance, and/or workload).
4. In a mixed reliability condition the agent type is expected to significantly impact crossover bias between the two agents. (H_O = There is no significant interaction between reliability and agent type).

a. Two agents perceived to be human will experience the least crossover bias in the mixed reliability condition. Thus, a low-reliability human aid will have little impact on a concurrent high-reliability human aid.

b. The same-type robotic agents will experience the most crossover bias in the mixed reliability condition. Thus, a low-reliability same-type robotic agent will have a strong impact on a concurrent high-reliability same-type robotic agent.

c. The different-type robotic agents will experience an intermediate level of crossover bias in the mixed reliability condition. Thus, a low-reliability different-type robotic agent will have an intermediate impact on a concurrent high-reliability different-type robotic agent.

5. The failure salience of the automation error is expected to influence the likelihood of relying on the aid in the future trials. Such that as the salience increases the lower temporal reliance becomes (temporal reliance is measured by the agreement with an aid on the trial following an aid error). (H_O = There is no significant effect between failure salience groups for temporal reliance).

    a. High salience failures (i.e., obvious errors) will cause a significantly less temporal reliance on the aid compared to less salient errors (moderate and low salience failures).

    b. Moderate salience failures will cause less temporal reliance compared to low salience failures but maintain higher temporal reliance than high salience failures.

c. Low salience failures will maintain the highest level of temporal reliance compared to the more salient errors.

6. It is expected that source characteristics of the agents and the salience of the agent errors will interact to affect temporal reliance. (H_O = There is no significant interaction between source characteristics and failure salience).
    a. Agents perceived to be human will experience drops in temporal reliance proportional to the increasing simplicity of the error made. Also it is expected that participants will be more forgiving of human errors compared to robotic errors, especially on more difficult stimuli.
    b. The computer agents will experience equivalent drops in reliance across all types of errors. This reflects automation bias, in which automation is expected to work perfectly or not at all. Participants will be unforgiving of all robotic errors regardless of error salience.

Independent Variables

Agent reliability will be manipulated so that there will be a low-reliability and a high-reliability condition. As reliability levels are highly dependent upon the task in question, Experiment 3 will serve to determine appropriate values for this study. As prior research has shown that prior experience with one reliability level impacts trust/reliance of subsequent reliability levels agent reliability was kept as a between-subjects measure (i.e., both high, both low, or mixed: one high and one low).

Source characteristics will be manipulated so that individuals are told that they are monitoring decisions from either two human agents, two similar computer agents, or two dissimilar computer agents. Images of the agents were placed on the display to increase the salience of this independent variable. To increase the believability of agent type, source characteristics was kept as a between-subjects measure.

Error salience will be manipulated so that the automation makes errors that are low, moderate, and high salience. The salience of the errors were determined by experiment 2 which will evaluate the difficulty of the clips, with more difficult clips (i.e., fewer participants correctly identifying) as being less salient (i.e., less obvious). Error salience will be a within subjects measure.

Dependent Variables.

Subjective Measures.

As Muir (1989) found people are able to generate meaningful subjective ratings of trust. Participant trust ratings are sensitive to the properties of the automation and related in a sensible way to those properties (Lewandowsky, Mundy, & Tan, 2000). Thus, the following studies employed previously used subjective measures of trust before and after experience with the agents. The pre-questionnaire, used in the fourth study, asked participants to estimate their and their automated aids expected performance on the coming trials (See Table 1). The post-questionnaire queried participants on their subjective experience of the trials they have just completed and asked them to make a decision regarding how their performance score would be calculated (See Table 2 for post-questionnaire questions; format varied depending on whether

questionnaire was used in experiment 3 or 4). The use of the choice for score calculation was selected because trust involves some degree of vulnerability on the part of the trustor. Additionally personality measures were obtained to examine general trust expectancies, anthropomorphic tendencies, and automation complacency potential.

Table 4. Pre-questionnaire questions for experiment 4. Questionnaire adapted from Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003.

Question	Scale
How well do you think the agent will perform during the 120 trials?	Likert Scale 1-9, endpoints "Not Very Well" – "Very Well"
How well do you think you will perform during the 120 trials?	Likert Scale 1-9, endpoints "Not Very Well" – "Very Well"
Who do you think will make more errors during the 120 trials? I will make…	Likert Scale 1-9, endpoints "Many More Errors" – "Far Fewer Errors"
How many errors do you think you will make during the 120 trials? I will make about _____ errors.	Numerical value entered by participant, range from 0 to 200.
How many errors do you think the agent will make during the 120 trials? The agent will make about _____ errors.	Numerical value entered by participant, range from 0 to 200.
To what extent do you believe you can trust the decisions the agent will make?	Likert Scale 1-9, endpoints "Very Little" – "A Great Amount"
To what extent do you believe you can trust the decisions you will make?	Likert Scale 1-9, endpoints "Very Little" – "A Great Amount"
How would you rate the expected performance of the agent relative to your expected performance? The agent will perform…	Likert Scale 1-9, endpoints "Better Than I Will Perform" – "Much Worse Than I Will Perform"

Table 5. Post-questionnaire questions. (Questionnaire adapted from: Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Master, Gramopadhye, Bingham, & Jiang 2000).

Question	Scale
Competence: To what extent does the agent perform this search-and-rescue task effectively?	Likert Scale 1-9, endpoints "Very Little" – "A Great Amount"
Predictability: To what extent can you anticipate the agent's behavior with some degree of confidence?	Likert Scale 1-9, endpoints "Very Little" – "A Great Amount"
Reliability: To what extent is the agent free of errors?	Likert Scale 1-9, endpoints "Very Little" – "A Great Amount"
Faith: To what extent do you have a strong belief and trust in the agent to do the search-and-rescue task in the future without being monitored?	Likert Scale 1-9, endpoints "Very Little" – "A Great Amount"
Overall Trust: How much did you trust the decisions of the agent overall?	Likert Scale 1-9, endpoints "Very Little" – "A Great Amount"
What percentage of responses by the agent do you think were correct?	Range 0% to 100%
How often did you notice an error made by the agent?	Likert Scale 1-9, endpoints "Not At All" – "Many Times"
To what extent did you lose trust in the agent when you noticed it made an error?	Likert Scale 1-9, endpoints "Very Little" – "A Great Amount"
Imagine that there are ten more video clips that need to be examined for terrorists, civilians, and IEDs. Also imagine that we were to offer you an additional compensation, of either $5.00 or an extra credit point for *each* of these ten additional video clips that is correctly identified. However, due to a software problem only you *or* Teammate B can make the decisions. Would you prefer that this additional compensation be based on the decisions made by the automated aid or the decisions made by you? (circle one)	"Agent's Decisions" or "My Own Decisions"
We would like to know what led to your decision to base your performance on either your decisions or on the decisions of the aid. Please tell us everything you thought of in coming to this decision. Do not worry about spelling or grammatical errors. Use the back side of this paper if necessary.	Free Response. Previous study divided answers into 4 major categories. 1) Trust in computers ("I don't trust computers that much. I know a lot about their tendency for errors"), 2) detection of obvious errors ("There were a few times that I'm pretty sure I saw a terrorist, but the program said he was absent"), 3) confidence in self ("I was not that confident in what I saw" or "I chose to use 'my decisions' because I trust my observations, and I never second guess my self"), & 4) relative performance ("I had less errors than the computer", "The contrast detector made less errors", or "The computer made more mistakes compared to mine").

Behavioral Measures

To measure objective trust of an automated system, reliance was analyzed. Reliance was measured as the combined total of the times the participant agreed with the aid. Additionally, temporal reliance was examined by looking at the likelihood of automation reliance on a trial immediately following an automation failure trial. In all cases, the automation correctly worked on the trial immediately following an automation failure.

EXPERIMENT 1: METHODOLOGY

Experimental Purpose

Due to the requirement for certain features to address the overarching research questions in this line of research, it was required that an experimental platform be developed that could test these questions. An experimental platform was designed that could serve as an interface for users to monitor the progress of an Unmanned Ground Vehicle (UGV) through an office building. Experiment 1 served as a pilot study to determine that users monitoring the UGV are receiving adequate time to view the video clips (i.e., duration of the stimuli) and adequate time to respond to the video clips (i.e., inter-stimulus interval; ISI). The goal of this experiment is to ensure that the basic task itself was possible for participants to perform. That is, the task is set to a pace that is neither too fast nor too slow for participants. The selected video durations and ISI were maintained for the other studies in this dissertation.

Experimental Participants

Twenty-five participants were recruited through the University of Central Florida extra-credit website and they received course credit for their participation. Participation were limited to those with normal or corrected to normal vision. Total participation time did not to exceed 1 hour.

Experimental Procedure

Participants first completed an informed consent (See Appendix B), followed by a brief introduction to the task (See Appendix C) and a practice session. Participants then completed the experimental session, which is composed of 108 trials divided into 9 blocks. After each block the participant completed a brief questionnaire (See Appendix D). At the end of the experimental blocks the participant were thanked for their participation.

Training Procedure

The purpose of the training was to acquaint the participant with the nature of the task, the response buttons, and the stimuli. Participants received the training in the form of an experimenter read script (See Appendix C) and a computerized practice session. The script described in detail the scenario, what stimuli the participant would view, and how they were to respond. The computerized practice session had the same layout as the experimental display with the addition of three stationary images of the critical signals above the video feed (See Figure 2). The video feed presented 4 video clips during the practice session. These clips were presented for 5 seconds, with 5 second inter-stimulus intervals (ISIs). Four video clips were chosen to demonstrate each of the four potential stimuli: a terrorist (See Figure 3), an Improvised Explosive Device (IED; See Figure 4), a civilian (See Figure 5), and an empty room (See Figure 6). Participants were able to respond to the practice trials to become familiar with the interface.

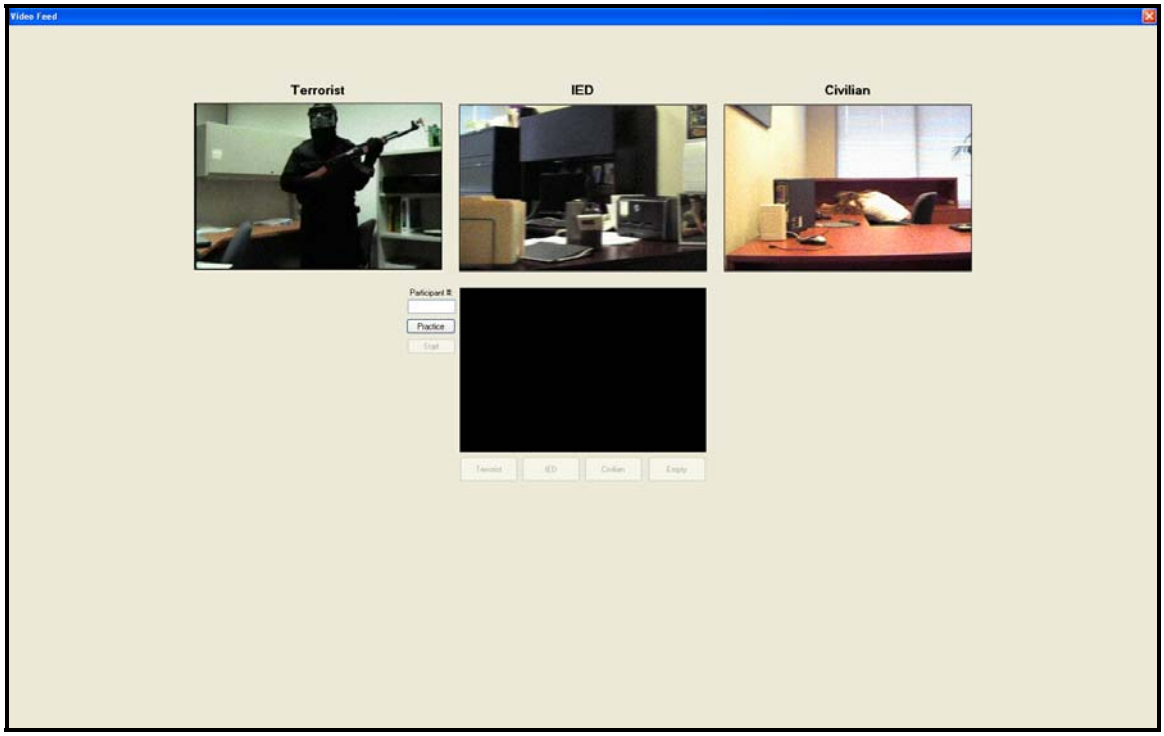


Figure 2. Practice interface for experiment 1.

Figure 3. Video clip demonstrating a terrorist.

Figure 4. Video clip demonstrating an unconscious civilian.

Figure 5. Video clip demonstrating an IED.

Figure 6. Video clip demonstrating an empty room.

Experimental Task

The experimental interface of the UGV search-and-rescue scenario was similar to the practice session, the main difference being the removal of the stationary stimuli images from the top of the screen (See Figure 7). Participants were able to respond after each video clip ends, by using the mouse to click on one of the response buttons (located beneath the video feed). Participants were able to respond only once per trial, this limitation was imposed by having the response buttons become deactivated after a participant had made a selection. Additionally, since the ISI was held constant during each block, participants were informed that it may take several seconds to move onto the next video after they have made their selection and that this was perfectly normal.

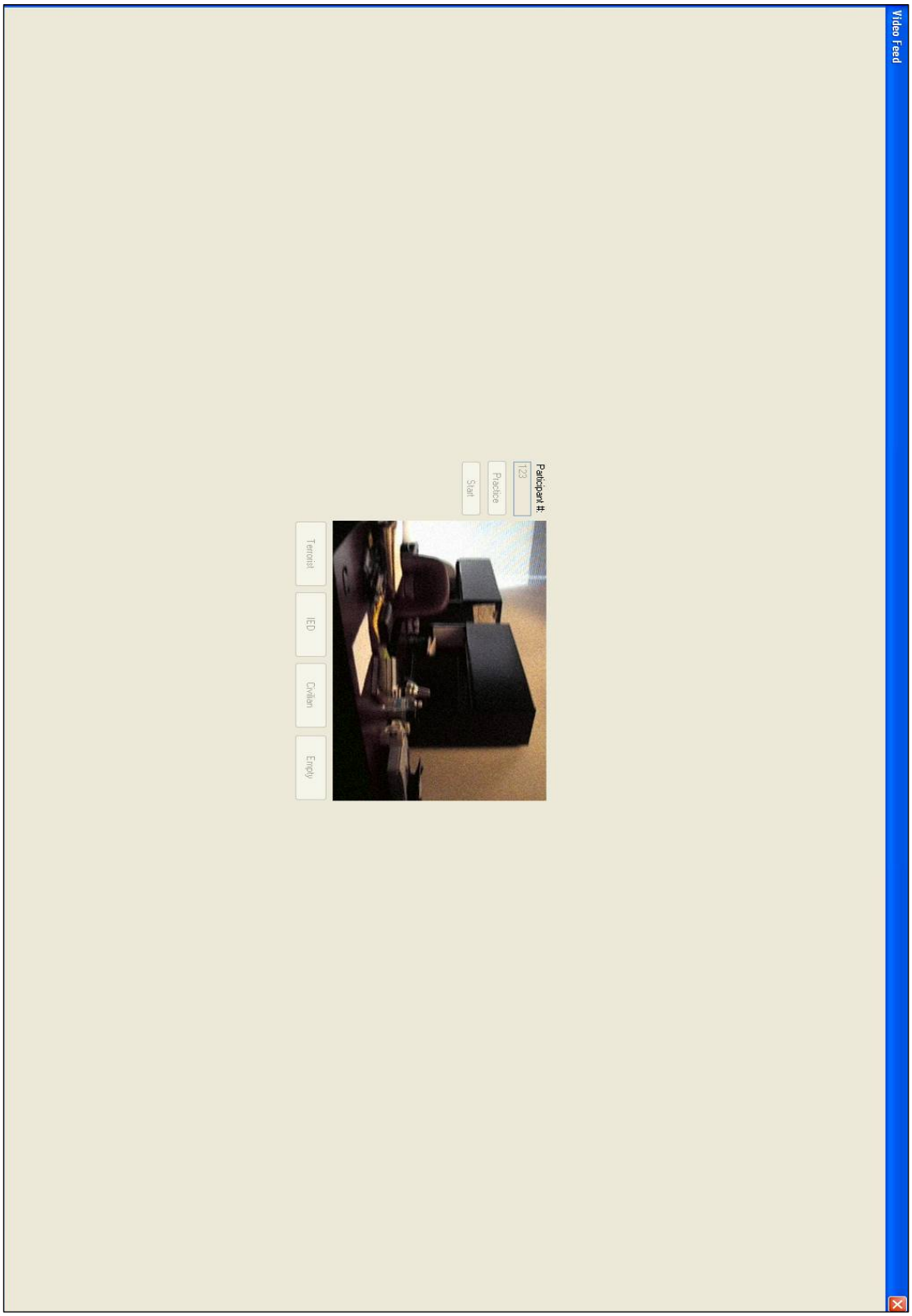


Figure 7. Video presentation interface for experiment 1.

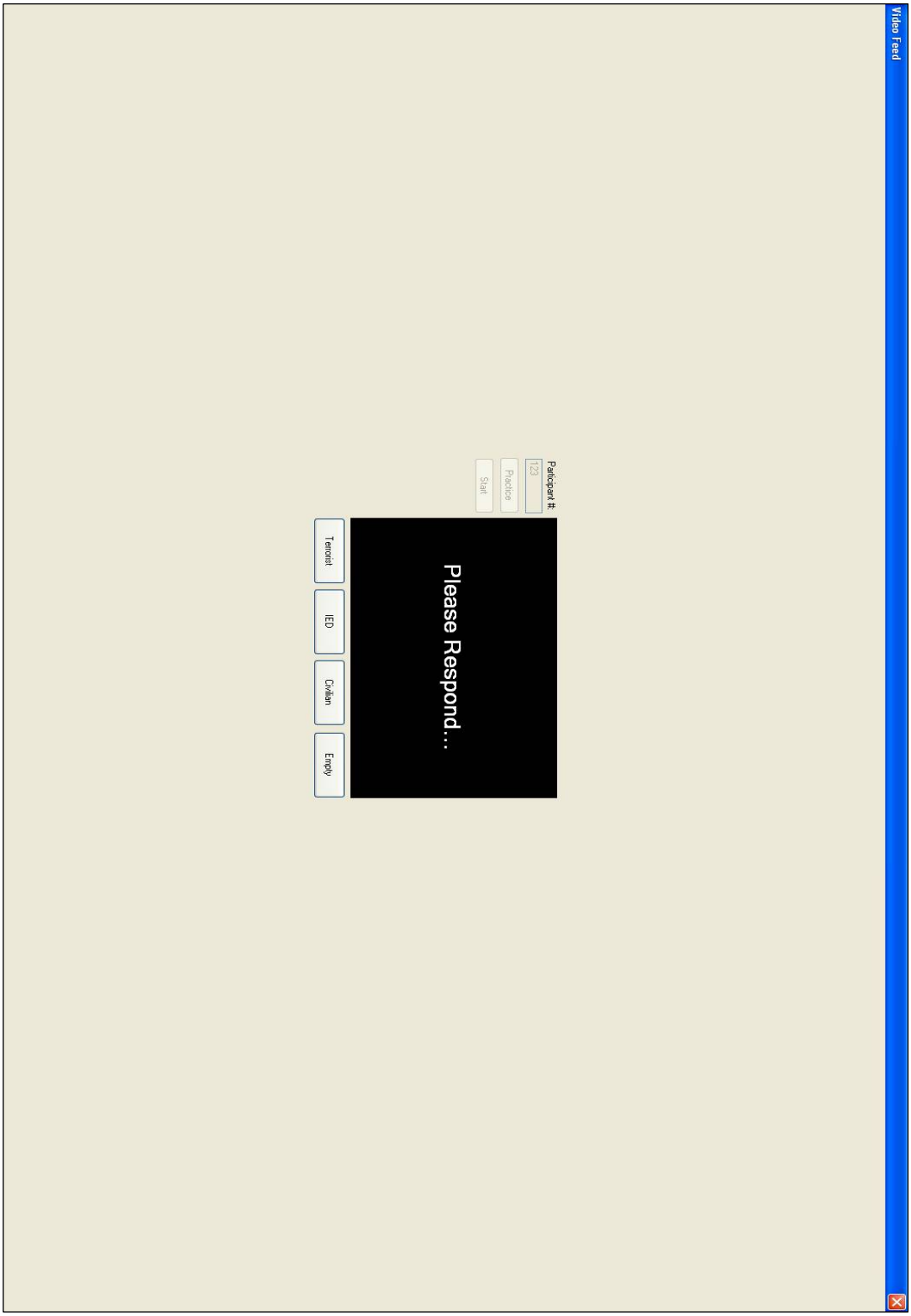


Figure 8. Response interface for experiment 1.

Experimental Conditions

The properties of video duration and ISI were manipulated in this study. Video duration was either 5, 7, or 10 seconds (video panning rate was held approximately constant). ISI was also either 5, 7, or 10 seconds in length. This lead to nine possible experimental conditions (See Table 6). Each experimental condition was composed of 12 randomly selected video clips, with the restriction that the 12 clips contained 3 of each kind of stimuli (i.e., terrorist, civilian, IED, and empty).

Table 6. Duration and ISI conditions.

Block	**Trial Duration**	**ISI**	**Total Time in Minutes (12 trials per block)**
1	5	5	2
2	7	5	2.4
3	10	5	3
4	5	7	2.4
5	7	7	2.8
6	10	7	3.4
7	5	10	3
8	7	10	3.4
9	10	10	4

Measurement and Analysis

Subjective Measures

Subjective measures were obtained after each block by using a post-block questionnaire (See Appendix D). All questions were presented in a Likert-style format, with the scale range of 0 to 10. Question 1 and 2 measured subjective satisfaction with the duration and ISI of each block, the scale endpoints were set so that an ideal satisfaction was rated in the midpoint of the

scale (5; with higher and lower values reflecting either too much or too little time respectively). Questions 3 thru 6 queried participants on whether they believed they would be able to monitor and respond to either 2 or 4 UGVs given the same durations and ISIs; scale endpoints were set to 'Definitely Yes' and 'Definitely Not.' This measure was used to reflect the participant's confidence in taking on a more complex task and provides some exploratory data as to whether individuals will be able to monitor multiple aids in following experiments.

The final six questions on the questionnaire, Question 7 thru Question 12, are the six rating scales from the NASA-Task Load Index (NASA-TLX; Hart & Staveland, 1988). The NASA-TLX uses six dimensions to assess workload: mental demand, physical demand, temporal demand, effort, performance, and frustration. Each dimension was rated by the participant on a scale from 0 to 10 with higher numbers reflecting greater workload. These values were then averaged into an overall rating of workload. Though the individual scales of temporal demand and perceived performance were of particular interest and were also examined individually.

Objective Measures

Objective measures were obtained for performance accuracy and reaction time (RT). In regards to performance accuracy I examined performance in terms of percentage correct across the different durations and ISIs manipulations. However, because the experiment was within-subjects different video clips were randomly selected for each block, thus performance comparisons may reflect differences in the inherent difficulty of the selected clips rather than differences due to the duration/ISI manipulation.

Reaction-time data was examined to determine the average reaction time needed to respond to a trial and its 95% confidence interval. This provided a general measure of the time most individuals would require for responding to the stimuli.

Experimental Equipment

The videos were recorded in three commercial and educational office buildings in the Central Florida area. Recordings were made using a standard digital video recorder set on a tripod dolly. To maintain maximum consistency a single operator, experienced with musical timing, controlled the pan rate of the camera. The terrorist in the video clips was held constant; such that, in each clip he was portrayed by the same individual, carried the same simulated assault rifle, an airsoft™ AK-47, and was outfitted in the same black outfit/mask to prevent any gender/racial/ethnic stereotyping (See Figure 3). Civilians were composed of a variety of volunteer participants of various genders and ages who were recruited at random from the three office locations. In all civilian clips the volunteers averted their faces; this was done to minimize the chance that participants might recognize any of the particular individual civilians (See Figure 4). The IED was held constant in all clips, and was composed of two metal canisters connected via wires to a timer (See Figure 5).

After obtaining the video stimuli it was then edited for length and noise using Adobe Premiere 2.0. Static was overlaid onto the video and the frame per second (fps) rate was reduced, from 30 fps to 15 fps, to develop brief choppy and realistic first-person video clips simulating an UGV exploring a commercial office building after a terrorist attack. The interface used to present the videos was created using Visual Basic.net. The interface contained one video display and

four response keys. Responses were recorded into a data file that records accuracy and response time (See Table 7). The simulation itself was presented on a desktop computer with a 20-inch widescreen monitor and an optical mouse for responding. Participants were instructed to wear headphones during the task to block out any extraneous noise.

Table 7. Recorded output from UGV simulation. All variables are recorded for each trial with the exception of participant # and date/time.

Name	Meaning	Example
Participant #	Identifies each participants data file	1
Date/Time	Records date and time of participant	7/27/2007 4:51:56pm
Clip	The video clip file name.	G5C1.avi
Group	The experimental condition (Duration and ISI)	5
Signal	The type of stimulus that is presented	Civilian
Response	The participants response	Empty
Correct	Whether the response is correct “C” or an error “E”	E
Reaction Time	The response time in seconds.	01.0156875

Hypothesized Outcome

The main determinant of ensuring adequate video duration and ISI are the responses to questions 1 and 2 from the subjective questionnaire. It was hypothesized that, while all the tested durations and ISIs would be sufficient for performance and reaction time measures, a subjective preference would emerge benefiting moderate durations and ISIs (e.g., 7 second duration and ISI; Hypothesis 1). It was further postulated that this would be reflected in both overall and subscale workload scores (Hypothesis 2). However, self-confidence in handling additional video feeds is hypothesized to be greater for longer durations and ISIs (Hypothesis 3; see Table 8).

These hypotheses were based on experimenter experience with the task during the development phase.

Table 8. Hypotheses for Experiment 1.

	Hypothesis Number		
Dependent Measure	1	2	3
Response Time	Ample Time = [D_5, D_7 , D_{10}] Ample Time = [$I_5 < I_7 < I_{10}$]		
Performance (% Correct)	Ample Performance = [D_5, D_7 , D_{10}] Ample Performance = [$I_5 < I_7 < I_{10}$]		
Satisfaction with Video Duration	$D_7 >$ [D_5, D_{10}]		
Satisfaction with Video ISI	$I_7 >$ [I_5, I_{10}]		
Overall and Subscale Workload Scores		$D_7 >$ [D_5, D_{10}] $I_7 >$ [I_5, I_{10}]	
Self-confidence in monitoring two feeds			$D_5 < D_7 < D_{10}$ $I_5 < I_7 < I_{10}$
Self-confidence in responding to two feeds			$D_5 < D_7 < D_{10}$ $I_5 < I_7 < I_{10}$
*D_5 = Duration 5 Second, D_7 = Duration 7 Second, D_{10} = Duration 10 Second, I_5 = ISI 5 Second, I_7 = ISI 7 Second, I_{10} = ISI 10 Second			

EXPERIMENT 1: RESULTS

The purpose of the first experiment was to ensure adequate video duration and response ISI for manual performance of the UGV monitoring task. Below the results are discussed for the performance and subjective data (see Table 9).

Table 9. Findings for hypotheses for Experiment 1.

	Hypothesis Number		
Dependent Measure	1	2	3
Response Time	**Ample Time = [D_5, D_7 , D_{10}]** **Ample Time = [$I_5 < I_7 < I_{10}$]**		
Performance (% Correct)	Ample Performance = [D_5, D_7 , D_{10}] **Ample Performance = [$I_5 < I_7 < I_{10}$]**		
Satisfaction with Video Duration	$D_7 > [D_5, D_{10}]$		
Satisfaction with Video ISI	$I_7 > [I_5, I_{10}]$		
Overall and Subscale Workload Scores		$D_7 > [D_5, D_{10}]$ $I_7 > [I_5, I_{10}]$	
Self-confidence in monitoring two feeds			**D_5** $< D_7 < D_{10}$ $I_5 < I_7 < I_{10}$
Self-confidence in responding to two feeds			$D_5 < D_7 < D_{10}$ $I_5 < I_7 < I_{10}$
*D_5 = Duration 5 Second, D_7 = Duration 7 Second, D_{10} = Duration 10 Second, I_5 = ISI 5 Second, I_7 = ISI 7 Second, I_{10} = ISI 10 Second			

Performance Data

Response Time

The 95% CI for overall RT was examined. The mean overall reaction time (RT) was 921.7 ms, with a lower bound of 852.2 ms and an upper bound of 991.3 ms. Additionally, a 3 (duration) by 3 (ISI) repeated measures ANOVA was conducted on RT. The main effect in all cases for duration, ISI, and the interaction between ISI and duration was not significant ($p > .05$).

Percent Correct

A 3 (duration) by 3 (ISI) repeated measures ANOVA was conducted on the percentage of correct detections. Video duration was statistically significant, $F(2, 42) = 43.48$, $p < .0005$, $\eta^2 = 0.67$. Pairwise comparison indicated that the 10 second duration had significantly fewer correct answers then the 5 or 7 second duration, which did not significantly differ from each other (See Figure 9). The interaction between duration and ISI was also significant, $F(4, 84) = 10.35$, $p < .0005$, $\eta^2 = 0.33$ (See Figure 10). The main effect for ISI was not statistically significant ($p > .05$).

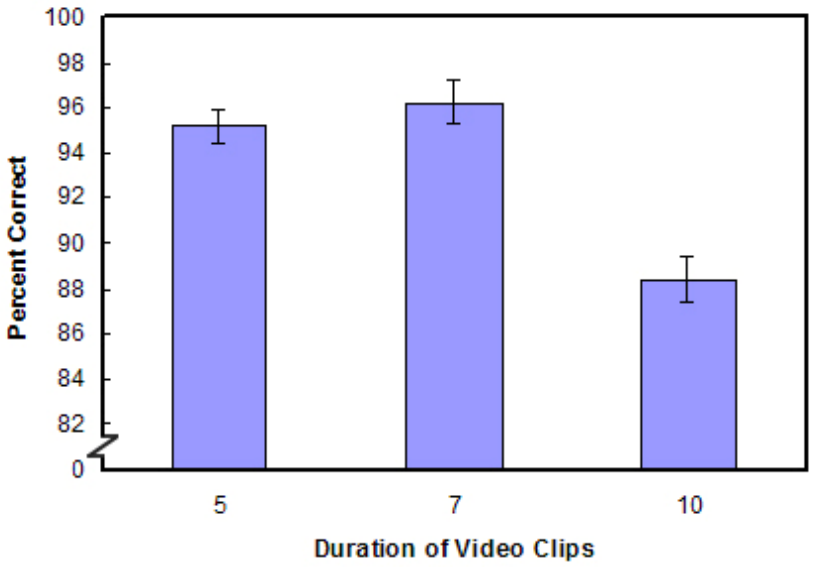


Figure 9. Percent correct as a function of duration of the video clips.

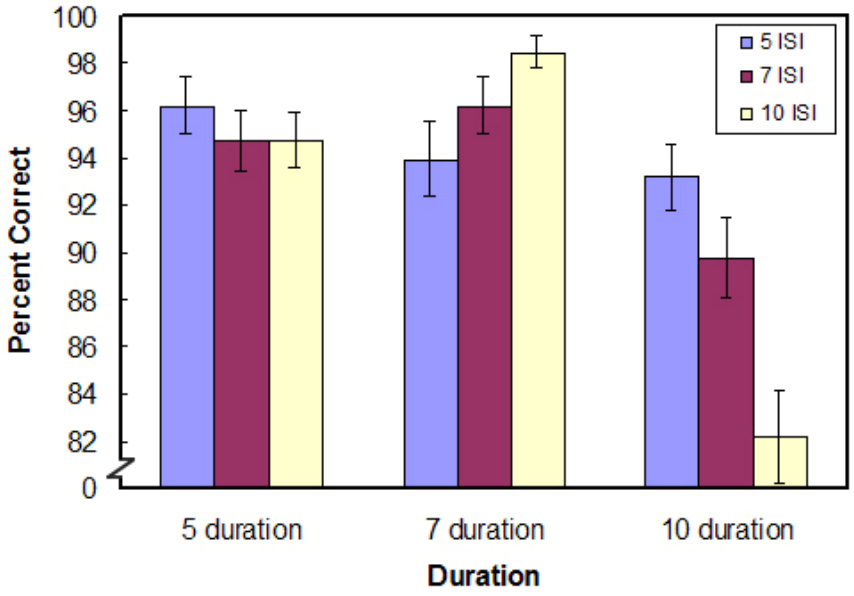


Figure 10. Percent correct as a function of video duration and ISI.

Subjective Data

Duration and ISI Subjective Satisfaction

The main determinant of ensuring adequate video duration was question 1 which concerned participant's subjective feeling of satisfaction with the amount of time they had to view each video clip. A 3 (duration) by 3 (ISI) repeated measures ANOVA was conducted on the data. The main effect for duration was significant, $F(2, 42) = 6.69$, $p = .003$, $\eta^2 = 0.24$. The main effect for ISI ($F(2, 42) = 0.08$, $p = .92$, $\eta^2 = 0.004$) and the interaction effect between duration and ISI were both not significant ($F(4, 84) = 2.09$, $p = .09$, $\eta^2 = 0.09$). Pairwise comparisons were conducted on duration, which indicated a significant difference between the 5 second and 10 second conditions, all other groups were not significantly different ($p > .05$; See Figure 11).

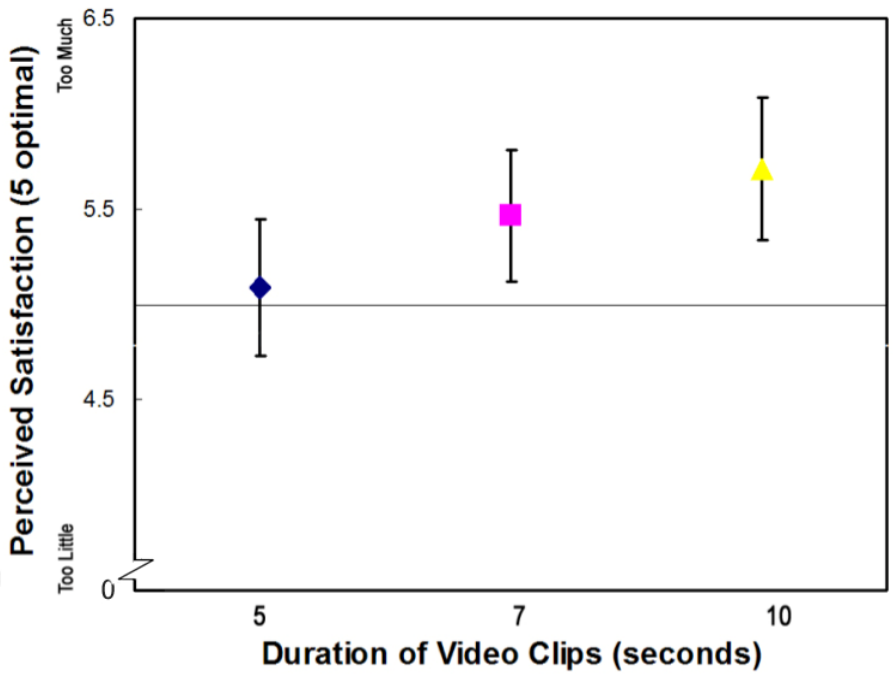


Figure 11. Perceived satisfaction of time to view each video clip as a function of video clip duration. Note that the line across the center represents optimal satisfaction with duration (a rating of 5), values above this line represent too much time, below this line too little time. Bars represent standard error.

The main determinant of ensuring adequate response ISI was question 2 which concerned participant's subjective feeling of satisfaction with the amount of time they had to respond to each video clip. A 3 (duration) by 3 (ISI) repeated measures ANOVA was conducted on the data. The main effect for ISI was significant, $F(2, 42) = 4.65$, $p = .015$, $\eta^2 = 0.181$. The main effect for duration ($F(2, 42) = 0.60$, $p = .55$, $\eta^2 = 0.028$) and the interaction effect between duration and ISI were both not significant ($F(4, 84) = 0.91$, $p = .46$, $\eta^2 = 0.042$). Pairwise comparisons were conducted on ISI, the 10 second ISI was significantly different from both the 5 and 7 second ISIs (the latter two did not significantly differ from each other, $p > .05$; See Figure 12).

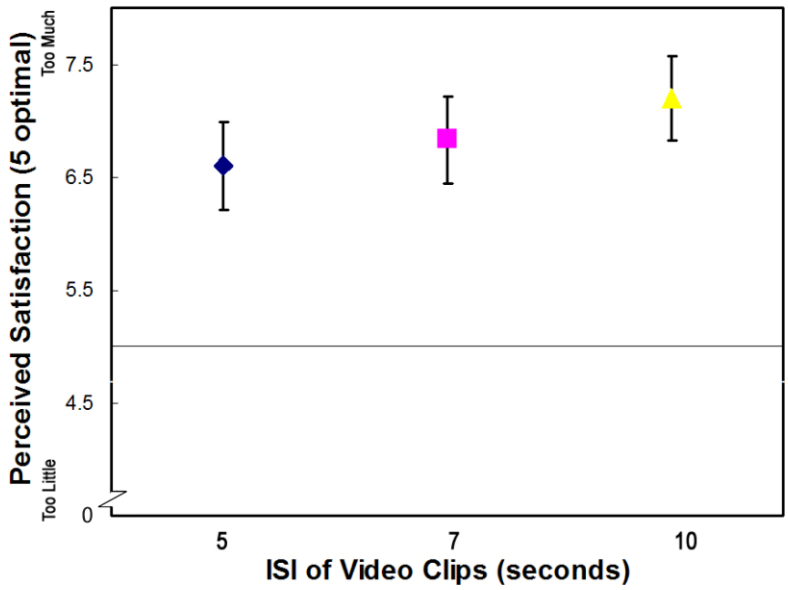


Figure 12. Perceived satisfaction of time to respond to each video clip as a function of ISI of each video clip. Note that the line represents optimal satisfaction with ISI (a rating of 5), values above this line represent too much time to respond, below this line too little time to respond. Bars represent standard error.

Overall and Subscale Workload

A 3 (duration) by 3 (ISI) repeated measures ANOVA was conducted on overall workload and each of the individual subscales. The main effect in all cases for duration, ISI, and the interaction between ISI and duration was not significant ($p > .05$).

Confidence in Handling Additional Video Feeds

To examine user's confidence in handling two agents Q3 and Q4 of the subjective questionnaire examined user confidence in monitoring and responding to two aids. A 3 (duration) by 3 (ISI) repeated measures ANOVA was conducted on Q3 regarding perceived confidence in monitoring two video feeds. Video duration was statistically significant, $F(2, 42) =$

5.10, p = .01, η^2 = 0.20. Pairwise comparison indicated that the 5 second duration was rated significantly lower in perceived confidence then the 7 or 10 second duration, which did not significantly differ from each other (See Figure 13). The main effect for ISI and the interaction between ISI and duration were both not significant (p > .05).

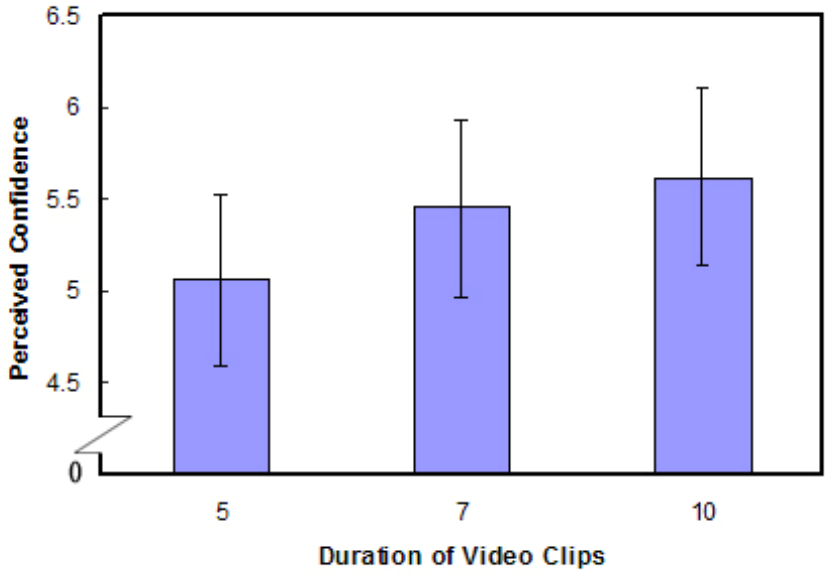


Figure 13. Perceived confidence in being able to monitor 2 video clips as a function of duration of video clips.

A 3 (duration) by 3 (ISI) repeated measures ANOVA was conducted on Q4 regarding perceived confidence in responding to two video feeds. The main effect for duration, ISI, and the interaction between ISI and duration were all not significant (p > .05).

EXPERIMENT 1: DISCUSSION

It was hypothesized that all the tested durations and ISIs would be sufficient for performance and reaction time measures but that a subjective preference would emerge benefiting moderate durations and ISIs (e.g., 7 second duration and ISI).

Duration Results

This original hypothesis was confirmed, in that all duration did not significantly impact participant RT. It was incorrect in that there was a main effect for duration in regards to percent correct. However, by examining the data it was concluded that the performance data was quite noisy. That is, by randomly distributing the video clips, some groups received easier or more difficult clips than others (See Appendix E). Visual inspection of the distribution of errors across videos in the conditions demonstrated that the 10 second duration groups (i.e., with 5, 7, and 10 second ISI) indicated not a random distribution of errors but instead a clustering of errors on just a few videos that proved to be particularly difficult. Therefore, it is believed that the performance duration main effect and the duration by ISI interaction are merely artifacts of random selection of videos without regard to their innate difficulty, which is the focus of study 2. Nevertheless, regardless of the difficulty of the videos that made up a condition it was apparent that user's had adequate time to view the videos as evident by their high accuracy across all conditions (all scores over 80% correct).

In examining the subjective data it was demonstrated that users preferred a shorter video duration (5 seconds) over a longer video duration (10 seconds; the 7 second duration was not significantly different from either of these conditions). In examining confidence in observing an

additional video feed, users rated themselves as less confident for the 5 second condition as compared to the other two conditions. While, this might indicate that a longer duration should be used to improve user confidence, it is actually in the interest of this research program, more experimentally useful to cause a drop in user's confidence in manually performing the task with the addition of added task complexity (i.e., increasing their need for automated aids).

While my initial hypothesis had suggested a 7 second duration, the results of this study indicate that a 7 second duration offers no measurable advantage over the 5 second duration. It was further evident that the 5 second duration did indeed offer a measurable advantage over the 10 second duration. Thus, the 5 second video duration was chosen.

ISI Results

The hypothesis was correct, in that all ISIs were more than adequate given user average response time (M = 0.92 seconds). Further, no significant differences in RT or percent correct were found across the main effects for ISI conditions. A duration by ISI interaction was found for percent correct but as previously mentioned this effect appears to be the result of error caused by the random distribution of video clip difficulties across conditions. However, regardless of the difficulty of the videos that made up a condition it was apparent that user's did have adequate time to respond (all scores over 80% correct).

In examining the subjective measures I found that users preferred the 5 second ISI over the 10 second ISI, which they reported as reflecting too much time to respond (i.e., the task seemed to drag). There was no significant difference in their satisfaction with the time between 5

and 7 seconds. In regards to the perceived workload and confidence data there was no significant difference among any of the ISI conditions.

While the initial hypothesis suggested employing a 7 second ISI, the results of this study indicate that a 7 second ISI offers no measurable advantage over the 5 second ISI. It was further evident that the 5 second ISI did indeed offer a measurable advantage over the 10 second ISI. Thus, the 5 second ISI was chosen.

EXPERIMENT 2: METHODOLOGY

Experimental Purpose

The first study has established that the basic task is set at a pace that allows manual performance. That is, when automation is added in the following experiments users may choose to employ it or they may manually complete the task themselves. However, as the first study demonstrated when performance accuracy was examined, the inherent difficulty of the video clips may not be sufficiently or uniformly sensitive (i.e., restriction of range of the task itself). In order to ensure sensitivity of the performance measure I conducted a second pilot study, using Item Response Theory (IRT; Inman, 2001) to ensure that a range of video clips difficulties (i.e., easy, moderate, and hard discriminations) for stimulus types was selected. The goal of this experiment was to prevent a possible ceiling or floor effect from stimuli difficulty. An additional purpose of having item difficulty quantified is that it allowed me to examine the impact of automation error salience on user reliance.

Experimental Participants

To determine item difficulty I examined the responses of sixty-five undergraduate students. Participants were recruited through the University of Central Florida extra-credit website and they received course credit for their participation. Participation was limited to those with normal or corrected to normal vision, and to those that had not participated in the prior experiment.

Experimental Procedure

Participants first completed an informed consent and demographic questionnaire (See Appendix B and F). Next participants received a short training and practice session, followed by the full experimental session. After completion of the experimental session the participants were thanked for their participation. The entire experiment took approximately 1 hour.

Training Procedure

Training was the same as in experiment 1, with minor exceptions (See Appendix G).

Experimental Task

The same computer-based simulation of a UGV search-and-rescue scenario that was used in experiment 1 was used in experiment 2. However, in this study all video clips were the same 5 second duration and the same 5 second ISI. Additionally, the number of trials was increased to 300. To minimize the effect of a vigilance decrement, participants were offered a short break every 10 minutes of participation. Additionally video clips were presented in a random order to each participant to further prevent a vigilance decrement from influencing only certain video clips.

Experimental Conditions

The properties of stimulus difficulty were examined in this study. That is, the video clips were altered using various levels of added static noise and statistically tested to obtain a range of

stimuli difficulties. The final goal of the experiment was to take the 225 video clips with signals embedded in them (75 videos were empty rooms and served as distracters during the task) and divide them into easy, moderate, and difficult categories for the three signals (i.e., terrorist, civilian, & IED) so that at least 8 clips fell into each category (See Table 10).

Table 10. Division of trial difficulties.

Stimuli	Difficulty	Clips
Terrorist	Easy	8
	Moderate	8
	Hard	8
Civilian	Easy	8
	Moderate	8
	Hard	8
IED	Easy	8
	Moderate	8
	Hard	8
	Total Clips	**72**

Measurement and Analysis

Item Response Theory

Stimuli were mapped for difficulty using the item difficulty index β_i from Item Response Theory (IRT). The index of item difficulty β_i is often used to determine the difficulty of multiple-choice questions; however, in this study it will be used to determine the difficulty of the stimulus (i.e., the video clips). That is the difficulty parameter β_i refers to the proportion of participants who answered an item correctly; thus, the smaller the value of β_i the harder the item (Inman, 2001). The equation for deriving item difficulty is presented below (see Equation 1). According to this equation a difficulty index of 100% indicates that all participants selected the correct

answer and that item was very "easy." A value of 0% indicates that none of the participants selected the correct answer and so that item was very "difficult" (Hotiu, 2006).

Equation 1. Difficulty index formula. Where c is the number who selected the correct answer and n is the total number of respondents.

$$\beta_i = (c/n)*100$$

In this study items were categorized into three distinct levels of difficulty; that is, low difficulty, moderate difficulty, and high difficulty. Item difficulty could range between 0 and 100, with higher values indicating a greater proportion of participants responding correctly to the item (i.e., an easy item). For the purposes of our research easy items were defined as those ranging in β_i from 67 – 100 (i.e., detected by 2/3 or more of the participants), moderately difficult items will have β_i scores from 34 - 66 (i.e., detected by 1/3 to 2/3 of the participants), and difficult items will have β_i scores from 0 – 33 (i.e., detected by 1/3 or fewer of the participants). These difficulty levels were selected across the full possible difficulty range to obtain a full array of item difficulties and prevent restriction of range.

Experimental Equipment

The video stimuli were 300 video clips obtained from three commercial and educational office buildings (75 clips of each potential stimulus: terrorist, civilian, IED, and empty). All features of the videos and interface were the same as those used in experiment 1, the only difference being that the duration of the videos were held constant in this study and the amount

of noise added to the clips in Adobe Premiere 2.0 varied to improve the differentiation between the difficulty levels.

Hypothesized Outcome

The outcome of this study will be a division of videos based on item difficulty that will be employed in experiments 3 and 4 of this dissertation. The purpose of this is to prevent restriction of range in the performance measure of these studies. Additionally, it is of interest to quantify item difficulty so that the impact of automation errors upon subsequent user reliance, in experiment 4, may be related to the salience of that error (with easier items being typically more salient than more difficult items).

EXPERIMENT 2: RESULTS

The purpose of the second experiment was to select the videos, based on their difficulty, for use in experiments 3 and 4. Item difficulty was determined for each of the civilian, terrorist, and IED clips. The resulting item difficulties for each of the 225 clips are presented in Appendix H.

The results of experiment two were not as predicted. People were far better at picking critical signals out of the video clips than originally anticipated. Initially it was desired to include 8 easy (more 2/3 participants detect), 8 moderate (more 1/3 less 2/3), 8 hard video clips (less 1/3 correct detection) from each kind of stimuli (terrorist, civilian, and IED) for a total of 72 signals. Unfortunately the data did not cooperate, after removing all 100% detection rates (which were not diagnostic; there were no 0% detection rates), there were five out of 9 divisions that did not contain the minimum number of clips (see Table 11). These results required that the video inclusion criteria be altered to allow for an equal selection of video clip difficulties while maintaining adequate number of trials for study power.

Table 11. Division of type of video clips into difficulty levels.

Stimuli	Difficulty	Clips Needed	Actual Clips
Terrorist	Easy	8	53
	Moderate	8	0
	Hard	8	0
Civilian	Easy	8	58
	Moderate	8	7
	Hard	8	1
IED	Easy	8	51
	Moderate	8	18
	Hard	8	5

Note: Video clips with 100% detection were removed as they were viewed as being not diagnostic.

EXPERIMENT 2: DISCUSSION

It was hypothesized that there would be sufficient video clips across the full spectrum of possible signal types and item difficulties (i.e., 8 clips from each difficulty level for each of the three types of signals: terrorist, civilians, and IEDs). However, results indicated that participants exceeded performance expectations and that the detection rate across subjects was quite high (M = 87.97, SD = 6.38). Thus, there were fewer hard and moderate difficulty clips than anticipated (i.e., only 6 total video clips meet current requirement for high difficulty compared to the desired 24 clips). Thus, several adaptations had to be made to the methodology.

First, due to effects outside of our experimental controls (e.g., human-beings exceptional detection of biological motion) no terrorist clips fit into either the hard or moderate difficult classification (see Table 11). Thus, in order to maintain the measure of item difficulty, signal type was collapsed. This was a viable solution since hypotheses for study 4 were concerned with the difficulty of the detection more so than 'what' per say was being detected.

Therefore, the selection of equal numbers of each type of stimulus in each difficulty level was abandoned and instead equal numbers of video clips in general from each difficulty level was used. However, due to the excellent detection rate across subjects (M = 87.97, SD = 6.38) there were fewer hard and moderate clips than anticipated. Study plans had called for 240 clips with a 30% event rate, thus requiring 72 videos with embedded stimulus (i.e., terrorist, civilian, or IED). However, even after collapsing over stimulus-type there were only 6 rather than the anticipated 24 high difficulty clips. This would substantially reduce the number of video clips in the following two studies from 240 to only 60 total video clips. Thus to increase the number of potential total videos (i.e., trials) in the following two studies the difficulty index associated with

low, moderate, and high difficulty was slightly adjusted. So that, low difficulty index became 75-100 (i.e., more than 75% of participants detect), moderate difficulty index 50-75, and high difficulty index 25-50. This resulted in the exclusion of a single video clip that had a 17 difficulty index. By adjusting difficulty level slightly, I was able to maintain a natural progression in difficulty of the video clips, but double the number of future trials from 60 (with 18 signals) to 120 (with 36 signals).

EXPERIMENT 3: METHODOLOGY

Experimental Purpose

The first and second experiments have been concerned with the stimulus durations, the given ISI for responding, and the difficulty of the stimulus. The third experiment will now examine the addition of an automated decision aid to a participant monitoring a single agent in the search-and-rescue task. The purpose of this study was to establish appropriate high- and low-reliability levels for the automated-aids. While, aid reliability levels in the literature can vary a great deal, they are often task dependent. In order to maximize the potential effects of conflicting reliability levels in experiment 4 (i.e., improve power of the aid mixed-reliability manipulation) I tested seven potential reliability levels and compared them for user automation reliance and perceived automation trust.

Experimental Participants

To obtain 20 participants per reliability condition and a control condition (i.e., no aid), one-hundred-forty participants were recruited through the University of Central Florida. Participants received either course credit or cash payment for their participation (equivalent to 1pt extra credit or $5). Participation was limited to those with normal or corrected to normal vision, and who have not participated in any of the prior experiments. Participants were randomly assigned to one of the seven conditions with the restriction that equal genders were present in each group (70 male, 70 female). Average age of the participants was 21 years old (SD = 5).

Experimental Procedure

Participants first completed an informed consent and demographic questionnaire (see Appendix I and F). Next, participants received a short training session (see Appendix J and K). Finally, participants completed the experimental session, which was composed of 120 trials (approximately 20 minutes). After completion of the experimental session the participants completed the exit questionnaire (see Appendix L) and were thanked and compensated for their participation. The entire experiment took approximately 30 minutes to complete.

Training Procedure

The purpose of the training was to acquaint the participants with the basic task, as before, but also to familiarize them with the automated aid. This was accomplished using an experimenter read script (see Appendix J), a follow-along mission folder (see Appendix K), and a computerized practice session. The script and mission folder described in detail the scenario, what stimuli the participant would view, how to respond, and how the automated aids worked. The computerized practice session had the same layout as the experimental display (see Figure 14). For practice the participants were presented with 8 video clips, 4 without the aid (see Figure 15) followed by 4 with the use of the aid (see Figure 16). The video clips were all of easy difficulty and were drawn from the four types of potential stimuli, such that each type of stimulus appeared once without the aid and once with the aid. Participants were to respond to the practice trials to become familiar with the interface.

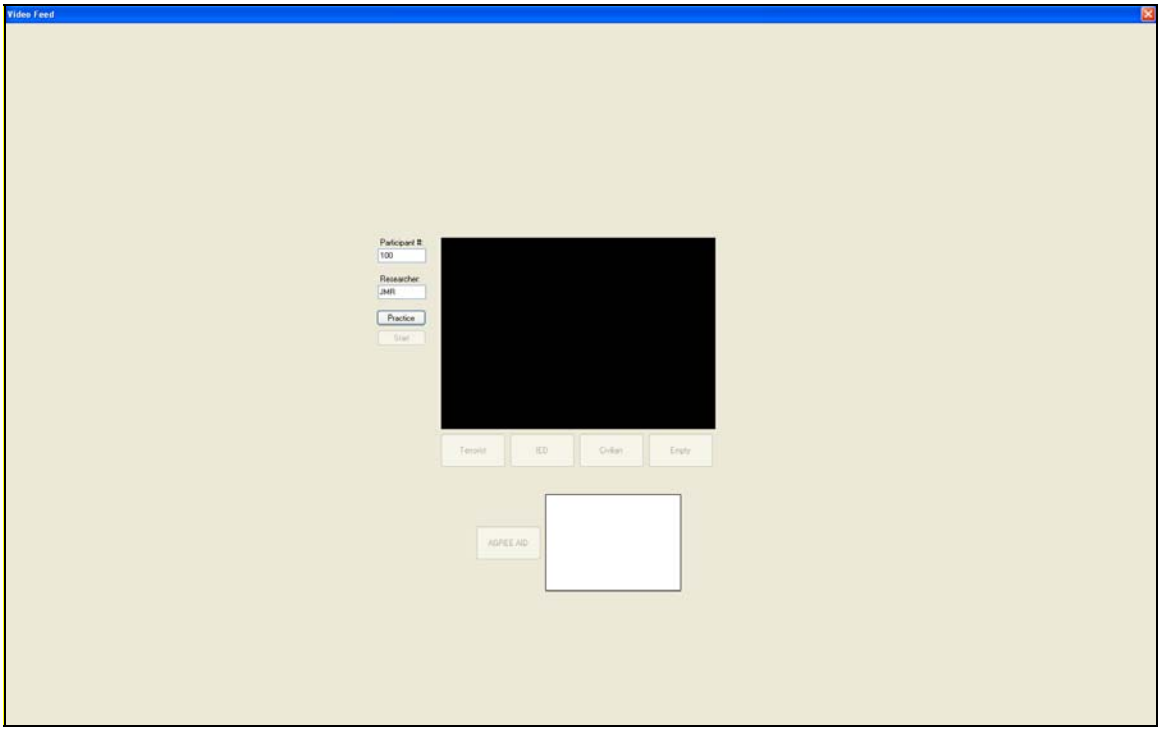


Figure 14. Practice interface for experiment 3.

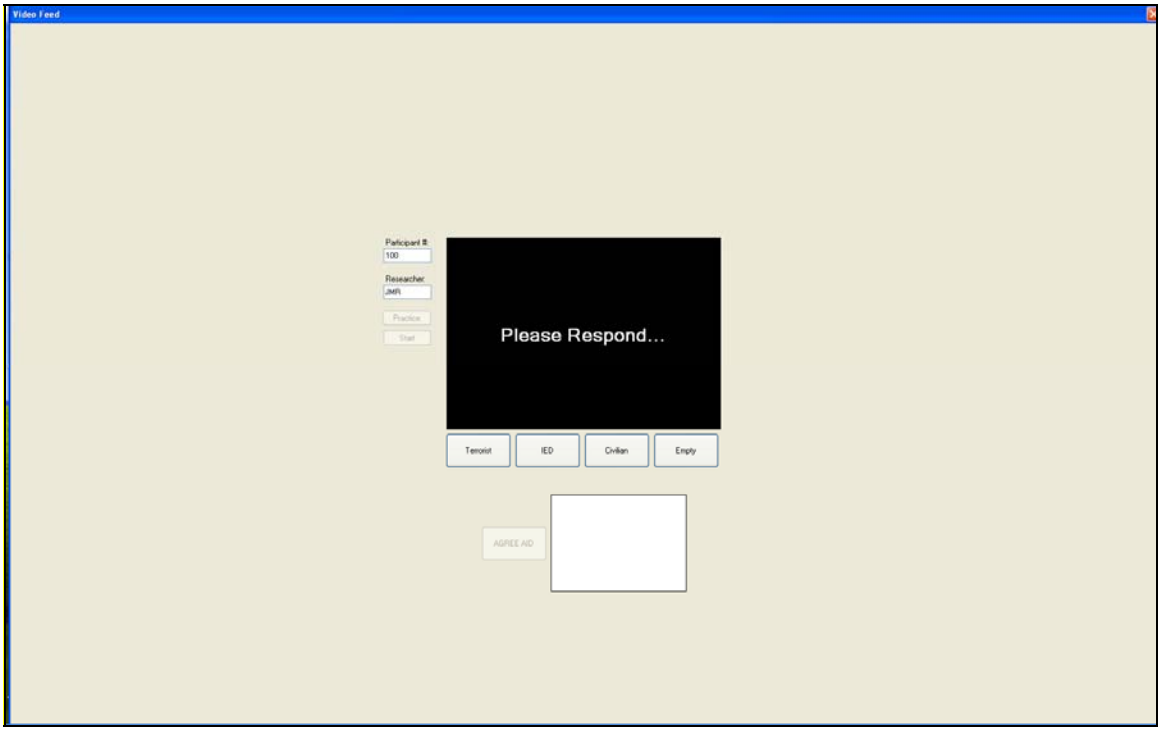


Figure 15. Experimental interface experiment 3 without the aid.

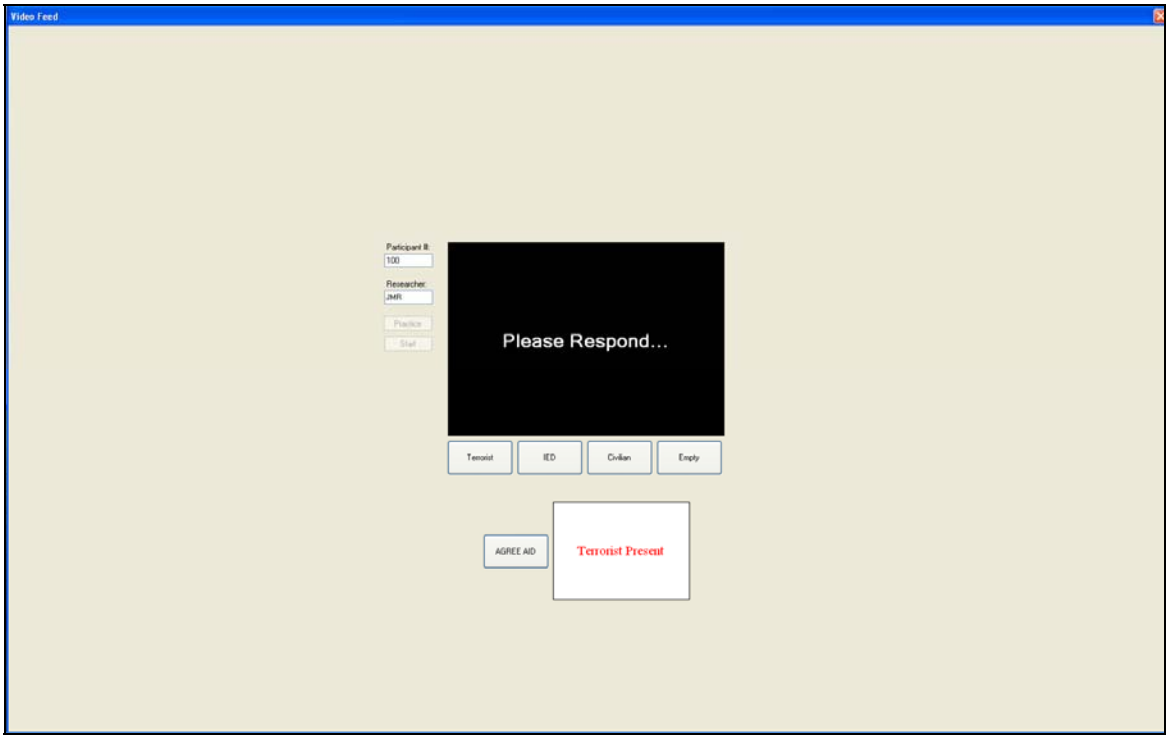


Figure 16. Experimental interface with the automated-aid. Note that: Aid recommendation reads "Terrorist Present."

Experimental Tasks

Participants were given the same basic search-and-rescue scenario from the prior experiments, with the addition of an automated decision aid. Participants were told that the automated aid works as a 'contrast detector' using an algorithm to identify certain patterns such as civilians, terrorists, and IEDs in complex scenes. Users were informed that use of the aid was completely optional and that the responsibility of the final decision was their own and that they could choose either to accept the aid's proposed diagnosis or to ignore it. Users were not informed about the aid's reliability.

Experimental Conditions

The property of aid reliability was examined in this study. The aid had a set reliability of either 75%, 80%, 85%, 90%, 95%, 99%, or Control (i.e., no aid recommendations). The aid had occasional misses and false alarms, within each reliability condition the number of misses and false alarms were equal (see Table 12). In all cases one-third of the trials (36 trials of the total 120) contain an embedded signal (i.e., terrorist, civilian, or IED). Automation errors were randomly distributed throughout the automation so as to prevent operators from developing a strategy for compensating for the automation errors. It is important to stress here that all participants received the same number of embedded signals the only variation is the accuracy of the automated decision aid in detecting those embedded signals.

Table 12. Reliability level false alarms and miss rates

Reliability Level	False Alarms	Misses	N
99%	1	1	20
95%	3	3	20
90%	6	6	20
85%	9	9	20
80%	12	12	20
75%	15	15	20
Control (No Aid Recommendations)	N/A	N/A	20

Measurement and Analysis

For performance I examined reliance (which is defined as the percent of times the users decision matched the aids decision) and performance. The two reliability levels chosen for the fourth study must have significantly different reliance, with higher reliance for the high-

reliability condition and lower levels of reliance for the low-reliability condition. Additionally, in terms of performance control performance (% correct) should be higher than the actual reliability of the low-reliability aid and lower than the actual reliability of the high-reliability aid.

In terms of trust I examined subjective evaluations of perceived trust after interacting with the automation (see Appendix L). The subjective trust ratings were based on the self-report measures used by Dzindolet et al. (2003) and Master et al. (2000) Dzindolet et al (2003) and administered after participants interacted with the automation (see Table 5). In the questionnaire participants were asked to rate their perceived trust in the automated decision aid. The two reliability levels chosen for the fourth study must have significantly different perceived trust, with higher trust for the high-reliability condition and lower levels of trust for the low-reliability condition.

Experimental Equipment

As in the previous studies the simulation was presented on a 20" widescreen monitor on a desktop computer. Participants responded using a mouse. The interface was created using VisualBasic.net.

Hypothesized Outcome

The outcome of this study was the selection of a high- and low-reliability level for study 4. The purpose of this was to improve the measure of mixed reliability in experiment 4 (i.e., have improved power of the measure). To accomplish this purpose it is required that the aids differ in perceived trust, reliance, and performance. It is hypothesized that differences will be consistently

obtained between reliability levels of 99% and 75%; however, it has been shown that if automation is faulty beyond a certain point operators will completely ignore it and focus solely on manual control. To prevent complete misuse of the low-reliability aid in experiment 4, it is desirable to use as high a reliability level for the low-reliability aid as possible that still maintains significantly less reliance and trust compared to the high-reliability condition. A final restriction is that the actual reliability for the low-reliability condition must be below control user performance and the actual reliability for the high-reliability must be above control user performance.

EXPERIMENT 3: RESULTS

The purpose of the third experiment was to ensure appropriate high- and low-reliability levels for the automated aids in experiment 4. While, aid reliability levels in the literature can vary a great deal, they are often task dependent. Thus, in order to maximize the potential effects of conflicting reliability levels in experiment 4 (i.e., improve power of the aid mixed reliability manipulation) I tested six potential reliability levels and compared them for performance, reliance, and trust differences.

Performance and Behavioral Data

Percent Correct

It was required that the set reliability of the low-reliability aid be below average operator performance. Since, the control group indicated that average operator performance on this task was around 82% accuracy (SD = 5%), the 75% reliability level was selected to serve as the low-reliability level, as it was the only reliability level below average user performance. It was also required that the set reliability level of the high-reliability aid be significantly above average operator performance, this criteria was satisfied by the 90%, 95%, and 99% reliability conditions. However, these values were for the set reliability (i.e., actual reliability of the aid), a univariate ANOVA was conducted on all 140 participants for overall performance accuracy (i.e., how participants calibrated their performance with that of the aid; See Figure 17). A significant effect for reliability of the aid on user overall performance, as measured by percent correct, was

found, F(6, 133) = 9.72, p < .0005, η² = 0.31. Correlation data indicated that as aid reliability increased so did user performance (*r* = .34).

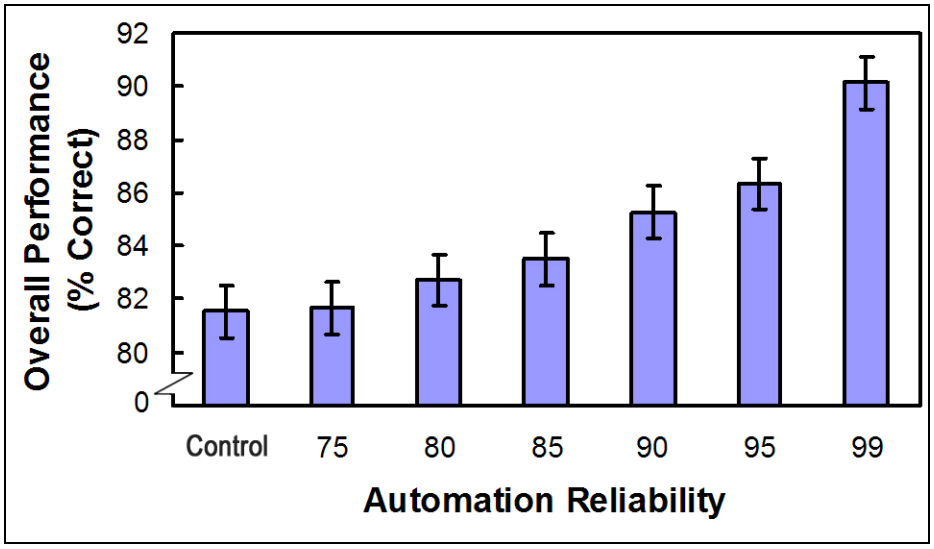


Figure 17. Percent correct as a function of automation reliability. Note that the control group that received no automated recommendations.

Reliance

A univariate ANOVA was performed on the 120 participants who interacted with the aid to examine reliance, as measured by the number of times the participant agreed with the automated aid. Aid reliability was found to have a significant effect on participant reliance, $F(5, 114) = 19.62$, $p < .0005$, $\eta^2 = 0.46$. Correlation data indicated that as aid reliability increased so did user reliance on the aid ($r = .66$; see Figure 18). Given that the 75% reliability condition has been selected for the low reliability, it is important that the high reliability condition is relied on significantly more than the 75% reliable condition. All reliability conditions, except the 80% reliable condition, had significantly higher reliance than the 75% reliable condition ($p < .05$ in all cases).

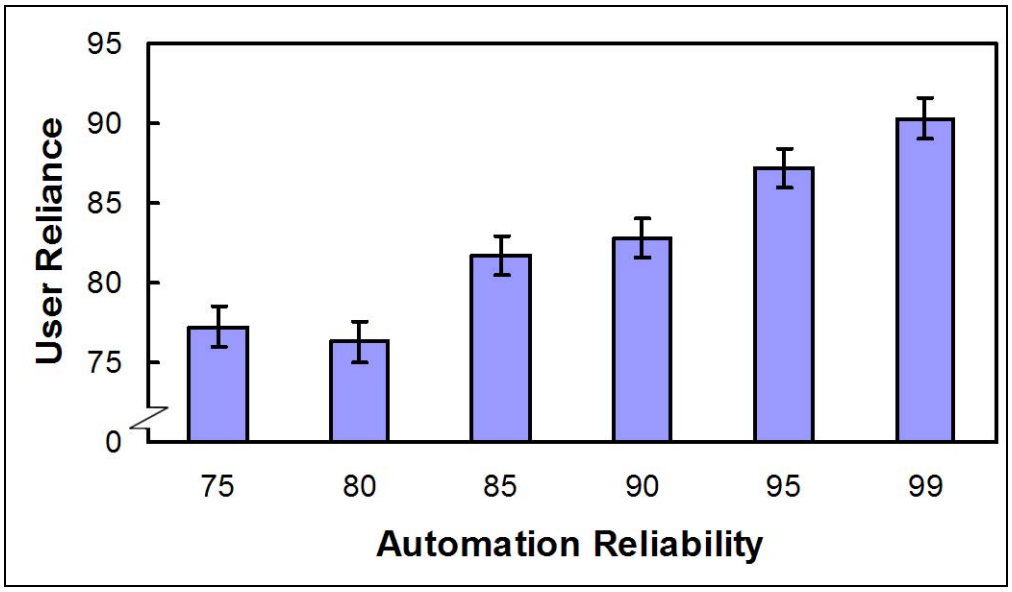


Figure 18. User reliance as a function of automation reliability. Note that user reliance is measured as the percent of time the participant agreed with the aid.

Subjective Data

Perceived Trust

A univariate ANOVA was performed to examine trust of the aid, as measured by a 9-point Likert scale, with higher numbers reflecting greater trust. Aid reliability was found to have a significant effect on participant trust, $F(5, 114) = 2.86$, $p = .018$, $\eta^2 = 0.11$. Correlation data indicated that as the aid's reliability increased, so did user perceived trust of the aid ($r = .29$; see Figure 19). Given that the 75% reliability condition had been selected for the low-reliability, it was important that the high-reliability condition garnered significantly more trust than the 75% reliable condition. Only the 95% and 99% conditions had significantly higher levels of self-reported trust compared to the 75% reliable condition ($p < .05$ in all cases).

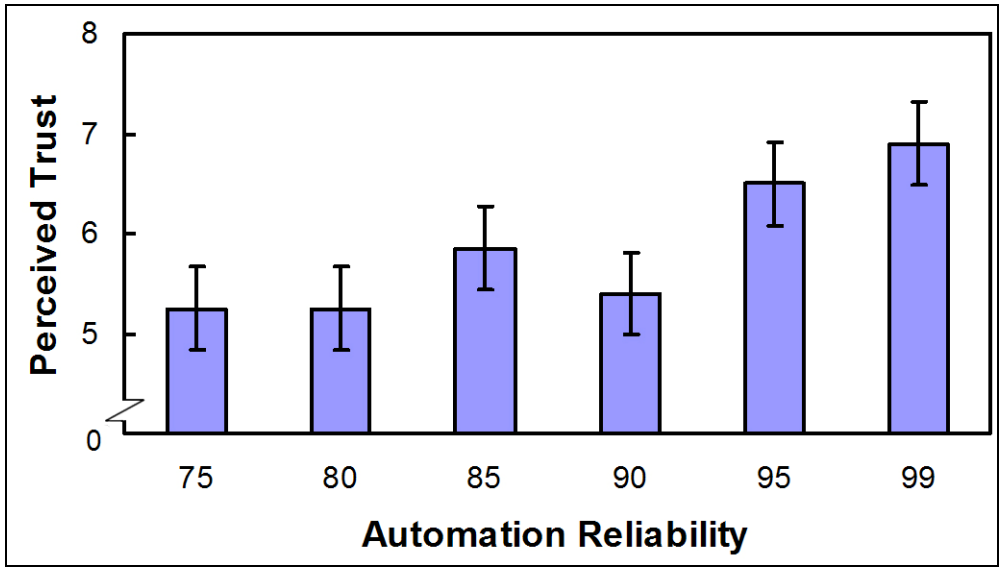


Figure 19. Participant perceived trust as a function of reliability of aid.

EXPERIMENT 3: DISCUSSION

Given the findings for experiment 3 it was decided that the low-reliability condition would be 75% based on the fact that it was the only reliability condition with an actual set reliability level *below* average user performance. Either the 95% or 99% reliability conditions would work for the high-reliability aid, in that they both had set reliabilities above average user performance, both were significantly more trusted than the low-reliability aid, and both were relied upon significantly more often than the low-reliability aid. However, it was decided to go with the 95% reliable aid, as it would allows for 6 automation errors during the 60 trials in study 4 (3 for each agent) and thus allow the examination of error salience (i.e., high difficulty, moderate difficulty, or low difficulty) on subsequent automation reliance. Whereas, use of the 99% reliable measure would allow for only 2 automation errors during the 60 trials (1 for each agent), thus forcing the measure of error salience to be dropped from study 4 (since making error salience a between-subjects measure would be prohibitive in terms of the increase to sample-size; i.e., from 300 to 840 participants).

EXPERIMENT 4: METHODOLOGY

Experimental Purpose

The first three experiments resulted in the creation of a test bed for the fourth experiment; which examined operator trust and reliance on automatic decision aids when working with multiple agents. In this experiment, participants monitored two video feeds and two concurrent automated decision agents. These agents were manipulated in terms of their reliabilities and agent type. The purpose of this study was to examine how inappropriate biasing of trust and reliance calibrations occur when an operator is exposed to two agents of different reliabilities (e.g., does disuse of a high-reliability aid occur when combined with a low-reliability aid and does misuse of a low-reliability aid occur when combined with a high-reliability aid). Additionally, it was of interest to examine whether this biasing effect was influenced by the perceived independence of the agents. That is, can 'what' one believes the agents are, influence how one reacts to them (i.e., reliance) and thinks of them (i.e., perceived trust)? In the following study this question was examined by looking at three levels of agent independence (i.e., two human agents – highest independence, two different-type robotic agents – moderate independence, and two same-type robotic agents – intermediate independence; see Figures 20, 21, & 22) and three levels of reliability (uniform low, mixed, and uniform high).

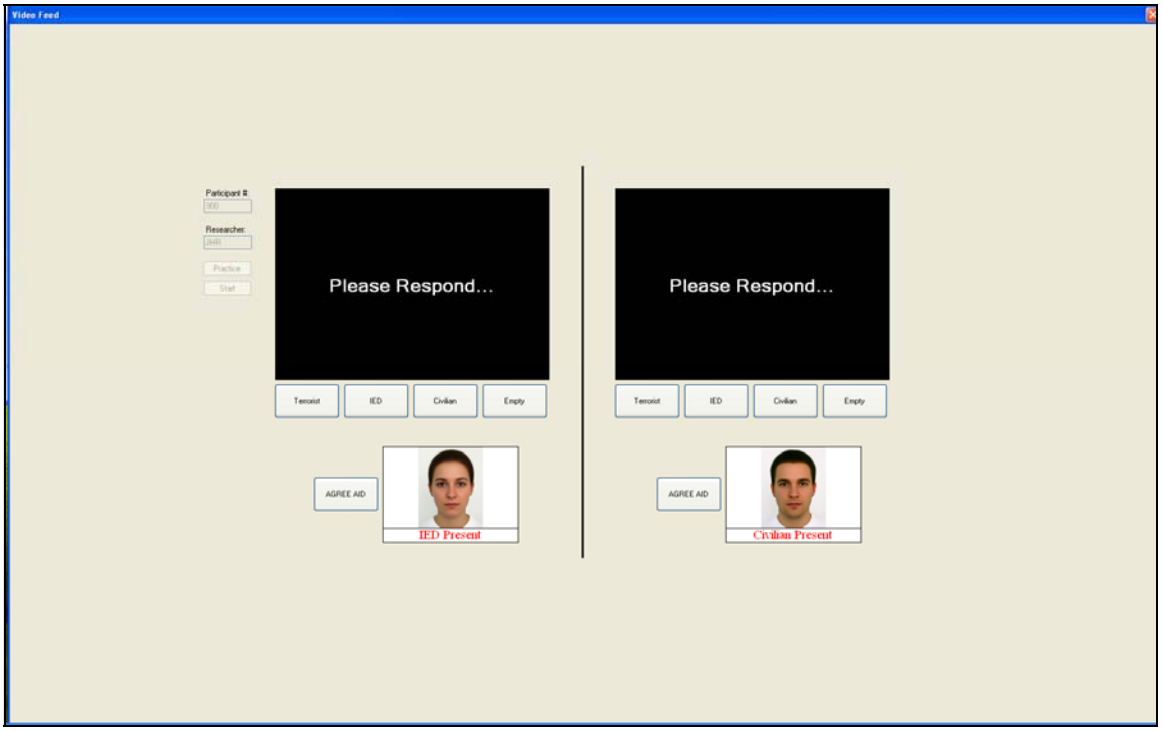


Figure 20. Human agent condition.

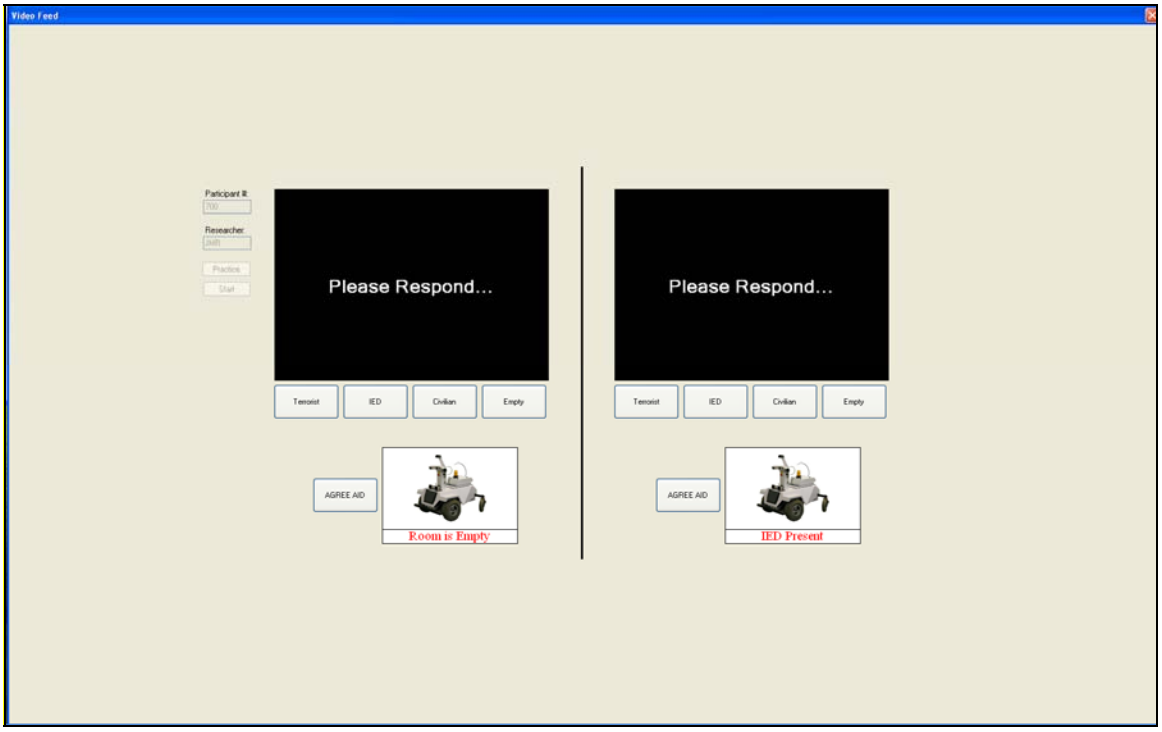


Figure 21. Same-type robotic agent condition.

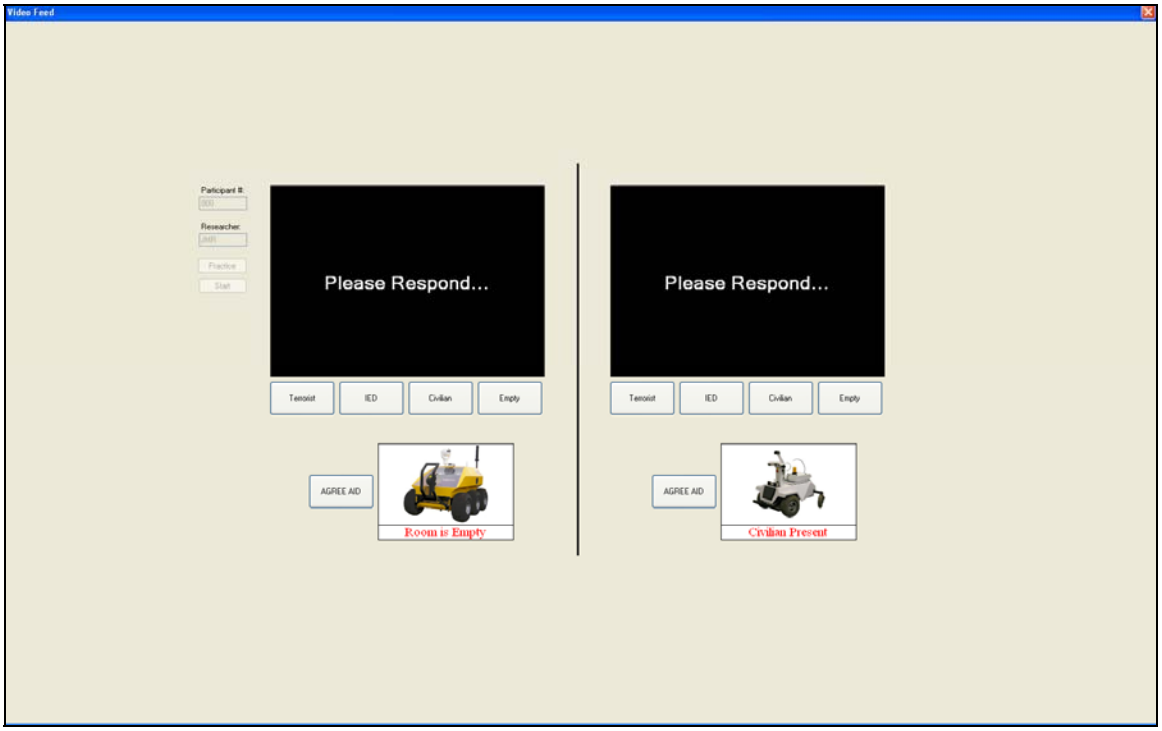


Figure 22. Different-type robotic agent condition.

Experimental Participants

A total of 330 participants (150 males, 180 females) from the University of Central Florida volunteered to participate in the study, this ensured adequate power of measurement (assuming $\Delta_I = 0.55$, $\alpha = .05$, & $\beta = .20$). Participants were compensated with course credit or cash payment for their participation (2pts course credit or $8 paid). Participation was limited to those with normal or corrected to normal vision and to those who have not participated in any of the prior experiments. Participants ranged from 18 to 57 years of age, with most subjects being close to the mean age of 21 years (SD = 5).

Due to the large sample size the laboratory was set-up to allow running of up to eleven participants at a time. Cubicle dividers and noise-canceling headphones were employed to

mitigate any visual or auditory interference between participants. Participants were randomly assigned to the cells of a 3 (source characteristics: humans, generic machines, unique machines) x 3 (reliability: both low, mixed, both high) between participant design (or a control condition), with the restriction that equal genders were equally distributed in each condition. One male participants data was lost due to a technical failure and the following results are thus based on 329 participants.

Experimental Procedure

Participants were tested in groups ranging in size from 1 to 11. Regardless of size of the group participants completed the same experimental order. That is, they first completed an informed consent (see Appendix M), demographic questionnaire (see Appendix F), anthropomorphism questionnaire (ATS; see Appendix N), interpersonal trust scale (ITS; see Appendix O), and complacency potential rating scale (CPRS; see Appendix P). Next participants completed a short training session (see Appendix Q) followed by a trust pre-questionnaire (see Appendix R). Finally, participants completed the experimental session, which entailed monitoring two video feeds with agent recommendations for 60 trials each (10 minutes; see Figure 23). After completion of the experimental session the participant completed three exit questionnaires. One questionnaire queried participants on their own performance by asking them to rate their own self-confidence in performing the task and to complete the NASA-TLX (which was computer based; see Appendix S). The other two exit questionnaires queried the participants on their trust in their Teammate A and Teammate B (see Appendix T). After completing the exit questionnaires participants were debriefed on the nature of the study (See Appendix U),

compensated, and thanked for their participation. The entire experiment took approximately 1 hour to complete.

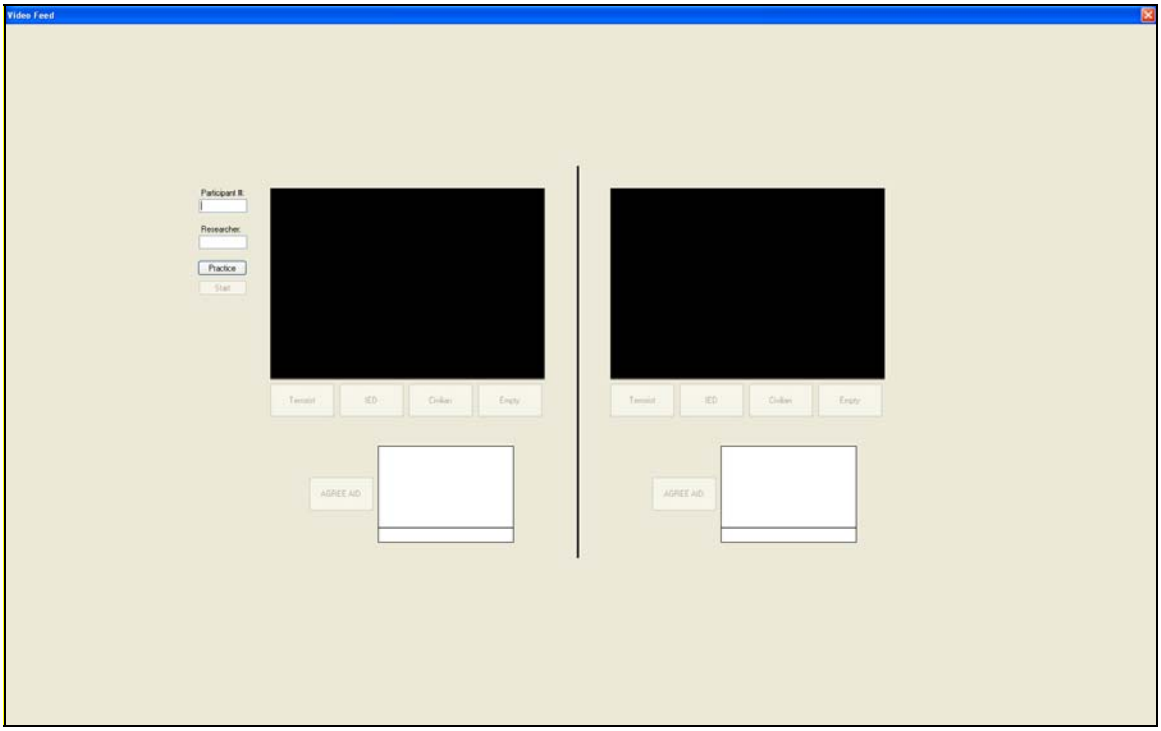


Figure 23. Experimental interface experiment 4.

Training Procedure

The training for experiment 4 was identical to the training session for experiment 3 except that participants were instructed on performing two monitoring tasks concurrently and informed that they would be interacting with a particular kind of agent. This was accomplished by using an experimenter read script (see Appendix Q) and a computerized practice session (see Figure 24). The practice session presented 8 sets of video clips, the first four without the aid of a teammate and the last four with the aid of a teammate (see Table 13). Participants were to

respond to the practice trials to become familiar with the interface. Participants were informed that the practice sessions were preprogrammed for demonstrational purposes and did not reflect the recommendations of their future teammates. During the practice session decision aids were held at 100% reliable.

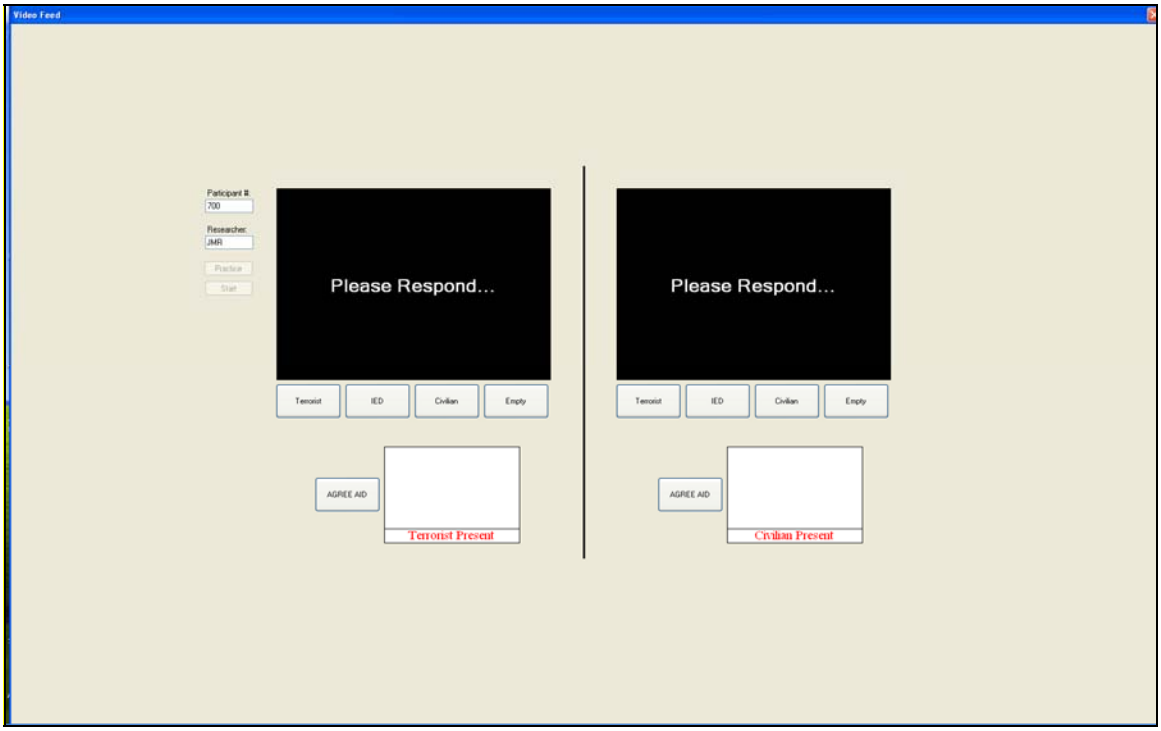


Figure 24. Practice interface for experiment 3.

Table 13. Video orders for experiment 4 practice session.

Trial	Video 1	Video 2
1	Terrorist	Empty
2	Empty	Civilian
3	IED	Empty
4	Empty	Empty
5	Terrorist	Terrorist
6	Civilian	Civilian
7	IED	IED
8	Terrorist	Civilian

Experimental Tasks

Participants were given the same basic search-and-rescue scenario from experiment 3, with the addition of a second video feed and automated decision aid (See Figure 23). The size of the display was held constant. The instructions participants received in the training condition differed depending on whether they were in the same-type robotic aid (see Figure 20), different-type robotic aid (see Figure 21), or human condition (see Figure 22). Participants in the robotic aid conditions were informed that they would be monitoring the responses of two robotic agents; they were informed that the robotic agents made decisions based on mathematical algorithms. To maximize the perceived difference between different-type and same-type robotic aids their different nature was emphasized in the instructions and also the UGVs were represented by either two of the same-type or two different-type robots (see Figure 25). The robots were both wheeled prototypical robots that differed in color and exact form. On the other hand participants in the human condition were informed that they would be interacting with two students who had previously completed this study. It was stated that the students had previously completed the study to account for the fact that their pictures were employed in the simulation. The pictures of the two students were actually facial compilations of 65 female faces and 35 male faces to give an ‘average’ male and an ‘average’ female. Averaged faces were used to minimize the chance of

participants recognizing the 'student' and to provide a level of control for the manipulation of human agents (see Figure 26). Male and female faces were used to allow for the examination of any differences in trust and reliance on the agents based on sex characteristics of the operator and the agents.

Figure 25. Robotic teammates. Note that robots were counterbalanced so that half of the participants in the same-type aid received the yellow robot and half the white robot.

Figure 26. Human agent facial compilations for male and female teammates.

Regardless of the specifics of the agent type (i.e., whether they are distributed human agents, unique computer agents, are identical computer agents) the participant received the same experimental task and always played the role of monitor (i.e., observing the scenario, and the recommendations by the agents, to make a final decisions). Users were explicitly informed that use of the aids was completely optional and that the responsibility of the final decision was their own and that they could choose either to accept the agents proposed diagnosis or to ignore them. This low level of automation was used as it has been argued that trust is only relevant in situations that can be characterized by a certain degree of free will in placing oneself in a situation of risk (de Vries et al., 2003). That is, the users are free to agree with or ignore the automation, but the automation will not take action independently of the operator. Participants were not informed about the agent's reliability levels.

Experimental Conditions

The properties of aid reliability and source characteristics were manipulated in this study. Agents were set at either the same reliability level (either low or high) or mixed reliability (one agent operates at high reliability and the other at low reliability). Additionally, participant attributions of the agent were manipulated so that they believed they are working with human teammates, same-type robotic teammates, or different-type robotic teammates. An additional condition in which the operator received no teammate recommendations served as a control. This results in a 3 by 3 between-subjects ANOVA (plus control condition). Between-subjects was used because it was believed that participants would be influenced by switching reliability levels

and agent source would become less effective as a within-subject variable (i.e., that agent source as a manipulation would become less believable if within).

Measurement and Analysis

It is critical to obtain subjective measures to measure the psychological construct of automation trust, as well as behavioral data to evaluate automation reliance, since often times performance and subjective measures are imperfectly calibrated (Wiegmann, Rich, & Zhang, 2001).

Subjective Measures

Exit questionnaires were administered to evaluate perceived workload, trust, and self-confidence based on interacting with the automated agents (see Appendix S & T). Automation trust, self-confidence, and perceived reliability of the aids were obtained using 9-point Likert-type scales.

The literature provides evidence that it is important to examine how personality differences (e.g., generalized trust expectancies, anthropomorphic tendencies, and complacency potential) affect trusting behavior. For example, studies have shown that those who score highly on interpersonal trust are generally more cooperative with other people (Rotter, 1967), it would be interesting to examine if interpersonal trust is related to being more cooperative (i.e., higher reliance) with robotic aids. One method to do this is by employing the Interpersonal Trust Scale (ITS; Rotter, 1967).

The ITS is a 25 item questionnaire that examines an individual's level of interpersonal trust. Some of the items on the scale measure trust in a variety of social objects and some items measure general optimism regarding society. Of the 25 trust items, 12 are written so that an "agree" response indicates trust and 13 are written so that a "disagree" response indicates trust (Rotter, 1967). The items use five Likert response categories from (1) strongly agree to (5) strongly disagree. Scores can range from 25 (lowest trust) to 125 (highest trust), with a neutral score or midpoint of 75. Test-retest reliability for the questionnaire has been found to be .56 or .68 (Rotter, 1967). The scale was designed to measure one's expectation that the behavior, promises, or (verbal or written) statements of other individuals can be relied upon (Wrightsman, 1991). The ITS is not significantly related to intellectual aptitude, but have been found to be related to birth order (youngest lower trust), self religion (any religious beliefs reflects greater generalized trust), parents religion (individuals with parents of differing religions have lower interpersonal trust scores compared to those whose parents are of the same religion), and socioeconomic level (individuals in lower socioeconomic levels have lower ITS scores compared to individuals in higher socioeconomic levels; Rotter, 1967). Additionally, scores on the ITS have been related to the likelihood of giving others a second chance (Rotter, 1980), but not to gullibility or dependence (Rotter, 1967).

Participants were also given the Complacency Potential Rating Scale (CPRS). The CPRS is designed to assess attitudes (favorable and unfavorable) toward everyday automated devices (e.g., automatic teller machines). An attitude can be defined as a personal disposition common to individuals, but possessed by them to different degrees, which impels them to react to objects or situations in favorable or unfavorable ways (Singh, Molloy, & Parasuraman, 1993). That is, the CPRS is designed to measure one's attitude toward automation (e.g., overconfidence) which may

in particular situations (e.g., high workload, routine, repetition) lead to complacent behavior. The concept of complacent behavior is defined by Parasuraman, Molloy, and Singh (1993) as inaccuracy and/or delay in detecting a failure in an automated system. The CPRS measures this attitude with an internal consistency (r = .87), overall reliability (r = .90), and test-retest reliability (r = .90). The scale measures four main factors which lead to complacency, they are: confidence, reliance, trust, and safety. This scale is composed of 12-items, each measured by a 5-point Likert-type scale with anchors ranging from strongly disagree (1) to strongly agree (5). As these anchors were the opposite direction of the ITS all participants were cautioned of the conflicting anchors prior to filling out the questionnaires. Mean CPRS scores in validation research were 57.69 (SD = 6.09), and scores range from 40 to 75.

Objective Measures

Automation reliance was determined by examining the agreement probabilities of the operator with the agents. Temporal reliance was determined by examining the likelihood of agreement with a correct aid recommendation following an aid error. Distribution of the errors was constrained so that each error was followed by a correct automation recommendation.

Experimental Equipment

Decision agents were referred to as ‘Agent A’ and ‘Agent B’ during the duration of the experiment. This was done to emphasize that the agent is conducting an activity that could conceivably be done by a person or machine, and to reflect the collaborative nature of the

operator's interaction with the agents (Beck, Dzindolet, & Pierce, 2002; Bowers, Oser, Salas, & Cannon-Bowers, 1996; Bubb-Lewis & Scerbo, 1997; Scerbo, 1996; Woods, 1996).

Agents were set to reliability levels as determined in experiment 3: 75% for the low-reliability agent and 95% for the high-reliability agent. Depending on the condition assigned participants interacted with two high-reliability agents, two low-reliability agents, or two agents of mixed-reliability. Reliability level was held constant for the duration of the experiment.

In regards to the interface for Experiment 4 (see Figures 20, 21, & 22), there were illustrations next to each agent recommendation. These illustrations were employed because past research has shown that teammates using video channels or face-to-face interaction established trust and cooperation more quickly than did teammates using only textual communication (Corritore, Kracher, & Wiedenbeck, 2001). Furthermore, in regards to the interface at large, past research using internet websites found that design quality as composed of strict grouping, formal language, the use of real photos, and employing empty space as a structural element, has been found to improve perceived trust (Karvonen & Parkkinen, 2001). Thus the use of these structural elements was utilized to minimize the negative effect of overall visual impression on perceived trust of the system, allowing participants to focus on the rational evaluation of the decision aids themselves (i.e., the utility and source of the recommendations) to guide their use of the agents.

The agents themselves were set at automation level 5 according to the level of automation classification of Parasuraman, Sheridan, and Wickens (2000; see Table 14). Automation level 5 was chosen based on research by Young & Stanton (2001) that found that ideally technological support systems should act like a driving instructor in the passenger seat – subtle enough so as not to cause interference, but accessible enough so as to provide assistance when needed. That is, the automation offered the operator a recommendation but did not automatically execute that

recommendation. Thus, the operator had to commit a voluntarily action of trusting the agent (Corritore, Kracher, & Wiedenbeck, 2001). For example, if an aid identified a terrorist agent in a video clip the operator had a limited amount of time to approve the automation's recommendation or enter their own decision before moving on to the next trial. Additionally, operators were told that using the automation was optional, and they could accept or ignore the automation on each trial during the experiment. Participants were not informed of the reliability level of the automated agent. Thus, the difficult position of determining whether or not one should rely on the decision aid was placed entirely upon the participant.

Table 14. Table of automation levels (adapted from Parasuraman, Sheridan, and Wickens, 2000)

Automation Level	Description
10	The computer decides everything, acts autonomously, ignoring the human.
9	informs the human only if it, the computer decides to
8	informs the human only if asked, or
7	executes automatically, then necessarily informs the human, and
6	allows the human a restricted time veto before automatic execution, or
5	executes that suggestion if human approves, or
4	suggests one alternative
3	narrows the selections down to a few, or
2	The computer offers a complete set of decision/action alternatives, or
1	The computer offers no assistance: human must make all decisions and actions.

Hypothesized Outcome

There were six central hypotheses to experiment 4 (see Table 15).

1. In a complex, dual-aid, condition there will be bias between two agents of mixed reliability compared to two uniform agents.
    a. Trust and reliance of a high-reliability agent will be negatively influenced by a concurrent low-reliability agent.

b. Trust and reliance of a low-reliability agent will be positively influenced by a concurrent high-reliability agent.

2. Operators experiencing high automation reliability will have significantly more subjective trust in the automation than those experiencing both low or the mixed reliability conditions. Additionally those with low automation reliability will experience significantly less subjective trust of the automation than those in the mixed reliability condition (H_O = There is no significant difference between reliability group trust scores).
    a. Increased levels of automation trust will be accompanied by increased levels of reliance on the aid and lower levels of reported workload.
    b. Decreased levels of automation trust will be accompanied by decreased levels of reliance on the aid and higher levels of reported workload.
3. Subjective levels of trust, automation reliance, and workload are expected to differ across agent type (i.e., human, similar computer agents, dissimilar computer agents). Such that human agents have increased trust, increased reliance, and decreased workload, compared to the computer agents. The computer agents are not expected to differ in overall trust, reliance, or workload (H_O = There is no significant difference between agent type group trust ratings, reliance, and/or workload).
4. In a mixed reliability condition the agent type is expected to significantly impact crossover bias between the two agents. (H_O = There is no significant interaction between reliability and agent type).
    a. Two agents perceived to be human will experience the least crossover bias in the mixed reliability condition. Thus, a low-reliability human aid will have little impact on a concurrent high-reliability human aid.

b. The same-type robotic agents will experience the most crossover bias in the mixed reliability condition. Thus, a low-reliability same-type robotic agent will have a strong impact on a concurrent high-reliability same-type robotic agent.

c. The different-type robotic agents will experience an intermediate level of crossover bias in the mixed reliability condition. Thus, a low-reliability different-type robotic agent will have an intermediate impact on a concurrent high-reliability different-type robotic agent.

5. The failure salience of the automation error is expected to influence the likelihood of relying on the aid in the future trials. Such that as the salience increases the lower temporal reliance becomes (temporal reliance is measured by the agreement with an aid on the trial following an aid error). (H_O = There is no significant effect between failure salience groups for temporal reliance).

   a. High salience failures (i.e., obvious errors) will cause a significantly less temporal reliance on the aid compared to less salient errors (moderate and low salience failures).

   b. Moderate salience failures will cause less temporal reliance compared to low salience failures but maintain higher temporal reliance than high salience failures.

   c. Low salience failures will maintain the highest level of temporal reliance compared to the more salient errors.

6. It is expected that source characteristics of the agents and the salience of the agent errors will interact to affect temporal reliance. (H_O = There is no significant interaction between source characteristics and failure salience).

a. Agents perceived to be human will experience drops in temporal reliance proportional to the increasing simplicity of the error made. Also it is expected that participants will be more forgiving of human errors compared to robotic errors, especially on more difficult stimuli.

b. The computer agents will experience equivalent drops in reliance across all types of errors. This reflects automation bias, in which automation is expected to work perfectly or not at all. Participants will be unforgiving of all robotic errors regardless of error salience.

Table 15. Hypotheses for Experiment 4.

	Hypothesis Number					
Dependent Measure	1	2	3	4	5	6
Perceived Trust	$M_{HR} \neq U_{HR}$ $M_{LR} \neq U_{LR}$	$U_{HR} > M > U_{LR}$	$H > [D, S]$	(ES: $HM_{HR} \neq HU_{HR}$) < (ES: $DM_{HR} \neq DU_{HR}$) < (ES: $SM_{HR} \neq SU_{HR}$) (ES: $HM_{LR} \neq HU_{LR}$) < (ES: $DM_{LR} \neq DU_{LR}$) < (ES: $SM_{LR} \neq SU_{LR}$)		
Reliance	$M_{HR} \neq U_{HR}$ $M_{LR} \neq U_{LR}$	$U_{HR} > M > U_{LR}$	$H > [D, S]$	(ES: $HM_{HR} \neq HU_{HR}$) < (ES: $DM_{HR} \neq DU_{HR}$) < (ES: $SM_{HR} \neq SU_{HR}$) (ES: $HM_{LR} \neq HU_{LR}$) < (ES: $DM_{LR} \neq DU_{LR}$) < (ES: $SM_{LR} \neq SU_{LR}$)		
Workload		$U_{HR} < M < U_{LR}$	$H < [D, S]$			
Temporal Reliance					$F_H <$ $F_M <$ F_L	$HF_H < HF_M < HF_L$ $[SF_L , DF_L] < HF_L$ $[SF_H , DF_H] <$ $[SF_M, DF_M] <$ $[SF_L, SF_L]$

*H = Human Agents, D = Different-Type Agents, S = Same Type Agents, U_{HR} = Uniform High-Reliability, U_{LR} = Uniform Low-Reliability, M = Mixed-Reliability, M_{HR} = Mixed High-Reliability, M_{LR} = Mixed Low-Reliability, F_H = High Failure Salience, F_M = Moderate Failure Salience, F_L = Low Failure Salience, ES = Effect Size

EXPERIMENT 4: RESULTS

The purpose of the fourth experiment was to examine the effect of agent type, reliability condition, and agent error salience upon subjective trust ratings, perceived workload, and behavioral measures (i.e., reliance; see Table 16). The following analyses focus first on main effects and then interactions. The final section of results examines findings regarding individual differences and how these may have influenced the results. Overall result means and standard deviations for each condition are given in Appendix V.

Table 16. Results for hypotheses for Experiment 4.

	Hypothesis Number					
Dependent Measure	1	2	3	4	5	6
Perceived Trust	$M_{HR} \neq U_{HR}$ **$\underline{M_{LR} \neq U_{LR}}$**	**$\underline{U_{HR} > M > U_{LR}}$**	H > [**$\underline{D, S}$**]	(ES: $HM_R \neq HU_R$) < (ES: $DM_R \neq DU_R$) < (ES: $SM_R \neq SU_R$)		
Reliance	**$\underline{M_{HR} \neq U_{HR}}$** $M_{LR} \neq U_{LR}$	**$\underline{U_{HR} > M > U_{LR}}$**	H > [D, S]	**(ES: $HM_R \neq HU_R$) <** **(ES: $DM_R \neq DU_R$) <** **(ES: $SM_R \neq SU_R$)**		
Workload		$U_{HR} < M < U_{LR}$	H < [D, S]			
Temporal Reliance					**$\underline{F_H \leq F_M \leq F_L}$**	**$\underline{HF_H < HF_M \leq HF_L}$** $[SF_L , DF_L] < HF_L$ $[SF_H , DF_H] = [SF_M, DF_M] = [SF_L, SF_L]$

*H = Human Agents, D = Different-Type Agents, S = Same Type Agents, U_{HR} = Uniform High-Reliability, U_{LR} = Uniform Low-Reliability, M = Mixed-Reliability, M_{HR} = Mixed High-Reliability, M_{LR} = Mixed Low-Reliability, F_H = High Failure Salience, F_M = Moderate Failure Salience, F_L = Low Failure Salience, ES = Effect Size

Hypotheses in bold and underlined were supported by the results. Hypotheses with plain text were not supported.

Subjective Data

Self-Rated Confidence

Contrary to anticipated results operator perceived self-confidence in performing the search-and-rescue task themselves was not related to actual reliance on the automated aids (r = 0.08, p = .20). Additionally, when self-confidence was subtracted from automation trust, agent correlations between trust-self-confidence and reliance were lowered or removed altogether (compared to direct automation trust and automation reliance correlations). Therefore, it was believed that self-confidence as measured in this study added more error than power to the analysis, and was therefore excluded from the rest of the analyses.

Self-Rated Trust

Self-Rated Trust Main Effect of Agent

Results were analyzed using a 3 (agent type) * 3 (reliability condition) univariate ANOVA on self-rated trust. The main effect for agent was not significant, $F(2, 287) = 0.41$, p = .66, $\eta^2 = .00$ (see Table 17).

Table 17. Self-reported trust of agents across agent-type.

Agent Type	Mean	SD
Human	6.27	1.56
Different-Type Robotic	6.43	1.56
Same-Type Robotic	6.21	1.75

Self-Rated Trust Main Effect of Reliability

Results were analyzed using a 3 (agent type) * 3 (reliability condition) univariate

ANOVA on self-rated trust. The main effect for reliability condition was significant, $F(2, 287) = 23.73$, $p < .0005$, $\eta^2 = .14$. Pairwise comparison indicated that the three reliability conditions were significantly different in the predicted direction (see Figure 27).

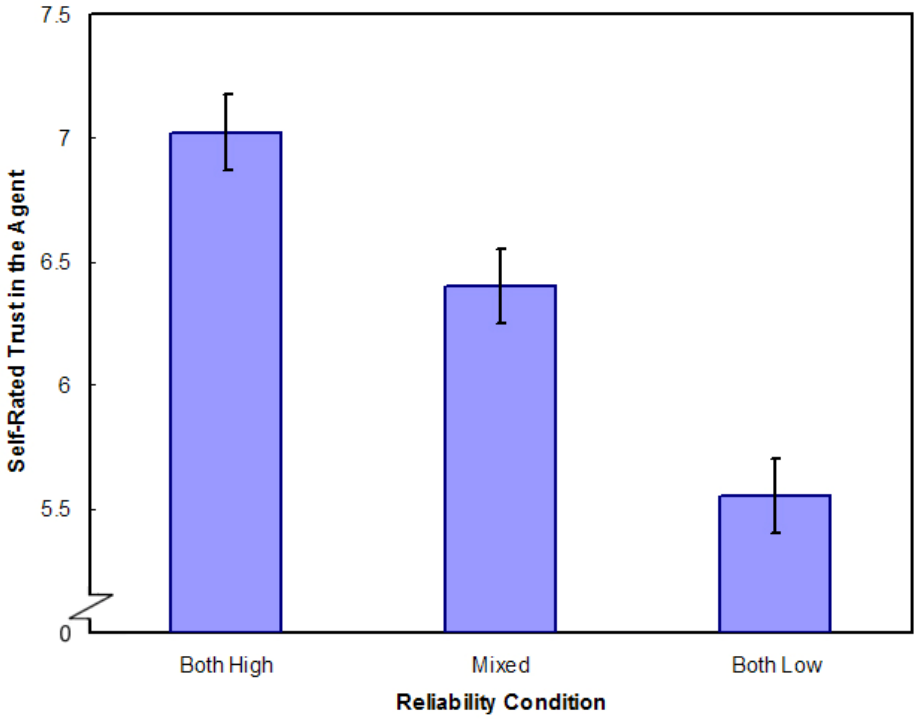


Figure 27. Agent trust as a function of reliability condition. Error bars represent standard error.

Self-Rated Trust Bias by Reliability Condition

It was hypothesized that there would be bias in the mixed reliability aids compared to the uniform reliability aids. That is, that a low-reliability aid would negatively affect the trust in a concurrent high-reliability aid, and that a high-reliability aid would positively affect the trust in a concurrent low-reliability aid. To measure this I first conducted paired-samples t-test to ensure that the trust ratings in the low and high-reliability mixed condition were significantly different, $t(97) = 3.72$, $p < .0005$, $g = 0.45$. Results were in the predicted direction with the low-reliability

aid (M = 6.02, SD = 2.07) being rated as significantly less trustworthy than the high-reliability aid (M = 6.84, SD = 1.60). I then conducted a paired-samples t-test for the two agents used in the uniform-high and uniform-low reliability conditions to ensure that they were sufficiently similar to take an average high and average low score. The uniform low ($t(97) = 0.47$, $p = .64$, $g = 0.05$) and high-reliability ($t(96) = 0.21$, $p = .83$, $g = 0.02$) aids were not significantly different in terms of self-reported trust. A one-tailed independent-samples t-test was conducted between the low reliability trust scores in the mixed-reliability condition and the averaged low reliability trust scores in the low-uniform-reliability condition. A one-tailed independent-samples t-test was also conducted between the high reliability trust scores in the mixed-reliability condition and the averaged high reliability trust scores in the high-uniform-reliability condition. A measure of the magnitude of the effect for each of the t-tests was obtained by calculating Hedges g from the means and standard deviations of each group. This gave me a non-significant result for the high reliability condition, $t(194) = 1.06$, $p = .15$, $g = 0.15$, though the means were in the right direction (High uniform: M = 7.07, SD = 1.43; High-mixed: M = 6.84, SD = 1.60). On the other hand, there was a significant result in the predicted direction for the low reliability condition ($t(196) = 1.71$, $p = .04$, $g = 0.24$; Low uniform: M = 5.59, SD = 1.56; Low-mixed: M = 6.03, SD = 2.06).

Self-Rated Trust Interaction between Agent Type and Reliability

These results were analyzed for the effect-size difference for each agent for their trust in the mixed reliability vs. trust in the uniform reliability. The same process from the previous section was used to calculate ES for each bias measure.

Human Agent and Trust

Limiting analysis to those participants in the human-agent condition only, a paired-samples *t*-test was conducted to examine trust in the mixed reliability condition. It was evident that the low reliability aid (M = 6.30, SD = 1.57) and high reliability aid (M = 7.21, SD = 1.29) had significantly different perceived rated trust, ($t(32) = 2.39$, $p = .02$, $g = 0.64$; See Figure 28 blue line). Paired-samples *t*-test were then used to examine trust in the two aids used in the uniform low-reliability condition (Agent A: M = 5.45, SD = 1.62; Agent B: M = 5.36, SD = 1.78; $t(32) = 0.28$, $p = .78$, $g = 0.05$; See Figure 29 blue line) and uniform high-reliability condition (Agent A: M = 6.70, SD = 2.02; Agent B: M = 6.61, SD = 1.98; $t(32) = 0.32$, $p = .74$, $g = 0.04$; Figure 30 blue line), both of which did not significantly differ. Since the perceived trust did not significantly differ, in terms of the *t*-test or ES values, for the Agents in either of the uniform reliability conditions, these values were combined to allow for comparison against the mixed-reliability condition (See Figure 31; uniform values are represented by hollow diamonds). Using an independent-samples one-tailed *t*-test the perceived trust for the low-reliability human agent in the mixed-reliability condition (M = 6.30, SD = 1.57) was compared against the averaged low-reliability human agent trust in the low-uniform condition (M = 5.40, SD = 1.42; $t(64) = 2.42$, $p = .009$, $g = 0.59$). Results indicated that the low-reliability agent in the human mixed-reliability condition was rated as significantly higher in terms of trust than the low-reliability human agents in the uniform low-reliability condition. Next the biasing effect on a high-reliability human agent was examined. Using an independent-samples two-tailed t-test, two-tailed was used because the means did not match the direction of the hypothesis, high-reliability human agent trust in the mixed-reliability condition (M = 7.21, SD = 1.29) was compared against the averaged high-

reliability human agent trust in the uniform high-reliability condition (M = 6.65, SD = 1.83; t(64) = 1.44, p = .16, g = 0.35). Results were not significant for the biasing effects in the human agent condition for perceived trust.

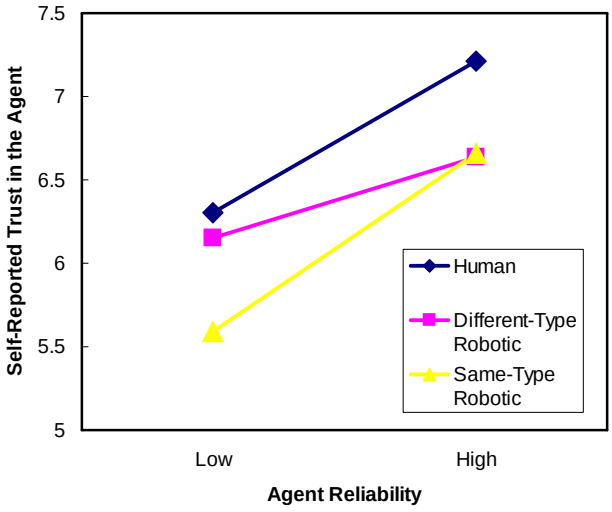


Figure 28. Perceived trust as a function of agent reliability by agent type in the mixed-reliability condition.

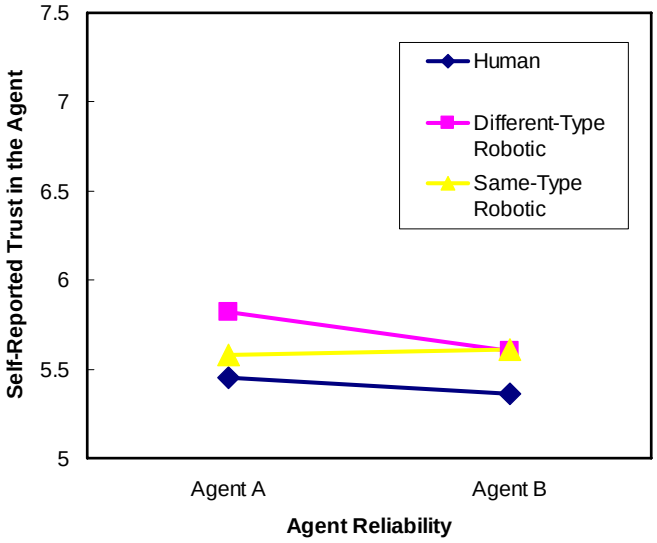


Figure 29. Perceived trust as a function of agent type in the low-reliability condition.

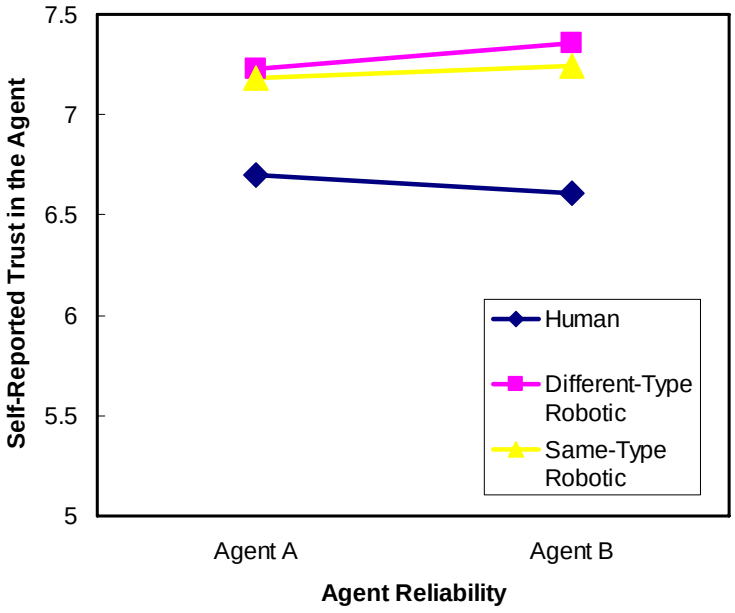


Figure 30. Perceived trust as a function of agent type in the high-reliability condition.

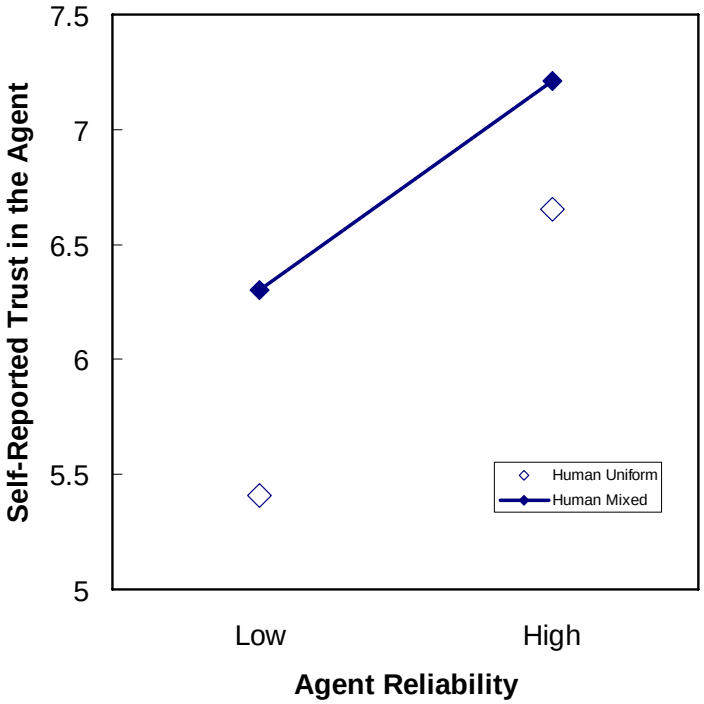


Figure 31. Perceived trust as a function of agent reliability for human agents. Note that mixed-reliability are the solid diamonds and uniform-reliabilities are represented by the hollow diamonds.

Different-Type Robotic Agent and Trust

The next sets of analyses were limited to those participants in the different-type robotic agent condition. To conduct this analysis a paired-samples one-tailed *t*-test was conducted to examine trust in the mixed reliability condition. Surprisingly the low-reliability different-type robotic agent (M = 6.15, SD = 2.25) did not significantly differ in terms of perceived trust from the high-reliability different-type robotic agent (M = 6.64, SD = 1.60; $t(32) = 1.33$, $p = .10$, g = 0.27; See Figure 28). However, additional analyses were still conducted to see if the degree of biasing in the agent scores was lower or higher in this agent compared to the other agents. Paired-samples t-test were used to examine trust in the two aids used in the low-uniform different-type robotic condition (Agent A: M = 5.82, SD = 1.81; Agent B: M = 5.61, SD = 2.00; $t(32) = 0.56$, $p = .58$, $g = 0.11$; See Figure 29) and high-uniform conditions (Agent A: M = 7.23, SD = 1.50; Agent B: M = 7.35, SD = 0.88; $t(30) = 0.50$, $p = .65$, $g = 0.10$; See Figure 30), both of which did not significantly differ. Since the perceived trust did not significantly differ, in terms of the *t*-test or ES values, for the Agents in either of the uniform-reliability different-type robot conditions, these values were combined to allow for comparison against the mixed-reliability condition. Using an independent-samples one-tailed *t*-test the perceived trust for the low-reliability different-type robotic agent in the mixed reliability condition (M = 6.15, SD = 2.25) was compared against the averaged low-reliability different-type robotic agent trust in the low-uniform condition (M = 5.71, SD = 1.56; $t(64) = 0.92$, $p = .18$, $g = 0.23$; See Figure 32). Results indicated that the low-reliability agent in the different-type robotic mixed condition was not significantly different in terms of trust than the uniform low-reliability agents in the different-type robotic condition. Next, the biasing effect on a high-reliability different-type robotic agent

was examined. Using an independent-samples one-tailed t-test high-reliability different-type robotic agent trust in the mixed-reliability condition (M = 6.64, SD = 1.60) was compared against the averaged high-reliability different-type robotic agent trust in the high-uniform condition (M = 7.34, SD = 0.98; $t(63) = 2.15$, $p = .036$, $g = 0.53$; See Figure 32).

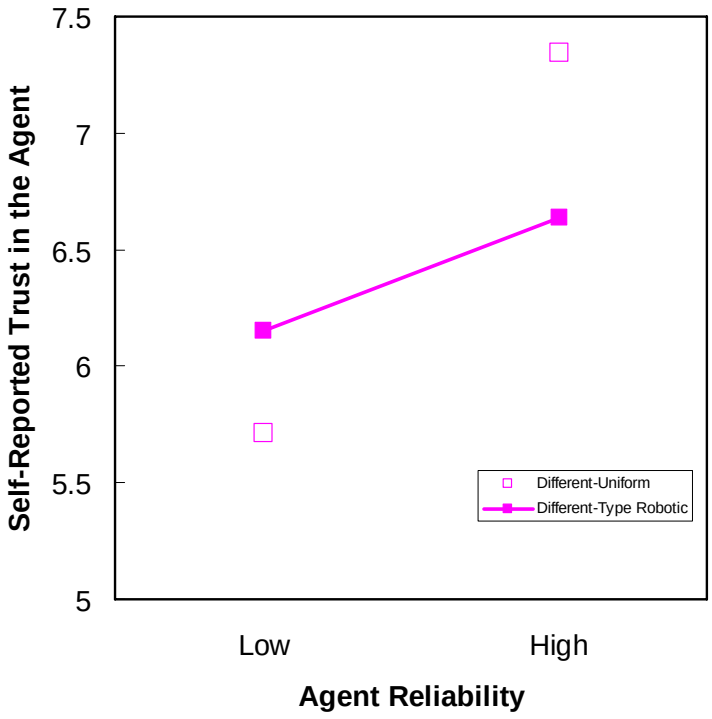


Figure 32. Perceived trust as a function of agent reliability for different-type robotic agents. Note that mixed-reliability are the solid squares and uniform-reliabilities are the hollow squares.

Same-Type Robotic Agent and Trust

The final set of analyses were limiting to those participants in the same-type robotic agent condition only. To conduct this analysis I used a paired-samples t-test to examine trust in the mixed-reliability condition, it was apparent that the low reliability aid (M = 5.63, SD = 2.33) and high reliability aid (M = 6.63, SD = 1.81) had significantly different perceived rated trust, ($t(31) = 2.46$, $p = .01$; See Figure 28). Paired-samples t-test were then used to examine trust in the two

aids used in the low-uniform (Agent A: M = 5.58, SD = 1.82; Agent B: M = 5.61, SD = 2.05; $t(31) = 0.10$, $p = .78$, $g = 0.02$; See Figure 29) and high-uniform conditions (Agent A: M = 7.18, SD = 1.40; Agent B: M = 7.24, SD = 1.39; $t(32) = 0.30$, $p = .77$, $g = 0.04$; See Figure 30), both of which did not significantly differ. Since the perceived trust did not significantly differ, in terms of the t-test or ES values, for the same-type robotic agents in either of the uniform reliability conditions, these values were combined to allow for comparison against the mixed-reliability condition. Using an independent-samples one-tailed t-test the perceived trust for the low-reliability same-type robotic agent in the mixed reliability condition (M = 5.67, SD = 2.31) was compared against the averaged low-reliability same-type robotic agent trust in the low-uniform condition (M = 5.63, SD = 1.73; $t(64) = 0.95$, $p = .48$, $g = 0.01$; See Figure 33). Next the biasing effect on a high-reliability same-type robotic agent was examined. Using an independent-samples one-tailed t-test high-reliability same-type robotic agent trust in the mixed-reliability condition (M = 6.63, SD = 1.81) was compared against the averaged high-reliability same-type robotic agent trust in the high-uniform condition (M = 7.21, SD = 1.27; $t(63) = 1.52$, $p = .07$, g = 0.37; See Figure 33), again results were not significant.

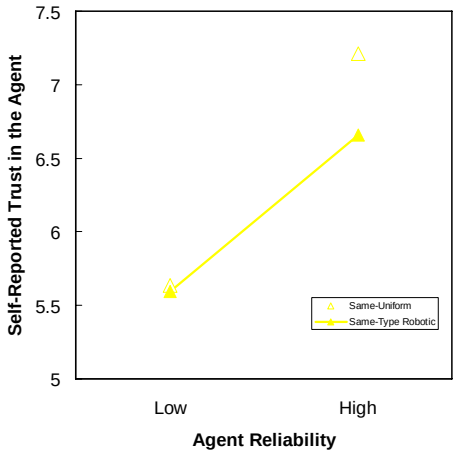


Figure 33. Perceived trust as a function of agent reliability for same-type robotic agents. Note that mixed-reliability are the solid triangles and uniform-reliabilities are the hollow triangles.

Effect-Size Analysis of Agent and Trust

The effect-sizes of the difference between the mixed and uniform agents of the same reliability are presented in Table 18. In absolute average terms, human agents demonstrated the largest average effect-size between the mixed and uniform conditions, meaning that they demonstrated the greatest biasing effect when presented in a mixed condition. The same-type robotic agents experienced the least biasing effect between the mixed and uniform conditions, meaning that these agents were the most insensitive to whether they were presented uniformly or in a mixed condition. Finally, different-type robotic agents experienced an intermediate level of effect-size biasing between the mixed and uniform conditions.

Table 18. Effect-size measures for degree of difference between mixed and uniform reliability conditions for trust. Note that negative values indicate that the mixed value is lower than the uniform value, while positive values indicate that the mixed value is higher than the uniform value.

Agent Type	Low Reliability ES	High Reliability ES	Absolute Average ES
Human	+0.59	+0.35	0.47
Different-Type Robotic Aid	+0.23	-0.53	0.38
Same-Type Robotic Aid	-0.01	-0.37	0.19

Workload

It was hypothesized that automation trust would be positively correlated to reliance on the agent and negatively correlated to workload. That is, with higher levels of trust in an agent reliance on the agent should increase and perceived workload should decrease. On the other

hand, it was also believed that with lower levels of trust in an agent, reliance on the agent would decrease, and perceived workload of the participant would increase. There was partial support for this hypothesis. In regards to reliance, there was a significant positive correlation to self-reported trust in the agent ($r = .37$; $p < .0005$). However, in regards to workload, there was not a significant relationship to perceived trust in the agents ($r = .03$, $p = .60$, $n = 294$) or participant reliance ($r = .004$, $p = .94$, $n = 294$). Thus, while users may rely more heavily on an agent's decisions with increased trust, this increased reliance is *not* associated with a decrease in workload.

Additionally, it was hypothesized that there would be a main effect for workload by agent type. This was examined using a 3 (agent type) by 3 (reliability condition) univariate ANOVA. Results indicated that there was not a main effect for agent type, $F(2, 285) = 0.63$, $p = .53$, $\eta^2 = 0.004$. All other effects were also not significant. Means and standard deviations for the NASA-TLX and its subscales are presented in Table 19. Note that two participants did not complete the NASA-TLX.

Table 19. NASA-TLX means and standard deviations for search-and-rescue task.

NASA-TLX Measure	Mean	Standard Deviation
Overall Workload	70.52	13.34
Mental Demand	80.44	16.15
Physical Demand	21.24	19.65
Temporal Demand	74.05	22.06
Performance	55.12	22.24
Effort	72.93	18.69
Frustration	58.45	25.37

Behavioral Measures

Automation Reliance and Trust

It was hypothesized that automation reliance would be significantly correlated to automation trust. A Pearson correlation was conducted and there was a significant correlation between automation trust and automation reliance ($r = .37$, $p < .0005$). That is, as self-rated agent trust increased so did user reliance as measured by agreement with the agent.

Automation Reliance Main Effect of Reliability Level

Reliance was first analyzed using a 3 (agent type) by 3 (reliability condition) univariate ANOVA. There was a main effect for reliability level, $F(2, 287) = 48.51$, $p < .0005$, $\eta^2 = 0.25$. Results were in the predicted direction with higher reliability levels having higher levels of reliance (see Figure 34). All other effects were not significant ($p > .05$ in all cases).

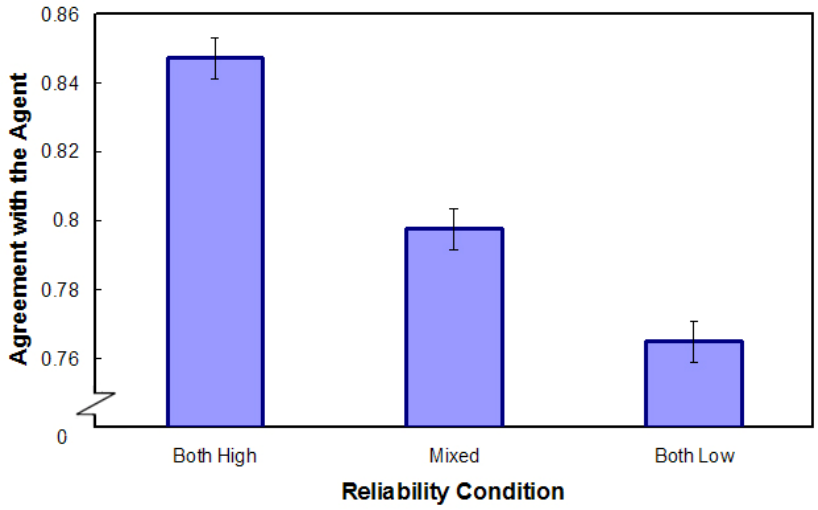


Figure 34. Reliance as a function of reliability condition. Note that error bars represent standard error.

Reliance Main Effect of Agent

In examining participant reliance with the univariate ANOVA, it was evident that the main effect of agent type was not significant, $F(2, 287) = 2.28$, $p = .10$, $\eta^2 = 0.02$. That is, the reliance scores across all three aids were approximately 80% (see Table 20).

Table 20. Reliance on agents across agent-type.

Agent Type	Mean	SD
Human	.79	.06
Different-Type Robotic	.81	.07
Same-Type Robotic	.81	.07

Reliance Bias by Reliability Condition

It was hypothesized that there would be reliance bias in the mixed reliability aids compared to the uniform reliability aids. That is, that a low-reliability aid would negatively affect the reliance in a concurrent high-reliability aid, and that a high-reliability aid would positively affect the reliance in a concurrent low-reliability aid. A paired-samples t-test indicated that the reliance between the low and high-reliability agents in the mixed condition was significantly different, $t(98) = 11.71$, $p < .0005$, $g = 1.04$. Results were in the predicted direction with the low-reliability aid ($M = 0.77$, $SD = 0.06$) being relied on significantly less than the high-reliability aid ($M = 0.83$, $SD = 0.06$). Conducting a paired-samples *t*-test for the two agents used in the uniform-high and uniform-low conditions indicated that they were sufficiently similar to take an averaged high and an averaged low reliability aid score. The uniform low-reliability aids were not significantly different ($t(98) = 0.62$, $p = .54$, $g = 0.05$). The uniform high-reliability aids were also not significantly different in terms of operator reliance, ($t(97) = 1.38$, $p = .17$, $g = 0.10$). A one-tailed independent-samples *t*-test was then conducted between low-mixed and average-low-

uniform as well as high-mixed and average-high-uniform reliance. The last step was to calculate Hedges *g* from the means and standard deviations of each group. This gave a non-significant result for low-reliability ($t(196) = 0.06$, $p = .96$, $g = 0.01$). There was a significant result in the predicted direction for high-reliability condition for reliance, $t(195) = 1.99$, $p < .05$, $g = 0.28$ (see Figure 35). Such that the uniform high-reliability condition ($M = 0.85$, $SD = 0.06$) was relied on significantly more often than the high-reliability in the mixed condition ($M = 0.83$, $SD = 0.06$).

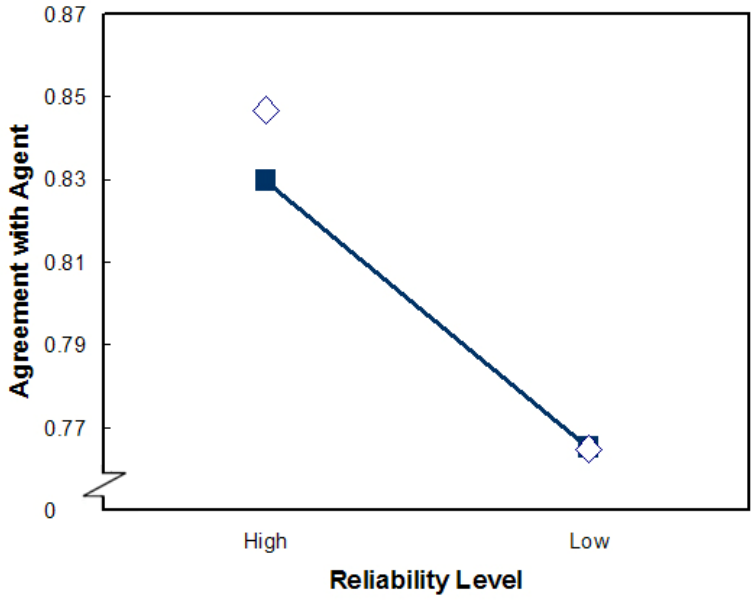


Figure 35. Reliance as a function of reliability condition. Note that the solid squares represent the mixed-reliability condition and the hollow diamonds represent the uniform conditions.

Agent Reliance: Interaction between Agent Type and Reliability

The next step was to examine the hypothesis on whether the type of agent impacts how a concurrent conflicting reliability agent can bias reliance. These results were analyzed for effect-size difference for each agent for reliance in the mixed-reliability vs. reliance in the uniform-reliability condition.

Human Agent and Reliance

Limiting analysis to those participants in the human-agent condition only, I used a paired-samples *t*-test to examine reliance in the mixed reliability condition It was apparent that the low reliability aid (M = 0.75, SD = 0.06) and high reliability aid (M = 0.82, SD = 0.06) had significantly different operator reliance, ($t(32) = 7.42$, $p < .0005$). Paired-samples t-test were then used to examine reliance in the two aids used in the low-uniform (Agent A: M = 0.75, SD = 0.05; Agent B: M = 0.76, SD = 0.07; $t(32) = 0.23$, $p = .82$, $g = 0.04$) and high-uniform conditions (Agent A: M = 0.85, SD = 0.06; Agent B: M = 0.83, SD = 0.06; $t(3237) = 2.$, $p = .02$, $g = 0.32$). While the agents in the low-uniform condition did not significantly differ, the high-uniform agents did significantly differ in terms of reliance. Thus, overall effect-size was calculated separately for the high-reliability agents. Using an independent-samples two-tailed t-test the reliance towards the low-reliability agent in the mixed-reliability condition (M = 0.75, SD = 0.05) was compared against the averaged low-reliability agent reliance in the low-uniform condition (M = 0.75, SD = 0.06; $t(64) = 0.28$, $p = .78$, g = 0.07). This indicates that the concurrent presence of a high-reliability human agent did *not* lead participants to rely any more on a low reliability human agent. Next the biasing effect of a high-reliability human agent was examined. Since the uniform-high-reliability agents differed significantly two analyses were conducted. In both cases the mixed-reliability condition was lower, but in only one case significantly. Using an independent-samples one-tailed t-test high-reliability agent reliance in the mixed-reliability condition (M = 0.82, SD = 0.06) was compared against the high-reliability agents reliance values of either the left aid M = 0.85 (SD = 0.06; $t(64) = 1.88$, $p = .03$, $g = 0.46$) or the right aid M = 0.83 (SD = 0.06; $t(64) = 0.68$, $p = .50$, $g = 0.17$). An average effect-size

difference between mixed-high reliability and uniform-high-reliability human agents is 0.32. Thus, while it appears there may be a trend for positive biasing of subjective trust ratings, as presented earlier, reliance was generally unsusceptible to the manipulation of agent type (see Figure 36).

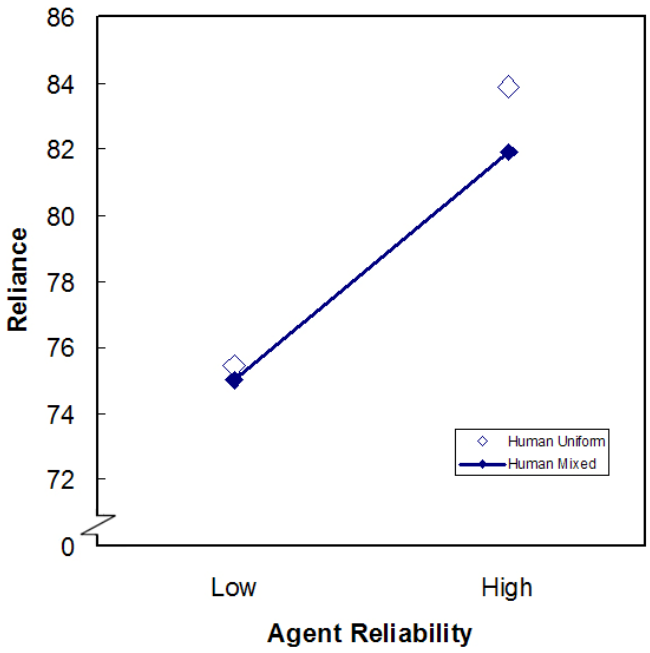


Figure 36. Reliance as a function of agent reliability for human agents. Note that mixed-reliability are the solid diamonds and uniform-reliabilities are the hollow diamonds.

Different-Type Robotic Agent and Reliance

Next I turn to the different-type robotic agents. Conducting a paired-samples *t*-test on different-type robotic agent reliance between the low and high-reliability agents, it was apparent that reliance on the aids did significantly differ, $t(32) = 7.05$, $p < .0005$, $g = 0.99$. Paired-samples t-test were then used to examine reliance in the two aids used in the low-uniform (Agent A: $M = 0.76$, $SD = 0.07$; Agent B: $M = 0.77$, $SD = 0.08$; $t(32) = 0.77$, $p = .45$, $g = 0.11$) and high-uniform

conditions (Agent A: $M = 0.85$, $SD = 0.07$; Agent B: $M = 0.85$, $SD = 0.07$; $t(31) = 0.06$, $p = .95$, $g = 0.01$), both of which did not significantly differ. Using an independent-samples one-tailed t-test the reliance towards the low-reliability agent in the mixed-reliability condition ($M = 0.78$, $SD = 0.07$) was compared against the averaged low-reliability agent reliance in the low-uniform condition ($M = 0.76$, $SD = 0.07$; $t(64) = 1.07$, $p = .14$, $g = 0.26$). Next the biasing effect of a high-reliability human agent was examined. Using an independent-samples one-tailed t-test high-reliability agent reliance in the mixed-reliability condition ($M = 0.84$, $SD = 0.06$) was compared against the averaged high-reliability agent trust in the high-uniform condition ($M = 0.85$, $SD = 0.06$; $t(63) = 0.83$, $p = .21$, $g = 0.21$; see Figure 37).

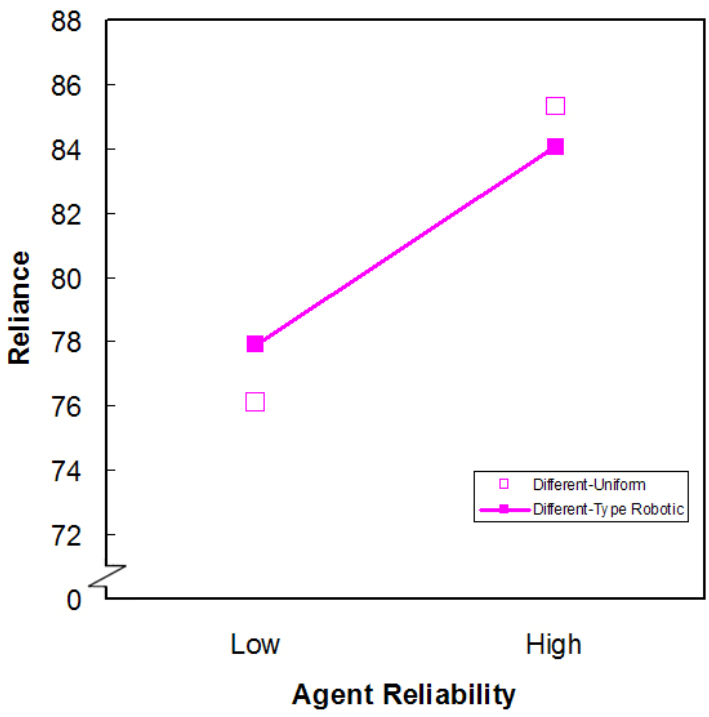


Figure 37. Reliance as a function of agent reliability for different-type robotic agents. Note that mixed-reliability are the solid squares and uniform-reliabilities are the hollow squares.

Same-Type Robotic Agent and Reliance

Finally, the reliance bias for same-type robotic agents was examined. Conducting a paired-samples *t*-test on same-type robotic agents reliance on the low-reliability and high-reliability aid in the mixed reliability condition, it was found that they were significantly different, $t(32) = 5.86$, $p < .0005$, $g = 1.02$. Next paired samples *t*-tests were conducted on uniform-low ($M = 0.78$, $SD = 0.06$ vs. $M = 0.78$, $SD = 0.07$; $t(32) = 0.05$, $p = .96$, $g = 0.01$) and uniform-high reliability conditions ($M = 0.85$, $SD = 0.07$ vs. $M = 0.85$, $SD = 0.08$; $t(32) = 0.13$, $p = .90$, $g = 0.07$). As these scores were not significantly different in terms of effect size or standard significance reliance scores within each uniform condition were combined for the next step of analysis. I then compared the reliance in the mixed-low-reliability aid ($M = 0.77$, $SD = 0.06$) to the averaged-uniform reliability aids ($M = 0.78$, $SD = 0.06$) using a two-tailed independent samples *t*-test, $t(64) = 0.83$, $p = .41$, $g = 0.20$). Results were also not significant for the mixed-high-reliability aid ($M = 0.83$, $SD = 0.06$) compared to the averaged-uniform reliability aids ($M = 0.85$, $SD = 0.07$) using a one-tailed independent samples *t*-test, $t(64) = 1.26$, $p = .11$, $g = 0.31$; see Figure 38).

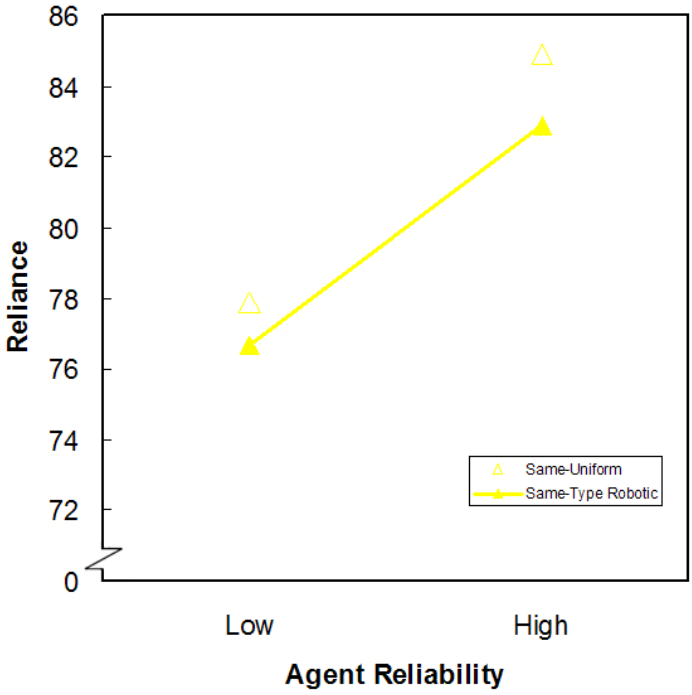


Figure 38. Reliance as a function of agent reliability for same-type robotic agents. Note that mixed-reliability are the solid triangles and uniform-reliabilities are the hollow triangles.

Effect-Size Analysis of Agent and Reliance

The effect-sizes of the differences by agent-type are presented in Table 21. In general all agents tended to have similar effect-size differences between the mixed and uniform conditions, but the pattern of results supported the research hypotheses. That is, human agents had the smallest effect-size differences for reliance; which means, that they had the least amount of difference in terms of agreement with a human agent when it appeared with another person of similar reliability or different reliability. In terms of agreement with robotic agents ES were slightly higher. Same-type robotic aids had the largest average effect-size difference (i.e., the most carryover bias), while different-type robotic aids feel in between the reliance bias of human and same-type agents. However, because the effect-sizes are so close the results provide only

limited support for our experimental hypothesis (i.e., that agent type impacts crossover bias between two agents, such that human agents are the most independent: smallest ES difference between mixed and uniform conditions, different-type robotic aids: moderate ES difference between mixed and uniform conditions, and same-type robotic aids: largest ES difference between mixed and uniform conditions). One point of possible contention of these results is that the single highest biasing component was the high-reliability human agents. That is, a concurrent low-reliability human dropped agreement with a concurrent high-reliability human agent by a third of a standard deviation, which was the single largest impact on reliance observed in this analysis! This result was especially surprising given the beneficial effects that mixed-reliability had on self-rated perceived trust in the human agents (both low and high).

Table 21. Effect-size measures for degree of difference between mixed and uniform conditions for reliance. Note that negative values indicate that the mixed value is lower than the uniform value, while positive values indicate that the mixed value is higher than the uniform value.

Agent Type	Low Reliability ES	High Reliability ES	Absolute Average ES
Human	-0.07	-0.32	0.20
Different-Type Robotic Aid	+0.26	-0.21	0.24
Same-Type Robotic Aid	-0.20	-0.31	0.26

Failure Salience on Reliance

The next analysis examines how failure salience (i.e., the obviouness of the agent's errors) influences the likelihood of relying on the aid in future trials. The hypotheses predicted that the more salient an error by the agent was the lower temporal reliance would be. Temporal

reliance is defined here as agreeing with the agent on the next correct trial. Unfortunately, due to programming errors experimental automation errors were not applied to high-difficulty video clips in the high-reliability condition. Therefore, analysis was limited to low-reliability conditions. As error salience is a within-subjects manipulation a repeated measures ANOVA was conducted on the three types of error, with temporal reliance on the agent during the following trial as the dependent measure. There was a significant effect for error type in the predicted direction, $F(2, 392) = 210.18$, $p < .0005$, $\eta^2 = 0.52$. Pairwise comparisons indicated that reliance on the agent after an error was significantly related to the salience of the error ($p < .05$ in all cases; see Figure 39).

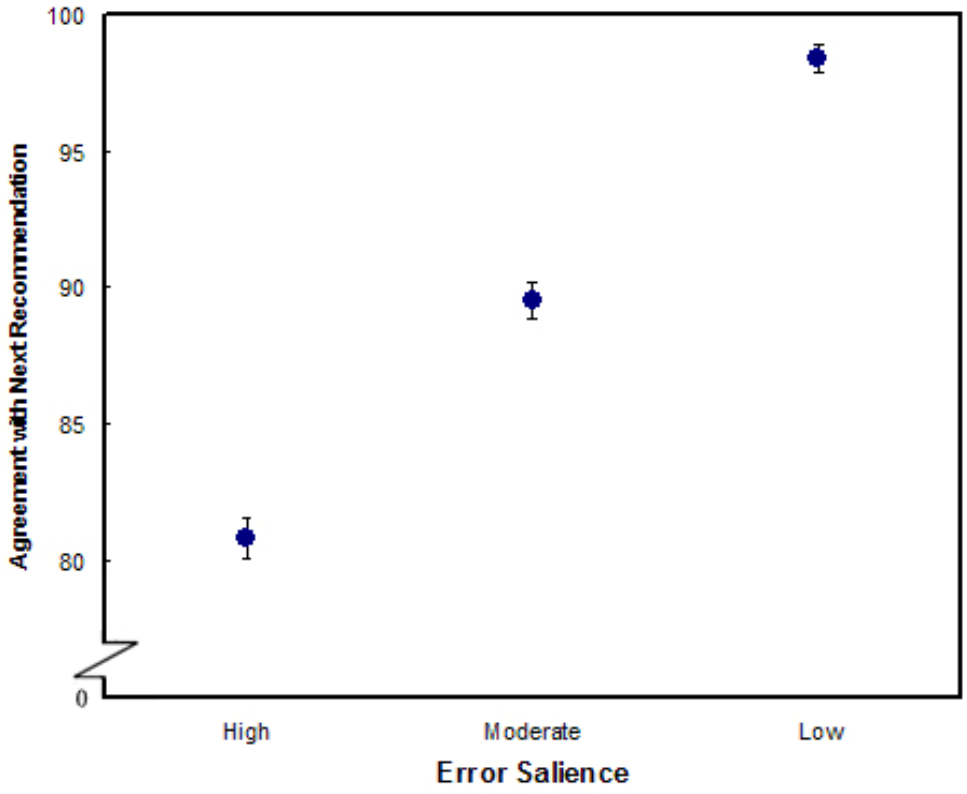


Figure 39. Temporal reliance as a function of error salience. Note that error bars represent standard error.

Failure Salience and Agent Type

Next the analysis on failure salience was conducted when Agent Type was added as a between subjects variable. There was a significant main effect for agent type on temporal reliance, $F(2, 194) = 4.82$, $p = .009$, $\eta^2 = 0.05$. However, as evidenced by the eta squared value this was a weak effect and pairwise comparisons indicated that a significant difference occurred only between the human agent and the different-type ($p = .055$) and same-type ($p = .002$) robotic agents (which did not significantly differ from each other; $p = .26$; see Figure 40). These results were contrary to the predicted direction, and actually indicated that participants were more distrusting of human agents following an error compared to robotic agents.

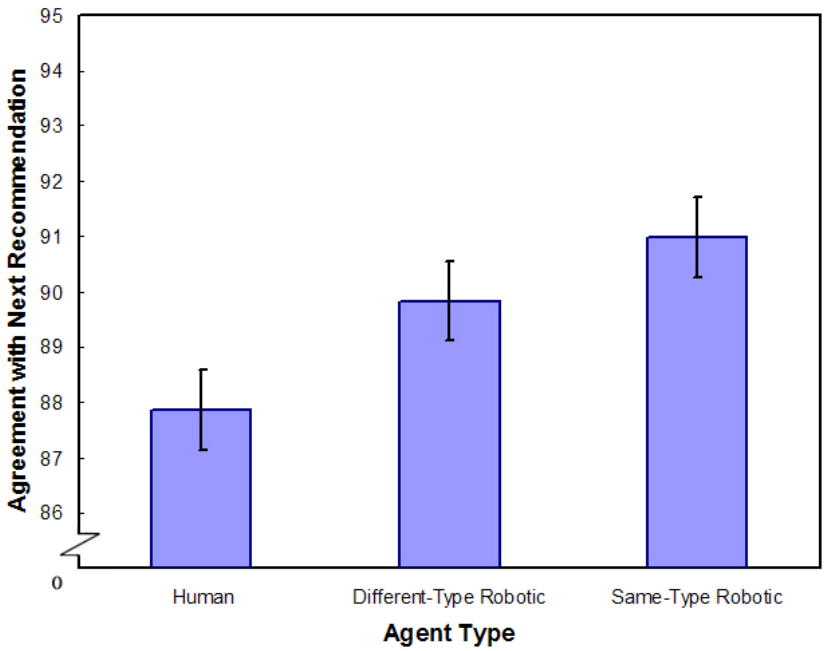


Figure 40. Temporal reliance as a function of agent type.

There was an agent by error type interaction, $F(4, 388) = 3.04$, $p = .017$, $\eta^2 = 0.03$. Visual inspection of the results (see Figure 41) indicated that temporal reliance varies more greatly for

agents when errors are more obvious (i.e., high or moderate salience). In these cases it appears that the human agents have less reliance following an error than the computer-agents. One-way ANOVAs confirmed this pattern of results. There was a significant difference within the high salience (i.e., obvious errors), $F(2, 195) = 3.94$, $p = .02$. Pairwise comparison indicated that the human agents were significantly different from the same-type robotic agents ($p = .006$) and there was a trend for them to be different from the different-type robotic agents as well ($p = .08$). The two robotic aids did not significantly differ from each other ($p = .31$). Next a one-way ANOVA was conducted on moderate salience errors. Results were significant, $F(2, 195) = 4.55$, $p = .01$. In this case pairwise comparison indicated that the human agents were significantly different from both robotic agents, which again did not significantly differ from each other. The final one-way ANOVA was conducted on the low-salience (i.e., least obvious errors), $F(2, 195) = 2.26$, $p = .11$. Examination of pairwise comparisons indicated that the same-type and different-type robotic agents significantly differed in terms of temporal reliance, such that different aids had less observer agreement following a low-salience aid error. In low-salience errors user temporal reliance did not differ from between the human and robotic agents ($p > .05$ in both cases).

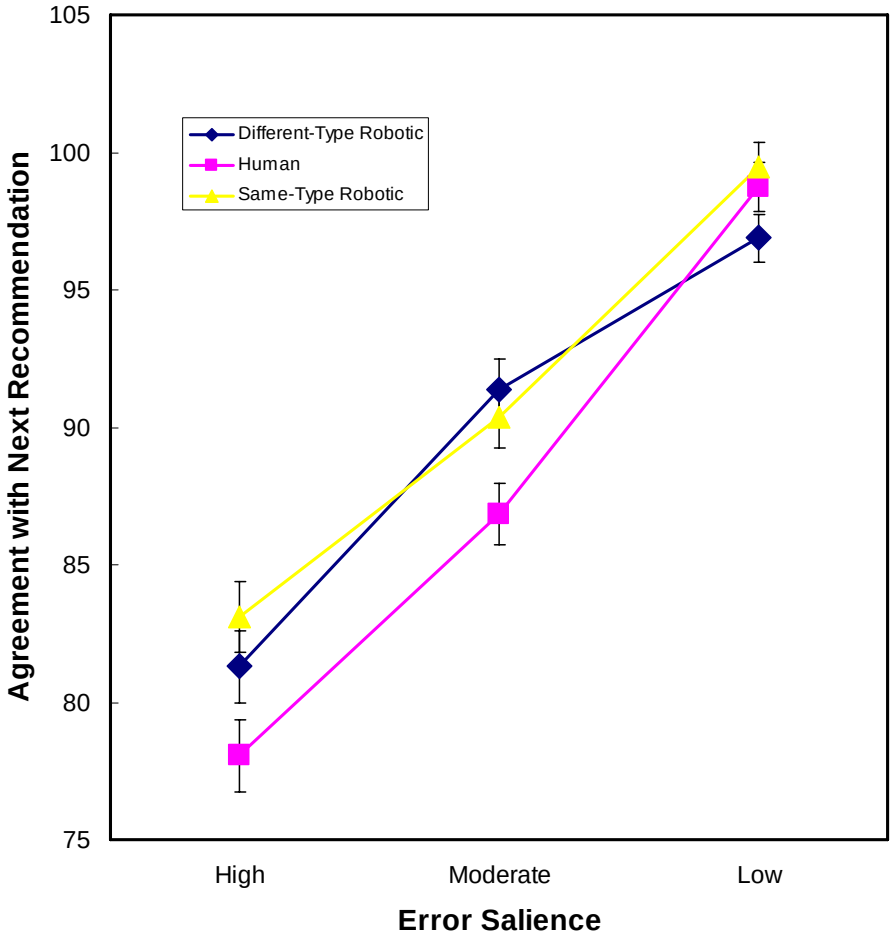


Figure 41. Temporal reliance as a function of error salience by agent type.

Miss and False Alarms

A paired samples *t*-test was conducted on temporal reliance following misses and false alarms, $t(295) = 15.41$, $p<.0005$. Contrary to prior research, in this study false alarms had significantly lower temporal reliance ($M = 80.00$, $SD = 14.35$) than misses ($M = 95.21$, $SD = 9.97$). The literature typically states that FA can be construed in several ways by participants making them more ambiguous and reducing trust and reliance levels less than misses which when noticed by participants indicates more clearly that the agent was indeed in error. However,

the literature supports that it is the clarity of the message that drives this relationship and this is more clearly typified in this research by the error salience. Thus, the current findings indicate that when error salience is controlled false alarms can be more detrimental to subsequent reliance compared to misses in this task.

Individual Differences

Participant Sex

Participants were assigned equally to control for any participant sex effects on the main factors of interest in the experiment: trust and reliance. An independent samples t-test indicated that participant sex did not influence trust ($t(294) = 0.36$, $p = .72$) or reliance ($t(294) = 0.18$, $p = .86$) in the study. This effect was also ns when broken down by agent type, reliability condition, and agent by reliability condition ($p > .05$ in all cases).

Questionnaire Data

Participants were assigned randomly to one of twelve between-subjects condition (3 reliability conditions * 3 agent types). As individual differences were a concern three trait questionnaires were administered at the start of the experiment: Anthropomorphic Tendency Scale (ATS), Interpersonal Trust Scale (ITS), and Complacency Potential Rating Scale (CPRS). The scales were given prior to study participation, and the questionnaires were designed to measure trait personality measures, thus by random assignment the groups should be approximately equal.

Anthropomorphic Tendency Scale (ATS)

There are four factors within the ATS: Extreme Anthropomorphism, Anthropomorphism of Pets, Anthropomorphism towards Gods or Deities, and Negative Anthropomorphism. A 3 (agent type) by 3 (reliability condition) univariate ANOVA was conducted on each of the four anthropomorphic factors to examine if there were any differences among the between-subjects groups. Additionally correlation analysis were conducted to examine if anthropomorphism scores correlated with the main variable of interest, these correlations were done overall, by agent, by reliability condition, and by agent*reliability condition (only significant correlations are reported in the text for a full list of correlations see Appendix W).

Extreme Anthropomorphism

Mean score for extreme anthropomorphism was 32.80 (SD = 9.78; coefficient $\alpha = 0.92$). The analysis on extreme anthropomorphism indicated that there were no significant difference between groups (in all cases $p > .05$; See Table 22). Additional analysis indicated that there were no sex difference in terms of extreme anthropomorphism ($t(327) = 0.29$, $p = .77$; See Table 23). In regards to correlations there was a nonsignificant correlation between extreme anthropomorphism to trust and reliance ($r = -.07$ and $r = -.05$ respectively). There was also no significant correlation for trust or reliance by reliability level. When broken down by agent there was a small negative correlation for self-rated trust in the different-type robotic agents ($r = -.29$). By breaking agent type down by reliability condition it was demonstrated that this effect was caused by a moderate negative correlation between extreme anthropomorphism and self-rated trust in the mixed-reliability different-type robotic agent condition ($r = -.47$). This effect was further examined by looking at how extreme anthropomorphism in the different-type mixed

reliability was significantly correlated to trust in the low-reliability aid (r = -.62, p < .0005) but not to trust in the high-reliability aids (r = -.09, p = .60). This indicates that when interacting with different-type robots in mixed reliability those high in extreme anthropomorphism had lower trust in the low reliability aid than those with lower extreme anthropomorphism scores (see Figure 42). However, this effect did not affect reliance or trust on high-reliability aids and was not apparent in the human or same-type robotic agent conditions (p > .05 in all cases).

Table 22. Extreme anthropomorphism among condition assignment. Note that SD are shown in parenthesis.

	Agent Type		
Reliability Condition	Human	Same-Type Robotic	Different-Type Robotic
Both High	30.99 (7.88)	32.67 (9.87)	33.00 (11.35)
Mixed	35.51 (11.04)	31.20 (7.59)	32.63 (9.64)
Both Low	31.66 (8.55)	32.76 (9.26)	32.63 (11.30)

Table 23. Anthropomorphism by participant sex.

ATS Factor	Participant Sex	N	Mean	SD
Extreme	Female	180	32.65	8.60
	Male	149	32.97	11.06
Pets	Female	180	39.56	6.03
	Male	149	38.10	6.35
Gods or Deities	Female	180	30.26	8.58
	Male	149	28.26	8.34
Negative	Female	180	12.18	4.29
	Male	149	11.61	4.47

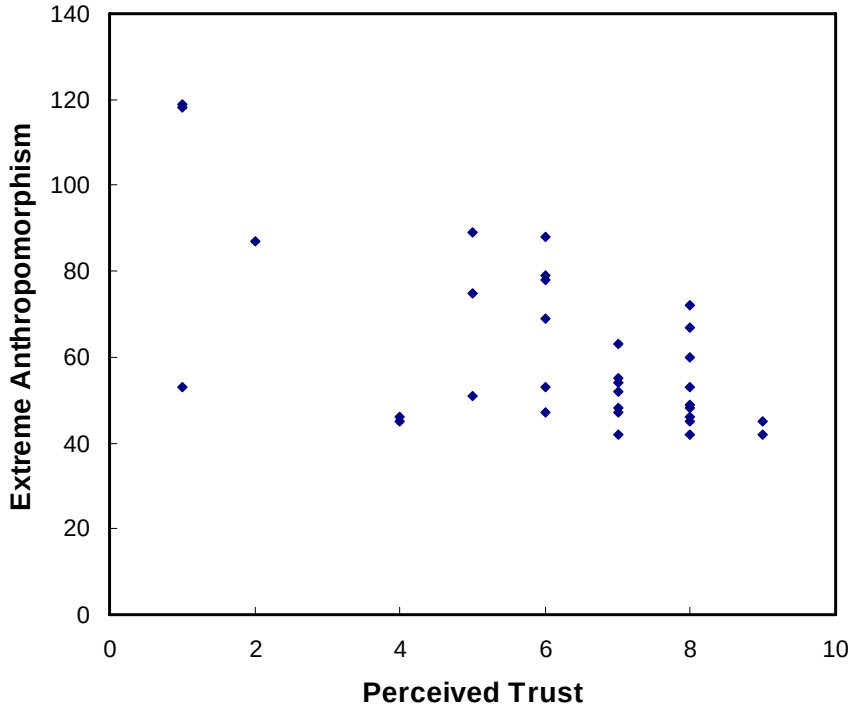


Figure 42. Extreme anthropomorphism as a function of perceived trust of the low-reliability aid. Note that results are for participants in the different-type robotic mixed condition.

Pet Anthropomorphism

Mean score for pet anthropomorphism was 38.90 (SD = 6.21; coefficient α = 0.90). The analysis on pet anthropomorphism indicated that there was a significant difference between agent conditions for levels of pet anthropomorphism, $F(2, 286) = 3.63$, $p = .028$, $\eta^2 = 0.03$. The extremely small eta squared value indicates that even though this effect was significant it was extremely small. All other effects were non-significant ($p > .05$ in all cases; See Table 24). Additional analysis indicated that there were sex difference in terms of pet anthropomorphism ($t(326) = 2.14$, $p = .03$). These results indicated that females have significantly higher pet anthropomorphism scores than males, though this was a small effect, $g = .24$ (See Table 23).

Table 24. Pet anthropomorphism among condition assignment. Note that SD are shown in parenthesis.

	Agent Type		
Reliability Condition	Human	Same-Type Robotic	Different-Type Robotic
Both High	39.33 (5.74)	41.63 (4.04)	36.77 (6.59)
Mixed	39.59 (5.77)	38.54 (5.53)	38.27 (6.35)
Both Low	39.19 (6.43)	39.01 (6.76)	37.57 (7.27)

In regards to correlations, the overall correlation of pet anthropomorphism to trust and reliance were not significant (r = .10 and r = .02 respectively). There was also no significant correlation for trust or reliance by agent type. However, when broken down by reliability level there was a small positive correlation for agent reliance in the mixed-reliability condition (r = .22). By breaking agent type down by reliability condition it was demonstrated that this effect was caused by a moderate positive correlation between pet anthropomorphism and participant reliance in the mixed-reliability different-type robotic agent condition (r = .42). This effect was further examined by looking at how pet anthropomorphism was significantly correlated to reliance in the low-reliability aid (r = .51, p = .003) but not to reliance in the high-reliability aids (r = .23, p = .19). This indicates that those high in pet anthropomorphism were more likely to rely on the low-reliability aid than those with lower pet anthropomorphism scores (See Figure 43). However, this effect did not affect trust ratings on high-reliability aids and was not apparent in the human or same-type robotic agent conditions (p > .05 in all cases).

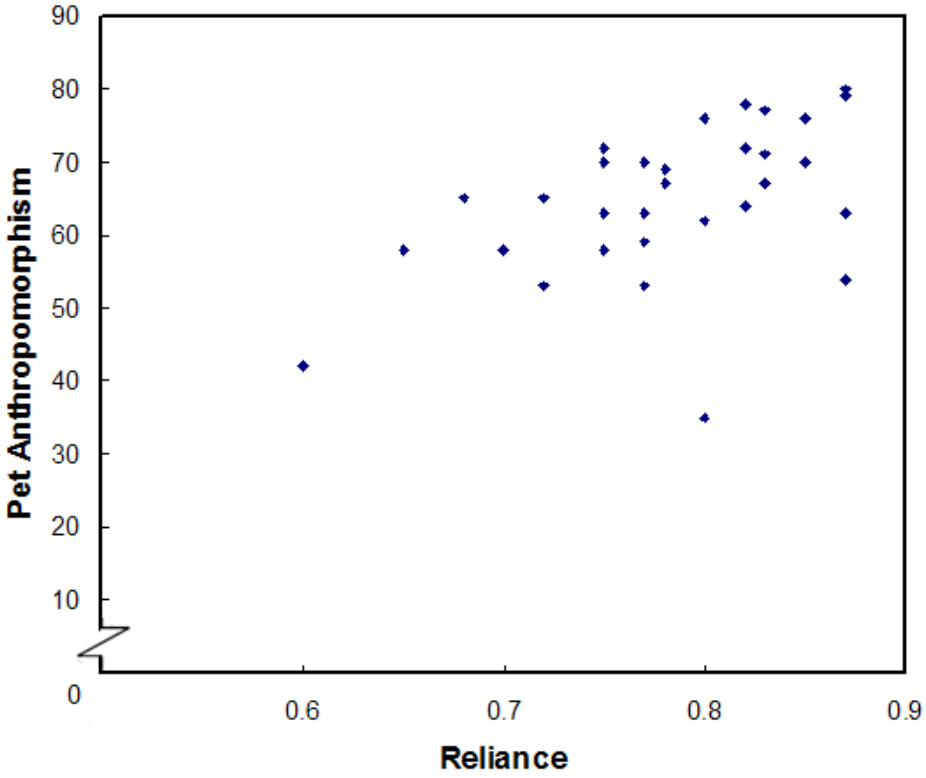


Figure 43. Pet anthropomorphism as a function of reliance on the low-reliability aid. Note that results are for participants in the different-type robotic mixed condition.

God or Deity Anthropomorphism

Mean score for God or Deity anthropomorphism was 29.36 (SD = 8.51; coefficient α = 0.93). The analysis on God or Deity anthropomorphism indicated that there was a significant difference between agent condition, $F(2, 287) = 3.45$, $p = .033$, $\eta^2 = 0.02$. The extremely small eta squared value indicates that even though this effect was significant it was of negligible size. All other effects were not significant ($p > .05$ in all cases; See Table 25). Additional analysis indicated that there were sex difference in terms of God or Deity anthropomorphism ($t(327) = 2.13$, $p = .03$). These results indicated that females have significantly higher God or Deity anthropomorphism scores than males, though this was a small effect size, $g = .24$ (See Table 23).

Table 25. God or Deity anthropomorphism among condition assignment. Note that SD are shown in parenthesis.

	Agent Type		
Reliability Condition	Human	Same-Type Robotic	Different-Type Robotic
Both High	27.44 (8.58)	29.69 (9.31)	29.51 (8.65)
Mixed	26.64 (9.38)	30.80 (7.82)	28.84 (7.75)
Both Low	26.64 (9.38)	28.93 (8.56)	34.11 (6.09)

In regards to correlations, the overall correlation of God or Deity anthropomorphism to trust and reliance were both not significant (r = .04 and r = -.07 respectively). There were also no significant correlations for trust or reliance by agent type or reliability condition. However, by breaking agent type down by reliability condition it was demonstrated that there was a moderate positive correlation between God or Deity anthropomorphism and participant trust in the high-reliability same-type robotic agent condition (r = .49) and the low-reliability different-type robotic agent condition (See Figures 44 and 45). However, this effect did not affect trust ratings on any other conditions (including mixed reliability conditions analyzed by low and high aid; p > .05 in all cases).

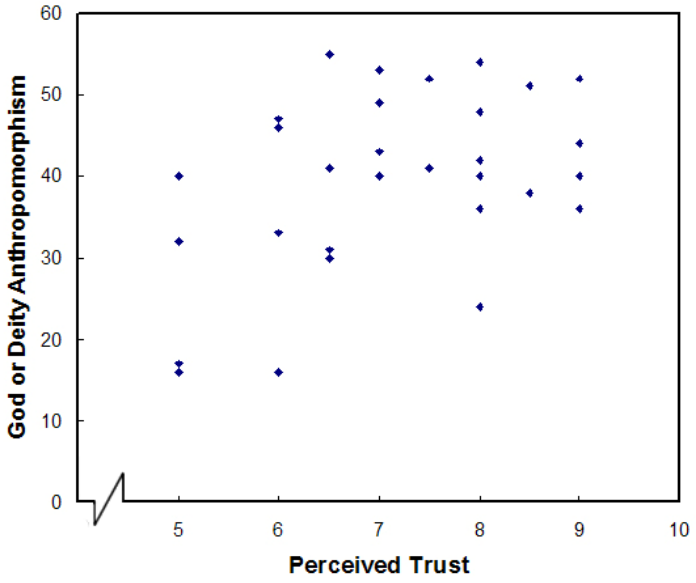


Figure 44. God or Deity anthropomorphism as a function of perceived trust. Note that results are for participants in the both high-reliability same-type robotic condition.

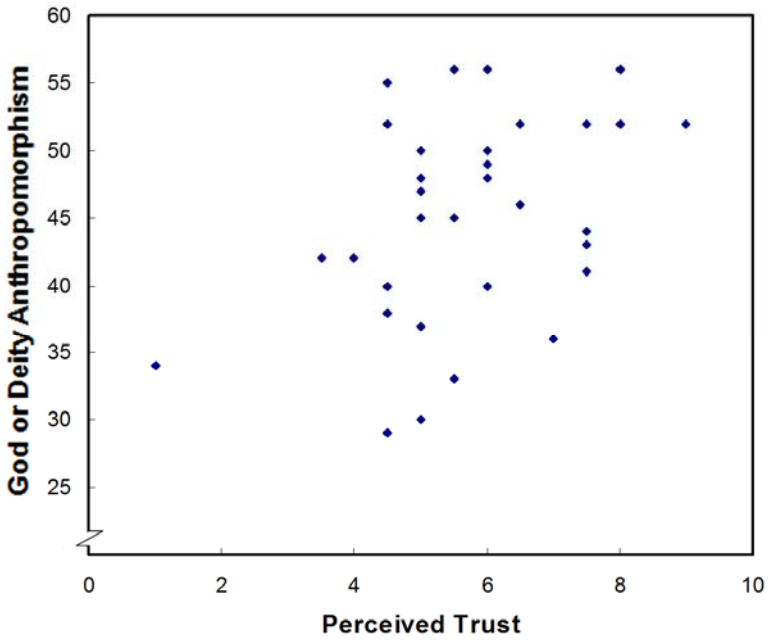


Figure 45. God or Deity anthropomorphism as a function of perceived trust. Note that results are for participants in the both low-reliability different-type robotic condition.

Negative Anthropomorphism

Mean score for negative anthropomorphism was 11.92 (SD = 4.38; coefficient α = 0.84). The analysis on negative anthropomorphism indicated that there was a significant difference between reliability conditions for levels of negative anthropomorphism, $F(2, 287) = 3.66$, $p = .027$, $\eta^2 = 0.03$. The extremely small eta squared value indicates that even though this effect was significant it was of negligible size. All other effects were not significant ($p > .05$ in all cases; See Table 26). Additional analysis indicated that there were no sex difference in terms of negative anthropomorphism ($t(327) = 1.18$, $p = .24$; See Table 23).

Table 26. Negative anthropomorphism among condition assignment. Note that SD are shown in parenthesis.

	Agent Type		
Reliability Condition	Human	Same-Type Robotic	Different-Type Robotic
Both High	12.03 (3.86)	12.55 (4.52)	12.29 (5.27)
Mixed	12.81 (4.46)	13.16 (4.01)	11.91 (3.99)
Both Low	11.02 (4.57)	11.55 (4.02)	10.60 (3.66)

In regards to correlations, the overall correlation of negative anthropomorphism to trust and reliance were both not significant ($r = .02$ and $r = .05$ respectively). There was also no significant correlation for trust or reliance by reliability condition. However, there was a small but significant positive correlation between negative anthropomorphism and agent reliance in the different-robotic agent condition ($r = .22$). By breaking agent type down by reliability condition it was demonstrated that there was a small to moderate negative correlation between negative anthropomorphism and participant reliance in the high-reliability same-type robotic agent

condition (r = -.35; See Figure 46) and a small to moderate positive correlation between negative anthropomorphism and reliance in the mixed reliability different-type robotic agent condition. By further analyzing the mixed-reliability different-type robotic agent effect by low and high reliability agent it was found that negative anthropomorphism was moderately positively related to reliance on the high reliability aid (See Figures 47). These results indicate that while negative anthropomorphism can lead to punishing the aid by not relying on it in inappropriate situations (i.e., both high reliability aids), it can also aid participants in allowing them to limiting their punitive efforts to only the unreliable aid in some conditions (i.e., mixed-reliability different-type robot condition).

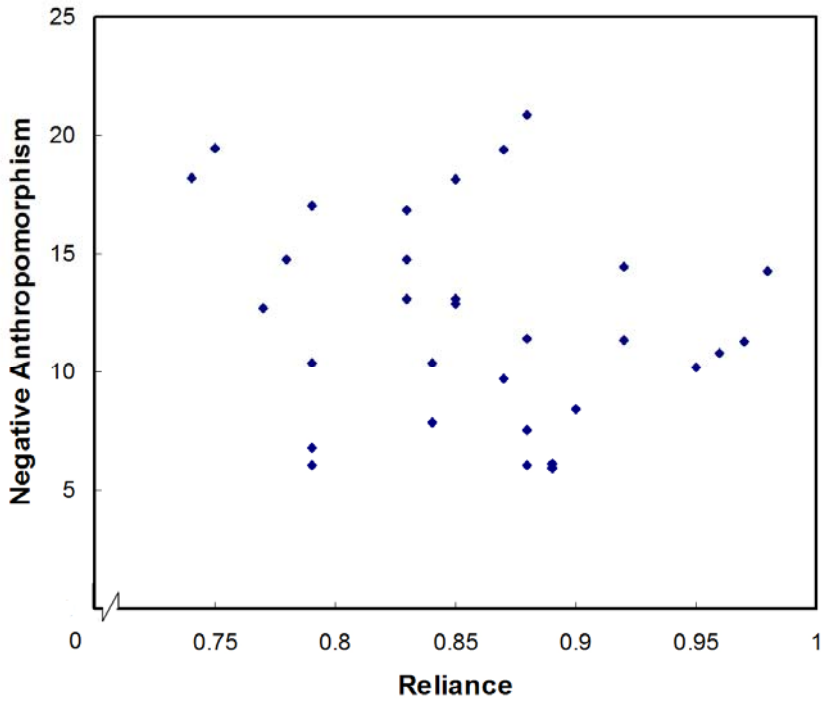


Figure 46. Negative anthropomorphism as a function of reliance. Note that results are for participants in the both high-reliability same-type robotic condition.

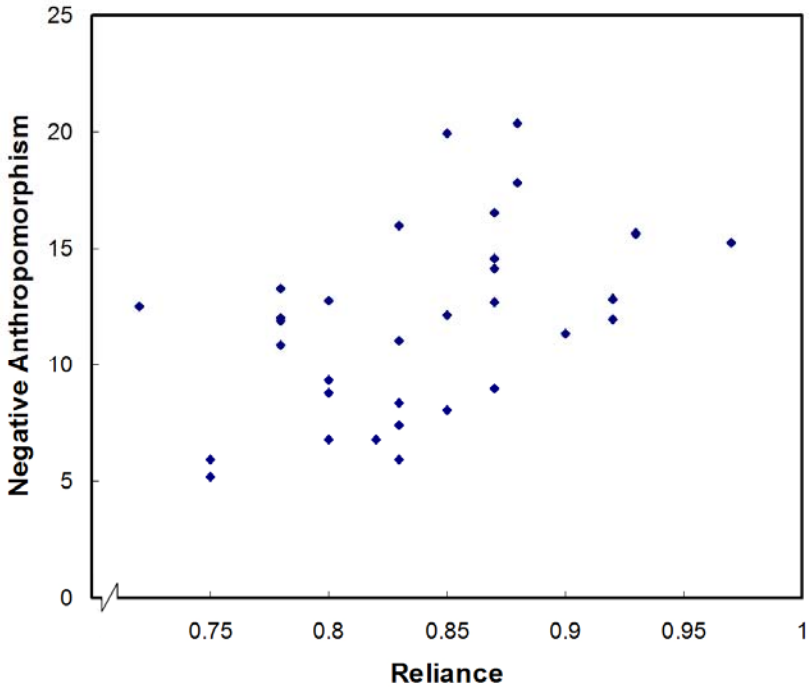


Figure 47. Negative anthropomorphism as a function of reliance on the high-reliability aid. Note that results are for participants in the mixed-reliability different-type robotic condition.

Interpersonal Trust Scale

Mean score for interpersonal trust was 85.10 (SD = 9.00; coefficient α = 0.52). The interpersonal trust scale was examined to see if the randomly assigned participants differed in terms of general trust level. Results from a 3 (agent type) by 3 (reliability level) univariate ANOVA with ITS score as the dependent measure indicated that agent type, reliability level, and the interaction between the two, did not differ in terms of ITS score ($p > .05$ in all cases). This indicates that by random assignment the experimental groups did not differ in ITS score distribution. Additional analysis indicated that there was a sex difference in terms of ITS score, ($t(327) = 2.61$, $p = .01$). These results indicated that females had significantly higher

interpersonal trust scores than males (M = 86.27, SD = 8.17 and M = 83.69, SD = 9.76 respectively). Analysis of effect size indicated that this was a small to moderate effect, g = .29.

In regards to correlations these correlations were done overall, by agent, by reliability condition, and by agent*reliability condition (only significant correlations are reported in the text for a full list of correlations see Appendix X). There were no overall significant correlations of interpersonal trust to rated trust or reliance on the agents (r = -.10 and r = -.07 respectively). In examining the data divided among agent type there was a negative correlation between ITS scores and reliance on human agents (r = -.21). Visual inspection of the data however indicated that this was a weak effect (see Figure 48). A second significant correlation was found in regards to agent-type. In this case there was a negative correlation between ITS score and trust in same-type robotic agents (r = -.28; see Figure 49). In examining across reliability conditions there was a significant negative correlation between ITS scores and trust in low reliability aids (r = -.20; see Figure 50); such that, individuals with higher interpersonal trust have significantly lower self-rated trust in uniform low-reliability agents. Further analysis of this relationship indicated that as ITS scores increase trust in human agents (r = -.35) and same-type robotic agents in the uniform low-reliability conditions decreases (r = -.43; see Figure 51 and 52 respectively). It is also useful to mention that in regards to pretrust measures there was no correlation between ITS scores and trust in the aids prior to interacting with them (r = -.08, $p > .05$).

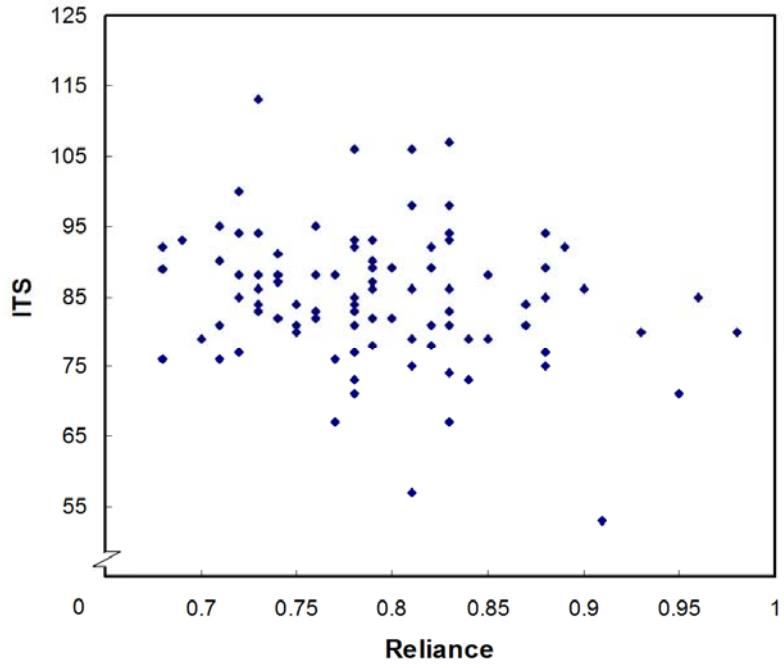


Figure 48. ITS as a function of reliance for human agents.

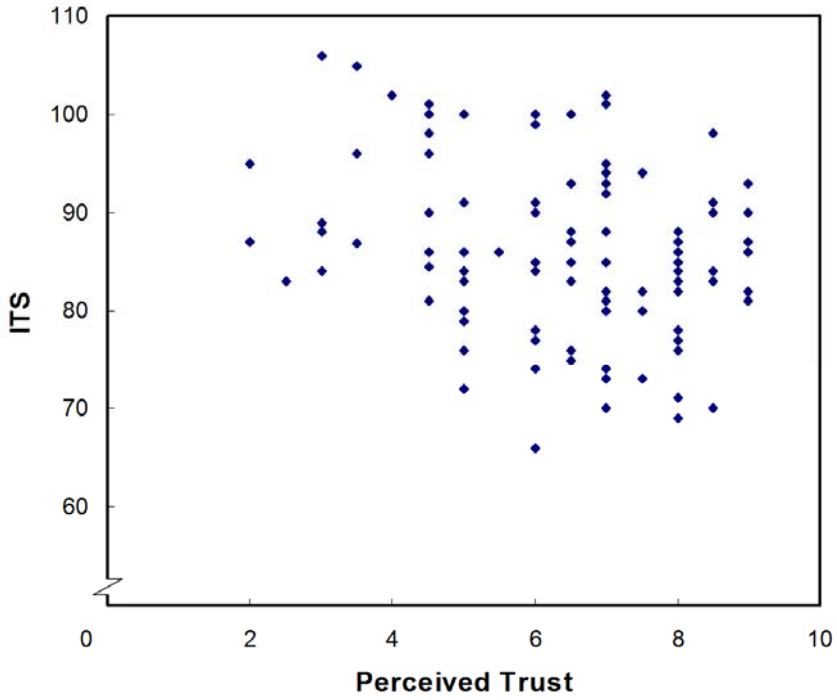


Figure 49. ITS as a function of perceived trust for same-type robotic agents.

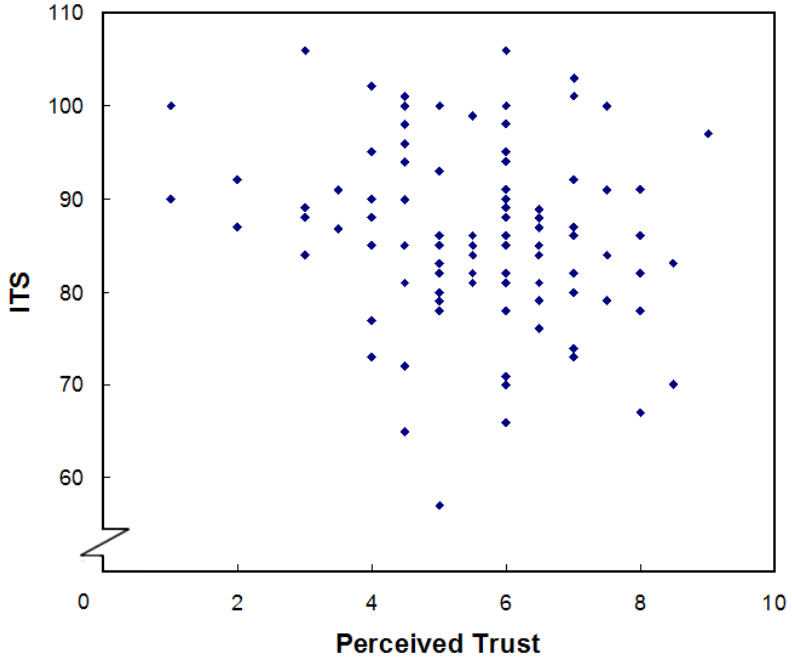


Figure 50. ITS as a function of perceived trust in the low-reliability condition.

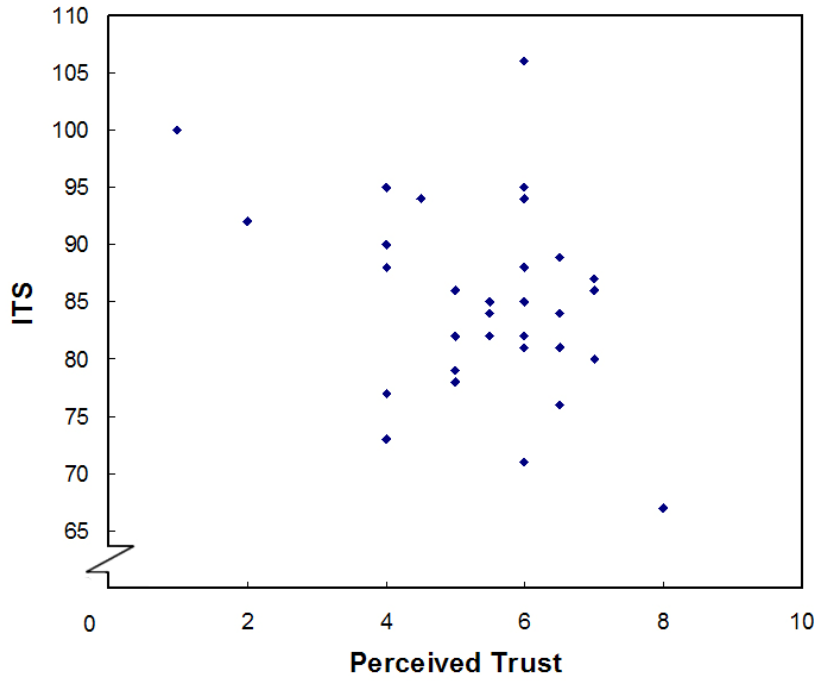


Figure 51. ITS as a function of perceived trust in the low-reliability human agent condition.

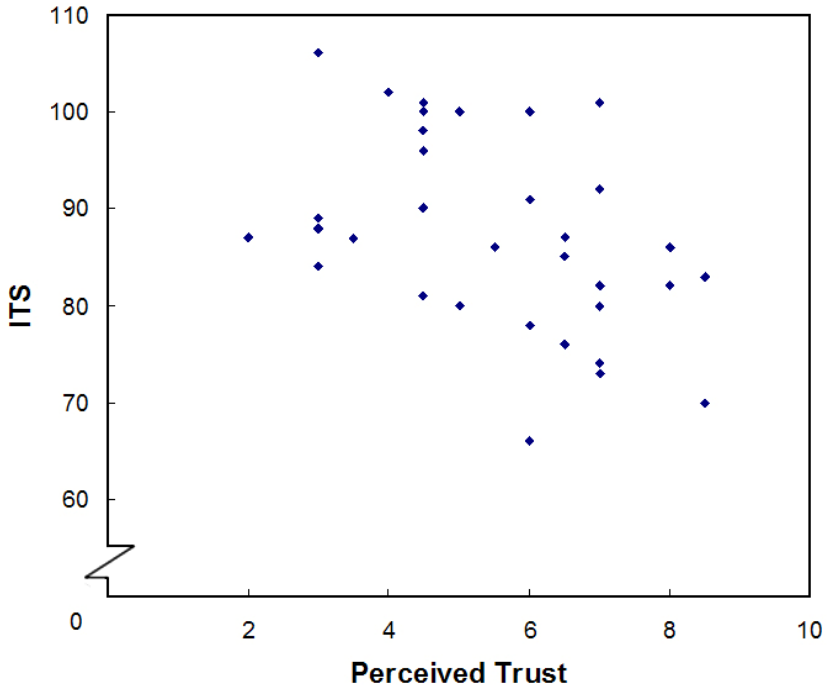


Figure 52. ITS as a function of perceived trust in the low-reliability same-type robotic agent condition.

Complacency Potential Rating Scale

In analyzing the CPRS there was an overall general score and four factors: Confidence, Reliance, Trust, and Safety. A 3 (agent type) by 3 (reliability condition) univariate ANOVA was conducted on each of the five divisions of the CPRS to examine if there were any differences among the between-subjects groups. Additionally, correlation analyses were conducted to examine if complacency potential scores correlated with the main variables of interest: trust and reliance. These correlations were done overall, by agent, by reliability condition, and by agent*reliability condition (only significant correlations are reported in the text for a full list of correlations see Appendix Y).

Overall CPRS Score

Mean score for the CPRS was 43.76 (SD = 5.49; coefficient α = 0.65). The analysis on overall CPRS score indicated that there was a significant difference between reliability condition, $F(2, 287) = 4.72$, $p = .01$, $\eta^2 = 0.03$. The extremely small eta squared value indicates that even though this effect was significant it was of negligible size. All other effects were not significant ($p > .05$ in all cases). Additional analysis indicated that there were no sex differences in terms of complacency potential, ($t(294) = 1.39$, $p = .16$).

In regards to correlations there was an overall correlation of complacency potential to self-rated trust ($r = .18$, $p = .002$) and reliance ($r = .14$, $p = .014$). In examining the data divided among agent type there was a positive correlation between CPRS overall score and trust ($r = .25$) and reliance ($r = .24$) for same-type robotic agents. Further analysis of the relationship indicates that this trust correlation is driven by the same-type uniform low-reliability condition in which there is a moderate correlation between overall CPRS score and trust in the agents ($r = .35$). A second significant correlation was found in regards to reliability condition, that is in the mixed reliability condition there were positive correlations between overall CPRS to trust ($r = .28$) and reliance ($r = .22$). Further analysis of the mixed reliability condition, examining trust and reliance in the high and low reliability aids, indicated that overall CPRS was significantly correlated to trust in the low-reliability aid ($r = .28$, $p = .006$) and reliance in the high reliability aid ($r = .28$, $p = .01$). Interestingly CPRS overall score was not significantly correlated to trust in the high reliability aid or reliance on a concurrent low-reliability aid ($p > .05$ in both cases). A significant correlation appeared for the different-type robotic mixed condition in which overall CPRS score was moderately correlated with average trust ($r = .56$). Further examination of this

effect, by examining the actual trust and reliance scores for the high and low reliability aids, indicated that trust in the low-reliability aid (r = .47, p = .006), trust in the high reliability aid (r = .49, p = .004), and reliance in the high reliability aid (r = .39, p = .02) were all significantly positively related to CPRS overall score. CPRS overall score was not significantly correlated to reliance on a concurrent low-reliability aid (r = .10, p = .57). These results indicate that complacency potential in general increases trust and reliance, especially in ambiguous situation (e.g., mixed reliability).

CPRS Confidence Factor

Mean score for the CPRS was 16.27 (SD = 2.36; coefficient α = 0.65). The analysis on CPRS confidence factor indicated that there were no significant difference between groups (in all cases p > .05). Additional analysis indicated that there was a trend for sex differences in terms of complacency potential factor confidence, (t(294) = 1.90, p = .06). The trend indicated that males (M = 16.59, SD = 2.32) were slightly higher than females (M = 16.09, SD = 2.23), but that it was a small effect (g = .22).

In regards to correlations there were no overall correlations of confidence complacency potential to self-rated trust (r = .09, p = .13) or reliance (r = .08, p = .18). In examining the data divided among agent type there was a positive correlation within same-type robotic agent condition for reliance (r = .20). There were no significant correlations for overall reliability level (p > .05 in all cases). However, when breaking the data down further into agent by reliability condition, it was found that within the different-type robotic aid the uniform-low reliability level's participant reliance was significantly correlated to CPRS confidence score (r = .37). Additionally, also in the different-type robotic mixed condition, CRPS confidence score were

significantly positively related to trust (r = .40) and reliance (r = .36). Analyzing this effect further by examining the mixed condition for trust and reliance on the high and low reliability aids themselves I found a positive correlation for CPRS confidence score and reliance on the high reliability aid (r = .21, p = .04). Taking this a step further and analyzing by agent it was apparent that the different-type robotic aid mixed condition had significant positive correlations between CPRS confidence and trust in the high reliability aid (r = .39, p = .02) and reliance (r = .44, p = .01) in the high reliability aid.

CPRS Reliance Factor

Mean score for the CPRS reliance factors was 10.90 (SD = 1.94; coefficient α = 0.15). The analysis on CPRS reliance indicated that there were no significant difference between agent or reliability grouping (in all cases p > .05). Additional analysis indicated that there was a small sex difference in terms of CPRS reliance, (t(294) = 2.01, p = .045). The effect indicated that males (M = 11.10, SD = 1.98) were slightly higher than females (M = 10.64, SD = 1.95) in terms of reported CPRS reliance, but that it was a small effect (g = .23).

In regards to correlations there was an overall correlation of reliance complacency potential to self-rated trust (r = .14, p = .02) but not on overall reliance (r = .06, p = .30). In examining the data divided among agent type there was a positive correlation within same-type robotic agent condition for trust (r = .27). Additionally in terms of reliability level there was a positive significant correlation for mixed-reliability trust (r = .30). When breaking the data down further by examining the trust and reliance within only the mixed condition by low and high reliability aid, it was found the CPRS reliance is significantly correlated to trust in a low reliability aid (r = .28, p = .01), trust in a high reliability aid (r = .21, p = .04), and reliance in a

high reliability aid (r = .24, p = .02). An overall correlation analysis of agent by reliability condition, found that the different-type robotic aid in the mixed reliability condition's trust score was significantly correlated to CPRS reliance rating (r = .43). By analyzing this in detail by examining how user rating differed between the individual low- and high-reliability aids, I found that the difference in trust at this level was determined primarily by trust in the high-reliability aid (r = .42, p = .02) rather than the low-reliability aid (r = .33, p = .06). This indicates that CPRS reliance factor is positively correlated with increased ratings of self-rated trust in general and also in conditions of ambiguity (e.g., interacting two-agents of the same-type, mixed reliability conditions, etc.).

CPRS Trust Factor

Mean score for the CPRS was 11.09 (SD = 2.11; coefficient α = 0.39). The analysis on the CPRS trust factor indicated that there were no significant difference between agent or reliability groups (in all cases $p > .05$). That is, random assignment allowed a relatively equal distribution of CPRS trust scores across between-subjects conditions. Additional analysis indicated that there were no sex differences in terms of trust complacency potential, ($t(294)$ = 0.06, p = .95).

In regards to correlations there was an overall correlation of trust complacency potential to self-rated trust (r = .18) but not on overall reliance. In examining the data divided among reliability condition there was a positive correlation within the low-reliability condition for trust (r = .20). There was no overall significant correlations among agent type. However, when results were examined by agent type and across reliability levels there were two significant conditions. These were, reported trust in the low-reliability same-type condition (r = .52; see Figure 53) and

trust in the mixed-reliability in the different-type aids (r = .43; see Figure 54). Further examination of CPRS trust within the mixed reliability condition indicated that in the different-type aid condition score was moderately correlated to trust in the low reliability aid (r = .40, p = .02) and trust in the high reliability aid (r = .37, p = .03).

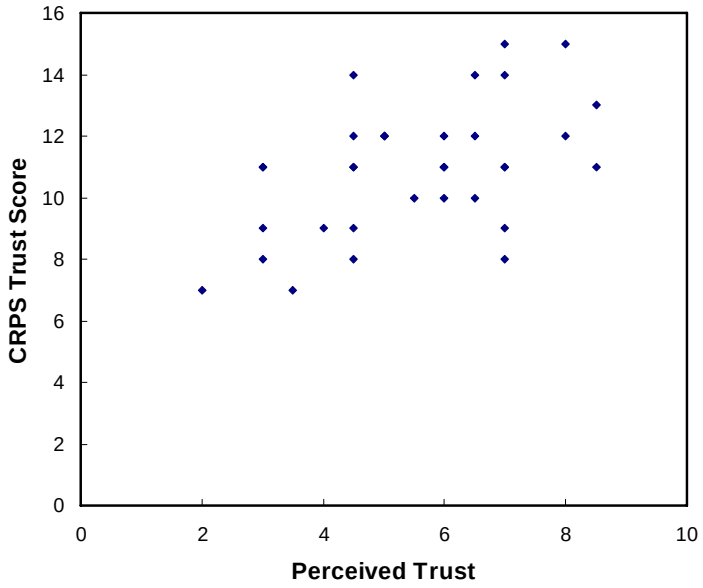


Figure 53. CPRS trust factor scores as a function of perceived trust in the low-reliability same-type robotic agent condition.

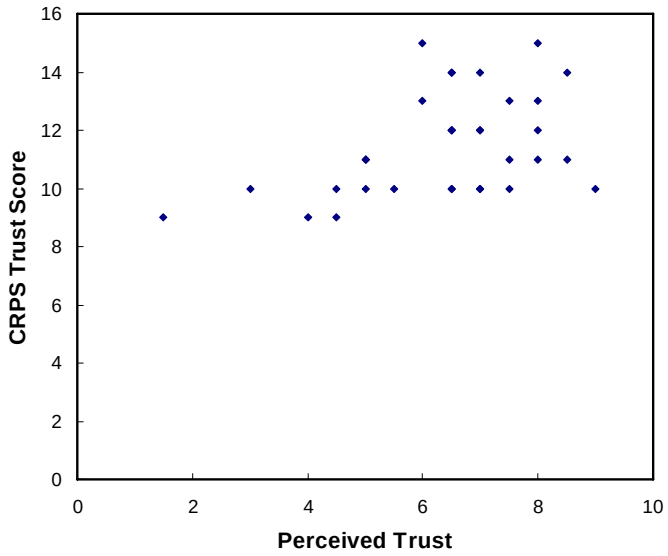


Figure 54. CPRS trust factor scores as a function of perceived trust in the mixed-reliability different-type robotic agent condition.

Overall Safety Factor

Mean score for the CPRS safety factor was 5.49 (SD = 1.66; coefficient α = 0.15). The analysis on the safety factor indicated that there was a significant difference among reliability conditions, $F(2, 287) = 5.14$, $p = .006$, $\eta^2 = 0.04$. The negligible eta squared factor however indicates that this, while significant, was trivial result. All other results were not significant (in all cases $p > .05$). Additional analysis indicated that there were no sex differences in terms of safety complacency potential, ($t(294) = 0.33$, $p = .75$).

In regards to correlations there was an overall correlation of reliance complacency potential to operator reliance ($r = .16$) but not on overall self-rated trust. In examining the data divided among agent-type there was a positive correlation within the same-type robotic aid condition for reliance ($r = .22$).All other correlations across agent type and reliability condition were not significant.

EXPERIMENT 4: DISCUSSION

Subjective Measures

The data concerning self-reported trust and automation reliance supported the hypotheses in several regards. First, participants were capable of accurately rating perceived trust and relying appropriately on the agents as a function of actual agent reliability. That is, even though the task was quite difficult, participants were carefully processing the responses of the agents and using these responses to rate perceived trust in the system. However, if trust and reliance always followed reliability level then the measures of mixed- vs. uniform-reliability for the low- and high-reliability aids would be equivalent. However, results demonstrated that there is biasing that does occur. Biasing occurs such that the low-reliability aid when it appears with a high reliability aid is viewed as significantly more trustworthy than when the low-reliability aids occur by them selves. On the other hand dissociation occurs because even though there is a subjective difference in the low-reliability aid depending on the reliability of the concurrent aid participant behavior toward the aid (i.e., reliance) does not change. Even though the low-reliability aid in the mixed condition is rated as significantly more trustworthy than the uniform low reliability aid, reliance on this *more* trustworthy aid is not different from the perceived *less* trustworthy aid. On the other hand, the high-reliability mixed- vs. uniform comparison indicates the opposite pattern of effects such that the agents are not rated significantly different in terms of perceived trust (though this effect was in the right direction), but do differ significantly in reliance, with participants relying less on high-reliability aids that occur in conjunction with a low-reliability aid.

These results can be construed in several ways. First in regards to perceived trust this indicates that the magnitude of the effect is much stronger for biasing trust upwards, when a low-reliability agent is portrayed with a high-reliability agent, then for biasing trust downwards when a high-reliability agent is portrayed with a low-reliability agent. This effect is particularly interesting when one takes into account that reliance on the aids differs for the high-reliability agents but not for low-reliability agents. The results of this study could be taken to indicate that operators respond in a more opened minded approach in a mixed-condition. That is, in the case of interacting with mixed-reliability participants are more critical in agreeing with high-reliability aids (reliance decreases – though their overall perceived trust in the agent is essentially the same), participants also become more willing to ascribe trust to a low-reliability agent (trust increase – though reliance does not change, that is operators still carefully weigh each of their agreements). This finding is supported by the fact that workload does not differ among reliability conditions, even under high-trust (i.e., high-reliability) situations workload is equivalent to workload in low-trust (i.e., low-reliability) situations. This indicates that operators are still mentally processing the task themselves regardless of their agents' reliability. Therefore it stands to reason that their reactions to the agents may be colored by their simultaneous processing of an agent of an alternative reliability. This effect may be more prevalent in this study since a low level of automation was used, that is while automation makes a recommendation an operator must select it to choose it. With higher levels of automation there may be a greater impact on reliance, such that operators become more complacent and less likely to process every trial when the automation is more autonomous. A higher level of automation should be studied to examine this issue.

Regarding agent-type and perceived trust and reliance it was believed that users would trust, rely on, and have less perceived workload when working with a pair of human agents. However, the main effect for agent-type was not significant across all three measures. This indicates that participants were not influenced by '*what'* the agent was when determining their overall trust, reliance, and workload. However, agent-type did influence temporal reliance when observing agent errors. While, this effect was expected its pattern was contrary to that hypothesized, in this study human agents had significantly less temporal reliance, compared to robotic-agents, following an observed easy or moderate salience error. It was originally believed that operators would be more forgiving of humans that made mistakes and less forgiving of machines that made mistakes (polarization bias), but this opposite effect occurred and indicates that users are actually more aware and punitively responsive to errors in other humans. Though the hypotheses that people would be more forgiving of people erring on more difficult trials but not simple ones was supported. However, the robotic agents did not followed the hypothesized pattern of results (i.e., that any error would cause an equivalent drop in reliance). There are several potential explanations for these findings. First, operators may assume that when human agents make simple errors that they are not focusing on the task (e.g., humans may be distracted or possibly not trying very hard); this could cause the operator to be negatively conditioned to agreeing with them on the next trial. On the other hand a robotic agent could make a simple error and this could be construed to be an accidental glitch (e.g., interpreting a stereo as an IED) that is not a byproduct of negligence or inattention of the aid but merely bugs in the program. While both would negatively impact overall reliance, purposeful and emotionally laden interpretations of human errors could lead to greater drops in temporal reliance then 'unintentional' robotic errors. This was supported by the fact that low-salience errors (i.e., difficult trials) did not

experience this decrement for human operators. That is, when the participant found the trial quite difficult themselves they became equally likely to negatively respond to the human on the next trial compared to the robots. There are two alternative reasons for this explanation, the first is that humans are more accepting of human-agent errors on difficult trials (i.e., they attribute faults less to negligence and more to the difficulty of the task), the second explanation is that at this level of difficulty many participants may have been unaware of the errors completely, thereby minimizing the effectiveness to detect this effect. It would be recommended in a future study to obtain a measure of participant error detection (i.e., whether the participant detected the agent failure) and to analyze the temporal reliance in only those conditions where users did indeed notice the failure. An alternative explanation could be that operators treat "intelligent" machines with the same or more forgiveness than they would treat humans with. While observers may be more critical of a calculator returning the correct answer every time, they may be more lenient to more complex forms of automation. In this way "intelligent" automation benefits from both worlds in that operators do not ascribe negative emotional connotations to the agent's errors and they also are forgiving of mistakes realizing that the system is imperfect but can on the whole work quite well. This theory should be examined by future research.

Biasing-Effects and Trust

Agent-type also appeared to effect trust biasing on the mixed vs. uniform reliability conditions. However, the pattern of trust biasing conflicts with the experimental hypothesis, which had predicted the opposite pattern of results: that humans would be viewed as the most independent agents (lowest ES difference), different-type robots would be viewed as semi-

independent (moderate ES difference), and same-type robots would be viewed as the least independent (largest ES difference). The rational for this argument was that two human beings are unique, and interacting with one person should not influence your trust in another person albeit if they are concurrent and of mixed reliability. However, two robotic agents of the same-type, whom you have been informed are operating under similar mathematical algorithms and created by the same company, should appear to be less independent. Thus, an error on the part of an inaccurate robot should be more likely to bias trust in a concurrent accurate robot, making an operator trust it less. The opposite effect could occur where an accurate robot could bias trust in a concurrent inaccurate robot, making an operator trust it more. As reported earlier, in general it was observed that the mixed reliability condition did cause a biasing effect which caused greater trust in inaccurate agents and less trust in accurate agents However, the predicted pattern of agent biasing was not supported by the experiment. It appears that people interact with automation in a complex manner when it comes to determining their perception of agent trust. While it was demonstrated that people could indeed differentiate the difference between a high reliability aid and a low reliability aid, depending on *what* they thought those aids were influenced their adjustment in their trust level. The data indicated that, contrary to the hypotheses, the human agents had the largest perceived changes in trust between mixed and uniform conditions. This indicates that human beings are actually very sensitive to performance differences between *people*, and that individuals change their criteria for trustworthiness in their human teammates quite dramatically based on the combination of people they are viewing. For example, high-reliability agents when paired with a low-reliability agent have significantly higher ratings of trustworthiness compared to the uniform-high-reliability human agents. In this way it appears that a human agent's stellar performance contrasted against a less reliable human performer

actually causes people to consistently rate the trust in the stellar performer much higher than they would if they just viewed two high-performers. Surprisingly though our less stellar performer in the human agent condition is not in contrast worse at the task, but instead benefits from association with the high performer. The moral of the story appears that if you are good at a task surround yourself with people who are not and you will be perceived as be more trusted by your colleagues, on the other hand if you are not good at a task it would be wise to surround yourself with people who are so that by association you can seem more trustworthy.

On the other hand when you interact with automation the story becomes slightly different. According to this study, when one interacts with two agents of the same-type, the high-reliability aid suffers in terms of trustworthiness by being associated with a lower reliability aid. This is the equivalent of losing faith in a particular device when you experience low reliability on a similar device. One becomes less trusting of the high reliability aid because of the now salient chance of errors. Additionally, when the inanimate aids are of similar make and model there is no trust benefit to the low reliability aid for occurring concurrently with a high reliability aid. This is a very cynical model of trust, such that mixed reliability only brings no change or decreased trust, a very strong contrast to the human-agent condition.

The final group of analysis for perceived trust is the different-type robotic aids. In this condition the hypotheses supported the hypotheses in terms of the different-type robotic aids having a mixed effect between what occurs for the human and same-type robotic agents. The pattern of results follows the originally anticipated direction, such that a concurrent high reliability aid raises trust in a low reliability aid (similar to what occurs with low-reliability human agents) and a concurrent low reliability aid decreases trust in a high reliability aid (similar to what occurs with high-reliability same-type robotic agents). However, it was unanticipated

that this biasing would occur to such a point that there was not a significant difference in regards to self-rated trust in the low- and high-reliability agents in the mixed-reliability different-type robotic agent condition. That is, while participants were able to determine high- and low-reliability in the uniform conditions, the degree of bias in the mixed condition made the trust ratings between the low and high reliability aids not significantly different. This could have detrimental consequences in an applied setting in which individuals could fail to identify inaccurate machine teammates because their inaccuracies are masked by the biasing effect of more reliable machine teammates.

Biasing-Effects and Reliance

As mentioned earlier reliance data demonstrated that people were able to adjust their reliance so that they could rely more on reliable aids and less on non-reliable aids. By examining the amount of bias that occurs when a reliability condition is paired with a concurrent different reliability, it was also apparent that the reliance bias between agents was minimal and in general followed the predicted pattern of results: humans the least bias, different-type robotic an intermediate amount of bias, and same-type robotic agents the most bias. This was particularly interesting considering the odd pattern of results for perceived trust in the agents. For example, ratings of trust for human agents were highly positively biased in the mixed-reliability condition; it appeared that by adding a comparison, both human agents increase in terms of trustworthiness. On the other hand, in terms of reliance, a mixed-reliability condition actually lead to less reliance on the high-reliability human agent and no change in reliance on the low-reliability human agent. Is there cognitive dissonance that is driving this hypocrisy? Why do participants report increased

trust in the agents but then not follow their subjective reports with changes in their behavior? One speculation is that individuals are more cautious with actions than with words, while trust ratings varied more significantly across the groups reliability ratings were more consistent (though this may also be due to the greater precision of the reliance measure). However, there is additional evidence for a cautious reliance approach, in that the low-reliability aid rarely benefited from its relationship to a high-reliability aid. Indeed, in all but one case individuals mitigated their reliance on both the high- and low-reliability aids when they were paired with an aid of conflicting reliability. Even in the case of high-reliability humans in whom trust was rated as significantly higher in a mixed-condition, participants still hedged their bets by not increasing their reliance. One possible explanation for this finding is that mixed-reliability allowed participants to be more open-minded about whether or not they agreed with the aid. As Table 21 demonstrates in all but one condition the mixed-reliability ead to less reliance on the agent (even compared to the low-uniform conditions). This indicates that teammate conflicting reliability levels make it more acceptable to disagree with either teammate's recommendation. This finding exemplifies why it is important to gather both subjective and objective data on user perceptions toward automation. If trust always followed reliance there would be little reason to collect them both. Thus how the subjective measure of trust links to the behavioral measure of reliance and the conditions in which trust and reliance dissociate are of distinct importance. There are practical reasons to predict the conditions that will cause dissociations, particularly to alert designers to the kinds of biases that they will encounter in operators of complex automated systems.

An alternative hypothesis concerns the temporal nature of the measurements themselves. Trust is measured at the end of the experiment, so as to gain a general measure of trust in the

agent. On the other hand, reliance is measured on a trial-by-trial basis. Human memory does not sum trust in the same way as a computer program sums their reliance score. Trust summed in this experiment means greater trust for the low-reliability aid. These after-the-fact ratings of trust indicate that biasing occurs to the benefit of the low-reliability aids. That is, participant's subjective evaluation of trust in the agents is positively affected by exposure to a higher reliability aid. Though its interesting that this effect does not extend to greater reliance on a trial-by-trial basis. These trial-by-trial reliance measures summed means less reliance for high-reliability aid in a mixed condition. On a trial-by-trial basis operators are more susceptible to negatively biasing their reliance on high-reliability aids when they are presented with a concurrent low-reliability aid. That is, observed errors in the low-reliability aid may prompt the observer to disagree with the high-reliability aid more often. Though again it is interesting that overall trust scores do not change.

Individual Differences

This last section examines individual differences and how they related to participant trust and reliance in the task. These analyses were done in an exploratory fashion.

Sex Differences

Participant sex overall did not affect user reliance on the aids or trust in the aids ($p>.05$ in both cases). This was expected because the automation literature has not demonstrated a sex effect. There were however some interesting sex effects in regards to several factors measured by the individual differences questionnaires, these included pet anthropomorphism, God or Deity

anthropomorphism, interpersonal trust, complacency confidence, and complacency reliance. However, these effects were relatively small (*g* ranged from 0.22 to 0.29) and may be more important predictors of performance in more specific cases (e.g., studies dealing with pets may show a slightly different pattern of results for female participants that is not present for male participants).

ATS

Through random assignment ATS scores were relatively equivalent across conditions, and if there were differences eta squared values indicated that the differences were negligible. This allowed for some interesting effects to be observed. For example, extreme anthropomorphism was significantly related to the rating of trust individuals would assign a low-reliability aid in the different-type mixed reliability condition. This is interesting because it indicates that people higher in extreme anthropomorphism have a stronger negative reaction to low reliability aids when those aids appear physically different. That is they appear more heightened to the independence of the aids in this condition than those with lower extreme anthropomorphism scores, and this is reflected in their subjective-trust ratings. However, higher levels of pet anthropomorphism had an opposite effect; participants became more likely to rely on a low-reliability aid in the different-type mixed condition. This effect is unusual, individuals with high pet anthropomorphism are more likely to ascribe human like traits to a familiar animate object (i.e., pets), but it is somewhat unclear how this trait relates to reliance upon faulty aids when those aids are different inanimate robotic agents. It is the author's speculation that this effect may be from anthropomorphism of pets being somewhat related to anthropomorphism of

inanimate objects. Some of the questions querying pet anthropomorphism query participants on whether they would reward a pet for doing something good and apologize for hurting a pet. In this manner if individuals rewarded a robot for doing something good, that could be construed as agreeing with the aid when it is correct. While apologizing for hurting a pet, could loosely be construed as being considerate to a pet or in this case considerate of an agent's recommendation. Therefore, those high in pet anthropomorphism may be more likely to agree with the aid when it is right to 'reward' it, while those low in pet anthropomorphism may not feel bad for ignoring (i.e., being inconsiderate of the aids recommendation) low-reliability agent recommendations, thus leading to the significant difference in reliance. However, this is only apparent in the different-type robotic condition, perhaps because two same-type robots may be too similar to activate pet anthropomorphism. However, this speculation should be studied further in future studies.

Another interesting effect uncovered by the ATS is that anthropomorphism of God or Diety leads to greater ratings of trust in two instances: high-reliability same-type robotic agents and low-reliability different-type conditions. While religious faith has been found to be positively related to generalized level of trust, these results were in very specific circumstances Visual inspection of trust graphed across all the conditions indicated that as anthropomorphism of God increased so did rated trust in the agents in general, but that possibly by chance these two groups had fewer deviations from this general pattern and more favorable pattern of scores. It was also a limitation that affiliated religion was not recorded; several participants reported trouble answering the God anthropomorphism questions because they were Atheist or Agnostic. It might clarify results if they were removed from analysis. It might also clarify results to divide among the remaining religions as some participants reported that they believe that God became

man thus they choose *high* anthropomorphism while other participants mentioned that God is much greater than man so that they reported much *lower* anthropomorphism. However, in regards to trust religious affiliation is know to be positively correlated to generalized trust, thus it would be interesting to extend this to examine whether this relationship is related to how one anthropomorphizes their God or Deity.

The final analysis of the ATS concerned negative anthropomorphism, that is how likely one is to lash out at an inanimate object when it does something you do not like. Results indicated that negative anthropomorphism can be beneficial or harmful depending on the situation. That is, negative anthropomorphism can lead to punishing the aid by not relying on it in inappropriate situations (e.g., both high reliability aids that look similar), but it can also facilitate participants in allowing them to limit their punitive efforts to only the unreliable aid in some conditions (e.g., the different-type aid mixed condition in which negative anthropomorphism was correlated with greater reliance on the high reliability aid). This indicates that negative anthropomorphism may help participants by having them harshly judge one inanimate object but not a concurrent more reliable inanimate object, but that this relationship is in part determined by external physical cues.

ITS

The results regarding interpersonal trust scores were quite surprising, the indicate that those high in interpersonal trust were in general actually less trusting of the agents after interacting with two unreliable aids then those scoring lower on the interpersonal trust scale. This provides empirical evidence that not only are high generalized trust individuals not gullible, but

that they also respond more harshly to those items that violate their trust (i.e., they rate perceived trust lower after interacting with low reliability aids).

CPRS

Analysis of the CPRS scores were found to be relatively lower than those found by Singh et al (1993; current study: $M = 43.76$, $SD = 5.49$ vs. Singh et al. study: $M = 57.69$ $SD = 6.09$). The CPRS tended to have a small positive correlation to gaming experience ($r = .11$, $p = .05$). However, the CPRS did not obtain significance in relation to participant age, sex, or computer experience ($p > .10$ in both cases). In regards to age and computer experience this was unusual because age and computer experience are typically related to CPRS scores. However, our lack of finding a correlation with age is most likely due to the restriction on the range of ages examined ($M = 21$, $SD = 5$). On the other hand, computer experience, as measured by number of hours a week spent on a computer, was normally distributed but still not related to CPRS scores ($r = -.05$). This indicates that computer experience does not necessarily lead to automation complacency, and that other factors are at work (e.g., type rather than quantity of computer experience) or other individual variables within the sample studied.

Across the main variables of interest, trust and reliance, overall CPRS was significantly positively related to trust and reliance on the agents. This was especially present in conditions of ambiguity (e.g., same-type agents, mixed reliability condition) and in cases in which the observer should not have relied upon the agents (e.g., when participants became complacent in the same-type low reliability agents and trust in the low-reliability aid in the mixed reliability condition).

However, in the mixed reliability condition overall complacency in general increased trust in both high and low reliability aids, but it only increased reliance on high reliability aids.

Examining the factors of CPRS it was not surprising to see that automation confidence was significantly related to reliance and trust in automation in conditions of ambiguity (i.e., same-type agents, different-type low reliability agents). Complacency reliance was surprisingly not related to reliance but was related to self-rated trust again in conditions of ambiguity (i.e., same-type agents, different-type low reliability agents, and mixed reliability conditions). Complacency trust was significantly correlated with trust overall, and again was significant in cases of ambiguity (i.e., same-type agents, and low reliability conditions). The last factor safety indicated that it was correlated to reliance especially in terms of an ambiguous situation (i.e., interacting with two of the same type aids).

Overall the results of the CPRS indicate that automation complacency does indeed increase reliance and trust in automation; however, these effects differ based on the nature of the task. It seems that in general complacency helps guide behavior when the task is ambiguous; that is, the operator is interacting with aids that differ in their reliability or appear physically the same. In this way operators who have higher levels of complacency may give-up their choice (i.e., rely or trust an agent) more quickly in cases of uncertainty because they trust that the agent will operate in their best interest. Interestingly, it appears that this complacency does affect trust in low-reliability aids but does not affect reliance in low-reliability aids as much. This may mean that those high in complacency are more likely to trust their teammates (regardless of their reliability) and actually rely on high reliability aids, but that often they do not typically have a greater predisposition to rely on low reliability aids.

Limitations to the Current Study

A limitation to the study is that some individuals questioned the human agents' similarity. The male and female human agents were strikingly similar, due to the facial compilation software, and some participants found that uncanny and it may have weakened the effect of the manipulation. It may be beneficial in future studies to actually *not* use averaged faces so as to increase the believability of the manipulation. While this reduces the control of the agent manipulation, in studies of trust it is imperative to limit skepticism in participants. However, the human agents were believable to many participants as they often used pronouns (e.g., "he" and "she") when discussing their teammates in the open ended question on the exit questionnaire.

Another limitation is this study is the measure of overall trust was a single question on a nine-point Likert scale given once at the end of the experiment. Many studies in this realm use the same or a similar Likert scale to garner information on trust but query participants continuously throughout the session, often after every trial. However, trust is an attitude that develops over time and by querying participants repeatedly on this attitude the researcher may not be measuring trust so much as belief the aid was just correct on the previous trial. By limiting the trust query to the end I minimized distraction to the participants and obtained an overall view of the agent (not a point by point report). However, other studies that have had their participants perform trial by trial ratings may have increased accuracy and power of this measure to detect an effect. In order to examine whether a temporal facet of the measures lead to the surprising trust and reliance dissociations it would be fruitful to replicate this study taking a point-by-point measure of trust. This would allow the examination of whether operators are more critical of their trust on a trial-by-trial basis.

Additionally, temporal reliance for low-salience errors was found to not significantly differ between humans and automation. However, this lack of effect may have been due to the low salience errors being so difficult that they were not detectable by the operators; hence, operators were unaware of the automation error because the trial was difficult artificially demonstrating no difference among the agents. It would be recommended in future studies to measure participant error detection and to limit analysis to users who did in fact detect the automation failure. Alternatively, the division of error salience could be shifted up to ensure participants noticing the errors and their evaluation of the difficulty of the error then impacting their perceived trust and reliance on the agents.

Proposed Future Research

The current research was a first step at examining how human operators calibrate their trust and reliance to fellow humans and/or robotic agents. For robotic agents responses were examined for whether people believed they were working with two agents of the same- or different-type. For human agents only different-type agents were used (i.e., a male and a female agent). It poses an interesting question about whether it is possible to vary the perceived independence of human agents. Would humans that are more similar, such as identical twins, clones, or more realistically individuals who are very similar based on their dress, training, and appearance, have different patterns of trust and reliance bias compared to two distinct individuals? The findings of this study also indicated that individuals did not respond differently to the agents if they were male or female; however, it would be noteworthy to examine when operators interact with teams of the same or different sex. In groups of all males is there more

distrust than groups of all females, additionally if the participant is the opposite sex/different age/different race than the human agents would that result in them feeling more like an outsider and relying more on their teammates? Would facial expressions impact users perception and behavior toward these human teammates? Additionally, what about the characteristics of the robotic agents? In the current study two standard but distinct robotic agents were employed; however, it would be interesting to examine how anthropomorphic robots (e.g., the Sony QRIO or AIBO) might bridge the gap between the differences in how trust and reliance spreads in human vs. robotic systems. These questions examine how different characteristics of the agent can influence the social interaction between the agents and the operator. The purpose of the current work was merely to see if there is a difference between how trust and reliance spreads in human compared to robotic teams, now that it is evident that it does spread differently the next step is naturally to see how characteristics of the agents can influence this spread (possibly through human agent conformity or increasing anthropomorphic characteristics of robotic agents).

The present study investigated the spread of bias in a system in which an operator monitored two agents; it would be of interest to investigate how trust and reliance were biased in more complex systems (3+ agents to monitor). Would bias between the agents decrease as more agents were being monitored, similar to an averaging out effect? Or alternatively would there be more bias because the complexity of the task may prevent users from developing accurate representations of each agent's reliability? The effect of the experimental test bed is another avenue for future research. The current study operated under a scenario that people's lives were in danger while many previous trust studies have investigated trust using juice pasteurization

tasks. Perceived importance of the task may effect how individuals allocate their trust and reliance in agents.

The addition of stress and subsequent examination of its impact on operator trust and reliance is another avenue of future research. I believe that stress would put participants in a situation of greater need for reliance on automation, and that while trust levels may still fluctuate reliance would be much more stable due to its greater need. Increased stress may however cause a stabilization of trust levels if the operator becomes so stressed that it impairs their ability to monitor the agents adequately to establish a set level of trust relative to their observed performance.

GENERAL DISCUSSION

The four studies entailed in this report allowed for the examination of task, which was calibrated to be difficult but possible for manual performance, when that task was paired with an automated aid that differed in reliability, perceived agent characteristics, and error salience. These results are based on a task in which operators monitored the decisions of agents on remote unmanned vehicles. Other possible applications of missions in which human operators would act through remote vehicles are hazardous material handling, emergency response operations (e.g., bomb removal), fire operations in searching burning buildings, extreme environments (e.g., Mars Land Rover), and even medical applications (e.g., nanomachines). For example even in the case of injecting nanomachines with the goal of clearing plague from arteries, much of the process could be preprogrammed but a physician/operator to monitor the activity and to provide ongoing regulation, especially in the cases of unexpected circumstances. The environment of operation and vehicle dynamics may be radically different, but the fundamental interface contains a number of commonalties (Mouloua, Gilson, & Hancock, 2003).

One of the most essential elements of any social organization, whether it is a professional soccer team or a military reconnaissance unit, is the willingness of the members of that social group to trust one another. The efficiency, adjustment, and even survival of any social group depend upon the presence or absence of such trust (Rotter, 1967). In fact, almost all of our daily activities, from buying gasoline, paying taxes, going to the dentist, flying to a convention involve explicitly trusting someone else (Rotter, 1971). Rotter (1980) has argued that as distrust increases the social fabric disintegrates, in order to support a complex society we must accept greater dependence on others. If trust in general weakens then this stands to reason that the social

interaction may also weaken and possibly collapse. Trust in regards to user reliance on human and robotic agents appears to be a fruitful area of investigation, and the previous analysis demonstrats that people respond in complex and intelligent ways to imperfect teammates. This interaction is of particular interest to engineers who should focus on how their design and the environment in which the operator will be interacting with the agent will influence trust and reliance on the aids. Without appropriate trust reliance on the system goes down, and the system may fall into disuse and eventually be abandoned. On the other hand, with too much trust operators may fall to detect automation failures and the safety of the system may come into question. Engineers must take the social-interaction into account to ensure that their systems are used safely.

These studies represent a first step in examining the complex interaction in how individuals cooperate to complete a task when paired with teammates. Applications of this work include understanding how subjective states impact reliance on automation and human colleagues. This has important connotations for human-human and human-machine systems in aviation, navigation, process control, military, and other applications. It also has important implications for automated tasks from human-human systems to human-machine systems. The present data suggest that while operators are able to differentiate between reliability levels in terms of trust and reliance, trust becomes quite biased when dealing with two agents of mixed reliability. However, individuals seem to be able to keep their reliance upon these agents to relatively nonbiased levels. So it appears that people are able to compensate their behavior to control for changes in subjective state at least in the bounds of this study. Additionally, some differences in agent type on biasing between trust and reliance, were found lending empirical support to the notion that humans and automation are *not* interchangeable and that users respond

differently to the exact same recommendations depending on how they respond socially to that agent (i.e., respond believing agent is robotic [same or different type] or respond believing the agent is human).

Guidelines

Drawing on the conclusions of this study several guidelines for system design, for when an operator and dual agents interact, have been created.

1. If possible use multiple agents of similiar high reliability.
2. If mixing agents of different reliabilities can not be avoided, expect and design for a drop in reliance across both the low reliability aid and the high reliability aids.
3. When using robotic aids, to prevent polarization bias stress the 'intelligent' aspect of the robotic automation.
4. In all cases, but particularly for those interacting with other human agents, there is a drop in reliance following obvious errors. So design for residual drop in reliance that may occur after teammate errors.
5. Dissociations can occur between trust and reliance, so even if operators report verbally trusting a system their actual use of that system should also be examined.

APPENDIX A: DEFINITIONS OF COMMONLY USED TERMS

DEFINITIONS

Automation: Any sensing, detection, information-processing, decision-making, or control action that could be performed by humans, but is actually performed by a machine (Moray, Inagaki, & Itoh, 2000, p. 44). Alternative definition: The execution by a machine agent of a function that was previously carried out by a human (Parasuraman & Riley, 1997).

Automation Reliance: Defined in terms of performance or behavioral measures such as automation utilization and efficiency (Wiegmann, Rich, & Zhang, 2001, p. 356).

Automation Trust: Defined in terms of subjective measures, such as users' confidence ratings in the automation or their verbal estimates of the automation's reliability (Wiegmann, Rich, & Zhang, 2001, p. 356).

Automation Use: The voluntary activation or disengagement of automation by human operators (Parasuraman & Riley, 1997).

Disuse: The neglect or underutilization of automation. Often represented by ignoring or turning off automated alarms or safety systems. A common cause of disuse is a high level of false alarms in the system (Parasuraman & Riley, 1997, p. 233).

Misuse: Overreliance on automation, that is using automation when it should not be used, which can result in failing to monitor it effectively (Parasuraman & Riley, 1997, p. 233).

Reliability: The accuracy of the machine or the likelihood that an objective can be achieved by automation (Beck, Dzindolet, & Pierce, 2002, p. 67).

Self-Confidence: Anticipated performance during manual control (Lee & Moray, 1994, p. 154).

Trust: Automation is seen as trustworthy to the extent that it is predictable, dependable, and inspires faith that it will behave as expected in unknown situations (Beck, Dzindolet, & Pierce, 2002, p. 68). Also defined as the expectancy held by an individual or a group that the word, promise, verbal or written statement of another individual or group can be relied upon (Rotter, 1967, p. 651).

APPENDIX B: INFORMED CONSENT TO EXPERIMENT 1 AND 2

Informed Consent Form

Please read this consent document carefully before you decide to participate in this study.
You must be 18 years of age or older to participate.

Project title: Empirical Examination of Trust in Automation across Multiple Agents in a Search and Rescue Operation.

Purpose of the research study: The purpose of this data collection effort is to determine the impact of using automated decision aids in a search-and-rescue scenario. The current effort seeks to determine under what conditions automated decision aids increases or decreases reliance upon these decision aids.

What you will be asked to do in the study: You will be asked to view a computer display running a simulated scenario of a search-and-rescue scenario using one unmanned ground vehicle (UGV). You will be asked to monitor the video images from the UGV for critical signals (e.g., enemy units, civilians, or weapons). At the end of the session you will be asked to complete several questionnaires about your experience performing the search-and-rescue scenario.

Time required: Approximately one (1) hour.

Risks: Minimal. The risks to you are no greater than operating any other computer.

Benefits/Compensation: Participants will be offered the benefit of 2 points of course credit in undergraduate psychology (equivalent to 1 hour research).

Confidentiality: Your identity will be kept confidential. Your information will be assigned a participant number. The list connecting your name to this number will not be released to anyone who is not directly involved in conducting this study. Your name will not be used in any report.

Voluntary participation: Your participation in this study is voluntary. There is no penalty for not participating. You have the right to withdraw from the study at any time without penalty.

Whom to contact if you have questions about the study: Dr James L. Szalma, Department of Psychology, University of Central Florida, Orlando, FL. Telephone (407) 823-0920, email jszalma@mail.ucf.edu.

Whom to contact about your rights in the study: Research at the University of Central Florida involving human participants is carried out under the oversight of the Institutional Review Board (UCF). For information about participants' rights please contact: Institutional Review Board, University of Central Florida, Office of Research & Commercialization, 12201 Research Parkway, Suite 501, Orlando, FL 32826-3246 or by telephone at (407) 823-2901.

___ I have read the procedure described above.

___ I voluntarily agree to participate in the procedure.

_____ /_____

Participant Date

_____ /_____

Principle Investigator Date

APPENDIX C: SCRIPT TO EXPERIMENT 1

Experimenter Script – Please Read Italic Sections Aloud to Participant

The goal of this study is to examine how interface design impacts one's interaction with a distributed robot, in this case an unmanned ground vehicle (commonly referred to as a UGV – these are typically similar to remote control cars that are equipped with special equipment such as webcams). UGVs are frequently used when the environment is too dangerous for a human operator.

*In the following simulation you are operating under the premise that a group of terrorist have released a dangerous chemical into a commercial office building, and we are sending in a reconnaissance UGV to ascertain the location of terrorists, improvised explosive devices (IEDs – basically a bomb), and unconscious civilians before reinforcements arrive. You can see examples of these objects at the top of your screen (*POINT OUT IED – PARTICIPANTS HAVE TROUBLE FINDING THIS*).*

We need you to monitor the video feed from the UGV and for each room report whether you detect the presence of a terrorist, IED, civilian, or if that the room is clear. Due to time constraints the UGV must automatically move through the building as quickly as possible, you will not be controlling the movement of the UGV, thus you will have only one chance to view each room.

After the robot has sent each signal, it will conserve battery by turning off the video while it is moving to the next room, during this short time period the response keys (POINT OUT RESPONSE KEYS – COMPARE ACTIVATED PRACTICE KEY TO DEACTIVATED RESPONSE KEYS) *will be activated and you can report your observation. You will not be able to change your answer after pressing a key. Additionally you may notice that after you respond it may take several seconds to move on to the next video, this is perfectly normal. Try to respond as quickly and accurately as possible.*

You will find a pair of headphones to the left on the monitor, please wear these during the experiment. No sound will come out of the headphones, they are merely meant to attenuate any extraneous noise.

Now the most important item I will mention is that at several points in the experiment a message will pop-up stating, "Please complete the form and press OK when you are ready to continue." When this message appears do NOT immediately click OK. I will give you a questionnaire; after you COMPLETE the questionnaire you may then click OK to resume the simulation.

Do you have any questions or concerns at this point?

APPENDIX D : BLOCK QUESTIONNAIRE TO EXPERIMENT 2

Participant #: _____
Experimenter: _____
Date: _____

1. Did you feel that you had enough time to view each video clip (with 5 being neither too much nor too little time)?

Definitely Not Enough Time 0 1 2 3 4 5 6 7 8 9 10 Definitely Too Much Time

2. Did you feel that you had enough time to respond to each video clip (with 5 being neither too much nor too little time)?

Definitely Not Enough Time 0 1 2 3 4 5 6 7 8 9 10 Definitely Too Much Time

3. Do you believe you would have been able to monitor **two** video feeds at the same time?

Definitely not 0 1 2 3 4 5 6 7 8 9 10 Definitely yes

4. Do you believe you would have been able to respond to **two** video feeds at the same time?

Definitely not 0 1 2 3 4 5 6 7 8 9 10 Definitely yes

5. Do you believe you would have been able to monitor **four** video feeds at the same time?

Definitely not 0 1 2 3 4 5 6 7 8 9 10 Definitely yes

6. Do you believe you would have been able to respond to **four** video feeds at the same time?

Definitely not 0 1 2 3 4 5 6 7 8 9 10 Definitely yes

7. Please rate the MENTAL DEMAND of the task: How much mental and perceptual activity was required?

Low 0 1 2 3 4 5 6 7 8 9 10 High

8. Please rate the PHYSICAL DEMAND of the task: How much physical activity was required?

Low 0 1 2 3 4 5 6 7 8 9 10 High

9. Please rate the TEMPORAL DEMAND of the task: How much time pressure did you feel due to the pace at which the task elements occurred?

Low 0 1 2 3 4 5 6 7 8 9 10 High

10. Please rate your PERFORMANCE: How successful do you think you were in accomplishing the goals of the task?

Low 0 1 2 3 4 5 6 7 8 9 10 High

11. Please rate your EFFORT: How hard did you have to work (mentally and physically) to accomplish your level of performance?

Low 0 1 2 3 4 5 6 7 8 9 10 High

12. Please rate your FRUSTRATION: How discouraged, irritated, stressed and annoyed did you feel during the task?

Low 0 1 2 3 4 5 6 7 8 9 10 High

APPENDIX E: ERRORS IN EXPERIMENT 1 ACROSS CONDITIONS

Group Condition	Video with Errors (parenthesis contain # participants missing the video)	Total Errors
1	G1C3.avi (1), G1I2.avi(7), G1N2.avi(1), G1T2.avi(1)	10
2	G2C3.avi(1), G2I1.avi(7),G2I2.avi(4), G2I3(3), G2N2.avi(1)	16
3	G3I2.avi(2), G3I3.avi(14), G3N2.avi(1), G3N3.avi(1)	18
4	G4C1.avi(7), G4C2.avi(1), G4I1(3), G4I3.avi(2), G4N2.avi(1)	14
5	G5C1.avi(1), G5I1.avi(1), G5I3.avi(3), G5N1.avi(1), G5N2.avi(4)	10
6	G6C2.avi(4), G6I1.avi(9), G6I3.avi(3), G6N1.avi(1), G6N3.avi(10)	27
7	G7I2.avi(10), G7I3(1), G7N2.avi(2), G7N3.avi(1)	14
8	G8I2.avi(3), G8N2.avi(1)	4
9	G9C3.avi(22), G9I1.avi(7), G9I2.avi(7), G9I3.avi(1), G9N1.avi(5), G9N2.avi(4), G9N3.avi(1)	47

Note – File names of video files are written so that the first two characters reflect the condition (e.g., G1 equals group 1, G2 group 2 and so forth). The third letter represents rather the clip presented a T for terrorist, C for civilian, I for IED, or N for nothing. The fourth, and final letter, represented which of the three clips of each stimuli type was presented (e.g., the three civilian videos for group one were named G1C1.avi, G1C2.avi, and G1C3.avi), the numbers were assigned only for organizational reasons only. The .avi simply is the file extension for the video files which were in AVI format.

APPENDIX F: DEMOGRAPHIC QUESTIONNAIRE

Participant Number:_____

Date:_____

Experimenter:_____

Condition:_____

Demographic Questionnaire

1. What is your sex? (circle one) Male Female

2. What is your age? _____

3. How many hours do you work on a computer per day? (circle one)

   0 <1 hour 1-2 hours 3-4 hours 5-6 hours 7+ hours

4. How many hours a day do you play video games on average? (circle one)

   0 <1 hour 1-2 hours 3-4 hours 5-6 hours 7+ hours

   IF YOU DO PLAY VIDEO GAMES, please describe what type:

   _____

   _____

5. Are you are have you ever been involved in a search-and-rescue operation? (circle one)

   Yes No

   IF YES, please describe:

   _____

   _____

6. Are you familiar with any Unmanned/Uninhabited Vehicle (UV) system?

   Yes No

   IF YES, please describe your experience:

   _____

   _____

7. Do you have normal or corrected to normal vision and hearing?

   Yes No

   IF NO, please describe:_____

8. You have just opened an airport. As part of your responsibility for running an airport you have to ensure that proper baggage screening procedures are in place to make sure that no illegal devices are allowed onto aircraft. You have two choices for how to screen bags. Company A sells an object recognition computer program that screens bags for illegal devices. Company B trains human operators to screen bags for illegal devices. Assuming that cost is not an issue, which service do you trust to do the task better?

   Company A: Computer Company B: Human Equal

APPENDIX G: SCRIPT TO EXPERIMENT 2

Experimenter Script – Please Read Italic Sections Aloud to Participant

The goal of this study is to examine how interface design impacts one's interaction with a distributed robot, in this case an unmanned ground vehicle (commonly referred to as a UGV – these are typically similar to remote control cars that are equipped with special equipment such as webcams). UGVs are frequently used when the environment is too dangerous for a human operator.

*In the following simulation you are operating under the premise that a group of terrorist have released a dangerous chemical into a commercial office building, and we are sending in a reconnaissance UGV to ascertain the location of terrorists, improvised explosive devices (IEDs – basically a bomb), and unconscious civilians before reinforcements arrive. You can see examples of these objects at the top of your screen (*POINT OUT IED – PARTICIPANTS HAVE TROUBLE FINDING THIS*).*

We need you to monitor the video feed from the UGV and for each room report whether you detect the presence of a terrorist, IED, civilian, or if that the room is clear. Due to time constraints the UGV must automatically move through the building as quickly as possible, you will not be controlling the movement of the UGV, thus you will have only one chance to view each room.

After the robot has sent each signal, it will conserve battery by turning off the video while it is moving to the next room, during this short time period the response keys (POINT OUT RESPONSE KEYS – COMPARE ACTIVATED PRACTICE KEY TO DEACTIVATED RESPONSE KEYS) *will be activated and you can report your observation. You will not be able to change your answer after pressing a key. Additionally you may notice that after you respond it may take several seconds to move on to the next video, this is perfectly normal. Try to respond as quickly and accurately as possible.*

You will find a pair of headphones to the left on the monitor, please wear these during the experiment. No sound will come out of the headphones, they are merely meant to attenuate any extraneous noise.

Do you have any questions or concerns at this point?

APPENDIX H: ITEM DIFFICULTIES FOR EXPERIMENT 2

Civilian	Dif Index	IED	Dif Index	Terrorist	Dif Index
C01.avi	98.46	I01.avi	63.08	T01.avi	96.92
C02.avi	96.92	I02.avi	75.38	T02.avi	100.00
C03.avi	100.00	I03.avi	69.23	T03.avi	100.00
C04.avi	90.77	I04.avi	90.77	T04.avi	100.00
C05.avi	89.23	I05.avi	89.23	T05.avi	100.00
C06.avi	96.92	I06.avi	69.23	T06.avi	96.92
C07.avi	98.46	I07.avi	98.46	T07.avi	98.46
C08.avi	98.46	I08.avi	84.62	T08.avi	100.00
C09.avi	100.00	I09.avi	80.00	T09.avi	100.00
C10.avi	89.23	I10.avi	100.00	T10.avi	95.38
C11.avi	98.46	I11.avi	49.23	T11.avi	96.92
C12.avi	98.46	I12.avi	93.85	T12.avi	98.46
C13.avi	95.38	I13.avi	29.23	T13.avi	96.92
C14.avi	100.00	I14.avi	87.69	T14.avi	98.46
C15.avi	98.46	I15.avi	93.85	T15.avi	96.92
C16.avi	96.92	I16.avi	64.62	T16.avi	95.38
C17.avi	100.00	I17.avi	92.31	T17.avi	98.46
C18.avi	95.38	I18.avi	86.15	T18.avi	96.92
C19.avi	100.00	I19.avi	92.31	T19.avi	98.46
C20.avi	96.92	I20.avi	73.85	T20.avi	98.46
C21.avi	96.92	I21.avi	76.92	T21.avi	98.46
C22.avi	90.77	I22.avi	86.15	T22.avi	96.92
C23.avi	95.38	I23.avi	47.69	T23.avi	98.46
C24.avi	93.85	I24.avi	64.62	T24.avi	98.46
C25.avi	81.54	I25.avi	98.46	T25.avi	100.00
C26.avi	56.92	I26.avi	81.54	T26.avi	98.46
C27.avi	87.69	I27.avi	86.15	T27.avi	98.46
C28.avi	87.69	I28.avi	78.46	T28.avi	96.92
C29.avi	98.46	I29.avi	64.62	T29.avi	96.92
C30.avi	89.23	I30.avi	93.85	T30.avi	100.00
C31.avi	98.46	I31.avi	78.46	T31.avi	96.92
C32.avi	98.46	I32.avi	75.38	T32.avi	98.46
C33.avi	98.46	I33.avi	30.77	T33.avi	98.46

Civilian	Dif Index	IED	Dif Index	Terrorist	Dif Index
C34.avi	95.38	I34.avi	63.08	T34.avi	100.00
C35.avi	96.92	I35.avi	52.31	T35.avi	96.92
C36.avi	60.00	I36.avi	53.85	T36.avi	98.46
C37.avi	81.54	I37.avi	66.15	T37.avi	100.00
C38.avi	92.31	I38.avi	67.69	T38.avi	100.00
C39.avi	98.46	I39.avi	27.69	T39.avi	100.00
C40.avi	98.46	I40.avi	47.69	T40.avi	96.92
C41.avi	95.38	I41.avi	33.85	T41.avi	100.00
C42.avi	96.92	I42.avi	50.77	T42.avi	98.46
C43.avi	98.46	I43.avi	76.92	T43.avi	98.46
C44.avi	56.92	I44.avi	66.15	T44.avi	100.00
C45.avi	63.08	I45.avi	92.31	T45.avi	96.92
C46.avi	98.46	I46.avi	92.31	T46.avi	100.00
C47.avi	98.46	I47.avi	87.69	T47.avi	96.92
C48.avi	73.85	I48.avi	75.38	T48.avi	98.46
C49.avi	80.00	I49.avi	60.00	T49.avi	100.00
C50.avi	98.46	I50.avi	64.62	T50.avi	98.46
C51.avi	84.62	I51.avi	32.31	T51.avi	98.46
C52.avi	16.92	I52.avi	83.08	T52.avi	98.46
C53.avi	67.69	I53.avi	83.08	T53.avi	96.92
C54.avi	100.00	I54.avi	69.23	T54.avi	96.92
C55.avi	87.69	I55.avi	84.62	T55.avi	98.46
C56.avi	73.85	I56.avi	66.15	T56.avi	100.00
C57.avi	89.23	I57.avi	75.38	T57.avi	95.38
C58.avi	87.69	I58.avi	84.62	T58.avi	95.38
C59.avi	89.23	I59.avi	73.85	T59.avi	98.46
C60.avi	95.38	I60.avi	80.00	T60.avi	98.46
C61.avi	98.46	I61.avi	84.62	T61.avi	98.46
C62.avi	98.46	I62.avi	87.69	T62.avi	98.46
C63.avi	100.00	I63.avi	83.08	T63.avi	95.38
C64.avi	86.15	I64.avi	98.46	T64.avi	100.00
C65.avi	93.85	I65.avi	92.31	T65.avi	96.92
C66.avi	100.00	I66.avi	90.77	T66.avi	98.46
C67.avi	96.92	I67.avi	83.08	T67.avi	100.00

Civilian	Dif Index	IED	Dif Index	Terrorist	Dif Index
C68.avi	93.85	I68.avi	61.54	T68.avi	95.38
C69.avi	80.00	I69.avi	87.69	T69.avi	98.46
C70.avi	83.08	I70.avi	61.54	T70.avi	100.00
C71.avi	49.23	I71.avi	98.46	T71.avi	100.00
C72.avi	100.00	I72.avi	98.46	T72.avi	96.92
C73.avi	89.23	I73.avi	96.92	T73.avi	95.38
C74.avi	36.92	I74.avi	89.23	T74.avi	100.00
C75.avi	41.54	I75.avi	92.31	T75.avi	93.85

APPENDIX I: INFORMED CONSENT TO EXPERIMENT 3

Informed Consent Form

Please read this consent document carefully before you decide to participate in this study.

You must be 18 years of age or older to participate.

Project title: Empirical Examination of Trust in Automation across Multiple Agents in a Search and Rescue Operation.

Purpose of the research study: The purpose of this data collection effort is to determine the impact of using automated decision aids in a search-and-rescue scenario. The current effort seeks to determine under what conditions automated decision aids increases or decreases reliance upon these decision aids.

What you will be asked to do in the study: You will be asked to view a computer display running a simulated scenario of a search-and-rescue scenario using either one or two unmanned ground vehicles (UGVs). You will be asked to monitor the video images from these UGVs for critical signals (e.g., enemy units, civilians, or weapons). You may receive automated decision aids while completing this task. At the end of the session you will be asked to complete several questionnaires about your experience performing the search-and-rescue scenario.

Time required: Approximately thirty minutes (0.5 hour).

Risks: Minimal. The risks to you are no greater than operating any other computer.

Benefits/Compensation: Participants will be offered the benefit of 1 point of course credit in undergraduate psychology (equivalent to 30 minutes research).

Confidentiality: Your identity will be kept confidential. Your information will be assigned a participant number. The list connecting your name to this number will not be released to anyone who is not directly involved in conducting this study. Your name will not be used in any report.

Voluntary participation: Your participation in this study is voluntary. There is no penalty for not participating. You have the right to withdraw from the study at any time without penalty.

Whom to contact if you have questions about the study: Dr James L. Szalma, Department of Psychology, University of Central Florida, Orlando, FL. Telephone (407) 823-0920, email jszalma@mail.ucf.edu.

Whom to contact about your rights in the study: Research at the University of Central Florida involving human participants is carried out under the oversight of the Institutional Review Board (UCF). For information about participants' rights please contact: Institutional Review Board, University of Central Florida, Office of Research & Commercialization, 12201 Research Parkway, Suite 501, Orlando, FL 32826-3246 or by telephone at (407) 823-2901.

___ I have read the procedure described above.

___ I voluntarily agree to participate in the procedure.

_____ /_____

Participant Date

_____ /_____

Principle Investigator Date

APPENDIX J : SCRIPT TO EXPERIMENT 3

Experimenter Script – Please Read Italic Sections Aloud to Participant

Welcome everyone, thank you for coming in today. Before we begin, please note that the task we will be involved in today is a *simulation*. We are conducting a scientific experiment which seeks to better understand how people interact with automated agents like unmanned ground vehicles. In order to obtain accurate results, we need to mimic a real-world situation as closely as possible. Therefore, today we will be working under the scenario that a terrorist organization has infiltrated a commercial office building somewhere in the United States. However, before we begin, I must again stress the fact that this is a *simulation*: there has been no real terrorist attack, nor is anyone's life truly in danger.

Are there any questions at this point?

We will now begin our background briefing. Please open your information packets to the first page. The person you see here is Augustus Sol Invictus, the merciless leader of the Invictus Terror Organization (or ITO), an extremist group bent on the destruction of the free world. There are no known photographs of his face; he, along with all of the members of the ITO are rarely seen, and when they are seen they always wear the black mask and uniform you see in the photograph, making our estimations of their numbers highly unreliable. We know very little of the ITO, other than that they are unpredictable, and very dangerous. This morning, the ITO infiltrated a commercial office building occupied by more than a hundred U.S. civilians. There is evidence that they released a gaseous chemical agent throughout the building. If you turn to page two of your information packet you can see an aerial surveillance photo of the building.

Preliminary intelligence reports indicate that there are probably civilians in the building who are still alive, but may have been rendered unconscious by the gaseous chemical agent. However, it is unclear how many civilians there are or where they are located within the building. Several of our own military forces have managed to covertly gain access into the building. They have reported seeing a number of IED's (or improvised explosive devices, which are basically bombs) placed throughout the building. We have received an image of the type of IED they have found, it is shown on page three of your information packet. Military intelligence estimates a high likelihood that a full assault on the building might lead to the detonation of the explosives after American forces have entered the building in order to maximize the number of casualties. Battalion headquarters has decided to deploy an unmanned ground vehicle (UGV) to identify the locations of IEDs, terrorist suspects, as well as the locations of any unconscious civilians. The UGV will patrol the building and transmit a video feed of each room. An example of the UGV is shown on page four of your information packet. This is where you come in.

At this point, please turn to page five in your information packet to read your

instructions. I will read these out loud and you should read along silently.

Your job will be to monitor the video feed from the UGV as it patrols through each room in the building. The navigation of the UGV is fully automated and does not require any control for its movement; that is, it has been programmed to move from room to room on its own.

The UGV will scan through each room, one at a time, and transmit the video feed to you. Your mission is to monitor the video feeds sent in by the UGV and report what is in each room. You may report your observation by selecting one of the response buttons below the video player while the UGV moves onto the next room.

Although it takes the UGV several seconds to move from one room to the next, we ask that you still respond as quickly as possible. If you do not respond by the time the UGV has moved on to the next room it will not wait for you, it will automatically begin presenting the next video feed and you will not have a chance to go back.

To prevent detection of the outgoing video feed by the Invictus Terror Organization, the UGV has been programmed to randomly switch the frequency at which it transmits its video feed. This is done randomly and may result in some video clips being presented at clearer frequencies than others.

To aid you in your mission, the UGV may be equipped with an automated object recognition system that allows it to recognize objects (e.g., IEDs). Please keep in mind that when the system is engaged, it will provide a recommendation that it believes is correct, but it is still ultimately your decision which response to select.

When you have completed the mission you may open the Exit Questionnaire envelope at your desk. Please open this only after you have completed the mission. When you have finished the questionnaire please come to the front of the room and you will be debriefed and receive your compensation (cash payment or extra credit).

To help you focus on the task we ask that, during the mission, you wear the noise canceling headphones, located to the left of the monitor. No noise will come out of the headphones, they are used solely to block out environmental noise.

Do you have any questions regarding your mission?

At this point you are ready to begin your training. Please close the information packet and place it somewhere where you can see the example pictures shown on the back. The training scenario will give you a chance to see how the task will work, and show you examples of the objects you will need to watch for. Please click the “Practice” button to begin the training. The first practice clip will contain a terrorist, press OK to view the clip, then when you are prompted to respond, click the “Terrorist” button below the video player.

(After participants have identified the terrorist, instruct them what to identify in the next

three clips as the videos are playing; first IED, then Civilian, then Empty)

The next four practice clips will show you what the automated object recognition system looks like. It will make a recommendation, but is still up to you to make the final selection (by clicking one of the four response buttons or the agree aid button). Use of the aid is completely optional and the responsibility of the final decision is your own and you can choose to either accept the aid's proposed diagnosis or to ignore it.

(Wait until participants identify all 4 clips, then continue).

The practice session is now complete. Does anyone have any questions about the task?

You may click the OK button to exit the training. The remainder of the experiment will be self-paced and I will not give you any more instructions. When you have competed the study fill out the questionairre about your experiences and come up to the front and I will compensate you for your participation. Does anyone have any final questions before we get started?

Please put your headphones on. They are adjustable, so take a minute to make them as comfortable as possible. I will come around to turn them on, once I have turned your headphones on you may begin the mission.

(Turn on all headphones)

APPENDIX K : PARTICIPANT FOLDER

Operation Silent Snake

Highly Sensitive Material

Do not open

until instructed

Operation Silent Snake

Figure 1: Augustus Sol Invictus, leader of the Invictus Terror Organization

Operation Silent Snake

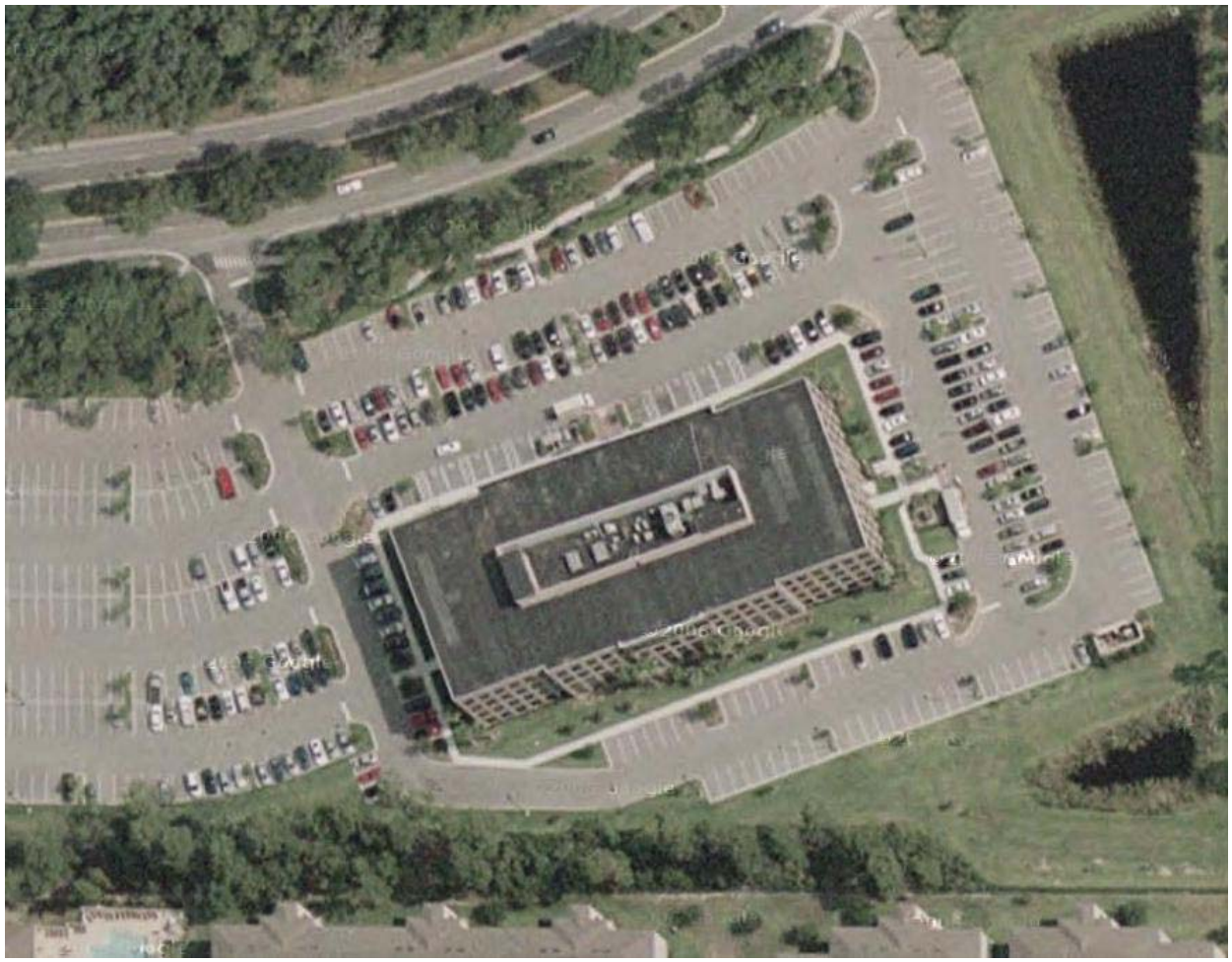


Figure 2: The attack site

Figure 3: Improvised Explosive Device (IED) found in the building

Operation Silent Snake

Figure 4: The tread driven all-terrain autonomous surveillance robot

Operation Silent Snake

Your Mission

Your job will be to monitor the video feed from the UGV as it patrols through each room in the building. The navigation of the UGV is fully automated and does not require any control for its movement; that is, it has been programmed to move from room to room on its own.

The UGV will scan through each room, one at a time, and transmit the video feed to you. Your mission is to monitor the video feeds sent in by the UGV and report what is in each room. You may report your observation by selecting one of the response buttons below the video player while the UGV moves onto the next room.

Although it takes the UGV several seconds to move from one room to the next, we ask that you still respond as quickly as possible. If you do not respond by the time the UGV has moved on to the next room it will not wait for you, it will automatically begin presenting the next video feed and you will not have a chance to go back.

To prevent detection of the outgoing video feed by the Invictus Terror Organization, the UGV has been programmed to randomly switch the frequency at which it transmits its video feed. This is done randomly and may result in some video clips being presented at clearer frequencies than others.

To aid you in your mission, the UGV may be equipped with an automated object recognition system that allows it to recognize objects (e.g., IEDs). Please keep in mind that when the system is engaged, it will provide a recommendation that it believes is correct, but it is still ultimately your decision which response to select.

When you have completed the mission you may open the Exit Questionnaire envelope at your desk. Please open this only after you have completed the mission. When you have finished the questionnaire please come to the front of the room and you will be debriefed and receive your compensation (cash payment or extra credit).

To help you focus on the task we ask that, during the mission, you wear the noise canceling headphones, located to the left of the monitor. No noise will come out of the headphones, they are used solely to block out environmental noise.

Do you have any questions regarding your mission?

APPENDIX L: EXIT TRUST QUESTIONNAIRE FOR EXPERIMENT 3

Participant #: _____

1. To what extent does the agent perform this search-and-rescue task effectively?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

2. To what extent can you anticipate the agent's behavior with some degree of confidence?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

3. To what extent is the agent free of errors?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

4. To what extent do you have a strong belief and trust in the agent to do the search-and-rescue task in the future without being monitored?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

5. How much did you trust the decisions of the agent overall?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

6. What percentage of responses by the agent do you think were correct?

_____(enter a value between 0% to 100%)

7. How often did you notice an error made by the contrast detector??

Not At All 1 2 3 4 5 6 7 8 9 Many Times

8. To what extent did you lose trust in the contrast detector when you noticed it made an error?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

9. **Hypothetical Scenario:** Imagine that there are ten more video clips that need to be examined for terrorists, civilians, and IEDs. Also imagine that we were to offer you an additional compensation, of either $5.00 or an extra credit point for *each* of these ten additional video clips that is correctly identified. However, due to a software problem only you *or* the aid can make the decisions. Would you prefer that this additional compensation be based on the decisions made by the automated aid or the decisions made by you? (circle one)

Automated Aid Decisions My Own Decisions

10. We would like to know what led to your decision to base your performance on either your decisions or on the decisions of the aid. Please tell us everything you thought of in coming to this decision. Do not worry about spelling or grammatical errors. Please ask the experimenter for additional paper if necessary.

APPENDIX M: INFOMED CONSENT FOR EXPERIMENT 4

Informed Consent Form
Please read this consent document carefully before you decide to participate in this study.
You must be 18 years of age or older to participate.

Project Title: Robot Search-and-Rescue Study

Purpose of the Research Study: The purpose of this data collection effort is to determine the impact of using automated decision aids in a search-and-rescue scenario.

What you will be asked to do in the Study: You will be asked to view a computer display running a simulated search-and-rescue scenario using two unmanned ground vehicles (UGVs). You will be asked to monitor the video images from the UGV for critical signals (e.g., enemy units, unconscious civilians, or improvised explosive devices). You may, or may not, receive recommendations while completing this task. The study will take approximately 1 hour. At the end of the study you will be asked to complete a brief questionnaire about your experience.

Time Required: 60 minutes (1 hour).

Risks: Minimal. The risks to you are no greater than operating any other computer.

Benefits/Compensation: Participants will be offered the benefit of 2 point of course credit in undergraduate psychology (equivalent to 1 hour research) or $8.00 US paid compensation.

Confidentiality: Your identity will be anonymous. Your information will be assigned a participant number. The list connecting your name will not be released to anyone who is not directly involved in conducting this study. Your name will not be used in any report.

Voluntary Participation: Your participation in this study is voluntary. There is no penalty for not participating. You have the right to withdraw from the study at any time without penalty.

Whom to Contact if you have Questions about the Study: Jennifer Ross, Graduate Research Fellow, University of Central Florida, phone: 407-687-4435, e-mail: jmross@mail.ucf.edu, or Dr. James L. Szalma, Department of Psychology, University of Central Florida, phone: 407-823-2901, e-mail: jszalma@mail.ucf.edu.

Whom to Contact about your rights in the study: Research at the University of Central Florida involving human participants is carried out under the oversight of the Institutional Review Board. Questions or concerns about research participants' rights may be directed to the UCF IRB office, University of Central Florida, Office of Research & Commercialization, 12201 Research Parkway, Suite 501, Orlando, FL 32826-3246, or by campus mail 32816-0150. The hours of operation are 8:00 am until 5:00 pm, Monday through Friday except on University of Central Florida official holidays. The telephone numbers are (407) 882-2276 and (407) 823-2901.

I have read the procedure described above. I voluntarily agree to participate in the procedure.

- ☐ I elect to receive 2pts course credit for the course of my choosing through Sona System.
- ☐ I elect to receive $8 hour for my participation.

_____

Participant Name Printed

_____/_____

Participant Signature Date

APPENDIX N: ANTHROPOMORPHIC TENDENCIES SCALE

ATS

Please read each statement carefully. Indicate the strength of your agreement with each statement by filling in the blank using the following 5-point scale. There are no right or wrong answers to any of these statements. We are interested in your honest reactions and opinions.

1	2	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

_____ 1. I would yell at a COMPUTER if it did something I did not like.

_____ 2. I would not praise a GOD OR HIGHER POWER when it does something I like.

_____ 3. A GOD OR HIGHER POWER does not have a personality like a person has a personality.

_____ 4. I would hit a CAR if it did something I did not like.

_____ 5. A GOD OR HIGHER POWER has a spirit or life-force like people do.

_____ 6. I would hit a BACKPACK if it did something I did not like.

_____ 7. A GOD OR HIGHER POWER cannot communicate with people.

_____ 8. I would not praise a PET when it does something I like.

_____ 9. I would hit a MICROWAVE if it did something I did not like.

_____ 10. When I am clearly upset, a GOD OR HIGHER POWER does not know.

_____ 11. A BACKPACK does not have a personality like a person has a personality.

_____ 12. I do not act as if a GOD OR HIGHER POWER has a spirit or life-force like people do.

_____ 13. When I talk to a PET, I do not believe it understands me.

_____ 14. I would yell at a CAR if it did something I did not like.

_____ 15. When I am clearly upset, an OCEAN does not know.

_____ 16. A GOD OR HIGHER POWER is intelligent like a human is intelligent.

_____ 17. If I were to get rid of a BACKPACK, it would feel abandoned.

_____ 18. When I talk to a GOD OR HIGHER POWER, I do not believe it understands me.

_____ 19. I would hit a COMPUTER if it did something I did not like.

_____ 20. A PET has a spirit or life-force like people do.

_____ 21. I treat a BACKPACK like a human.

_____ 22. I would apologize to a GOD OR HIGHER POWER for accidentally hurting it.

_____ 23. I would talk to a CAR.

_____ 24. A PET does not have a personality like a person has a personality.

_____ 25. I would talk to a COMPUTER.

_____ 26. I would apologize to a PET for accidentally hurting it.

_____ 27. A PET is intelligent like a human is intelligent.

_____ 28. When I am clearly upset, a CAR does not know.

_____ 29. A CAR has a spirit or life-force like people do.

_____ 30. When I am clearly upset, a PET does not know.

_____ 31. I do not act as if a STOMACH has a spirit or life-force like people do.

_____ 32. A PET likes certain people better than others.

_____ 33. A PET cannot communicate with people.

	1	2	3	4	5
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

_____ 34. I would not buy a present for a PET.

_____ 35. I do not act as if a MICROWAVE has a spirit or life-force like people do.

_____ 36. A COMPUTER does not do things just to annoy me.

_____ 37. I would not apologize to a GOD OR HIGHER POWER for neglecting it.

_____ 38. If I were to get rid of a COMPUTER, it would feel abandoned.

_____ 39. I would not praise a HOUSE PLANT when it does something I like.

_____ 40. A MICROWAVE has a spirit or life-force like people do.

_____ 41. A MICROWAVE is intelligent like a human is intelligent.

_____ 42. When I am clearly upset, a COMPUTER does not know.

_____ 43. If a PET were to be destroyed, I would not mourn it like I would mourn the loss of a human.

_____ 44. I do not act as if a COMPUTER has a spirit or life-force like people do.

_____ 45. A COMPUTER does not have a personality like a person has a personality.

_____ 46. A STUFFED TOY is intelligent like a human is intelligent.

_____ 47. I would not buy a present for a HOUSE PLANT.

_____ 48. A MICROWAVE likes certain people better than others.

_____ 49. LUCK is intelligent like a human is intelligent.

_____ 50. I treat an INSECT like a human.

_____ 51. A STUFFED TOY does not have a personality like a person has a personality.

_____ 52. When I am clearly upset, a MICROWAVE does not know.

_____ 53. I would not praise a MICROWAVE when it does something I like.

_____ 54. A STUFFED TOY cannot communicate with people.

_____ 55. I would talk to a GOD OR HIGHER POWER.

_____ 56. I would not apologize to a COMPUTER for neglecting it.

_____ 57. An OCEAN does not do things just to annoy me.

_____ 58. I do not act as if an OCEAN has a spirit or life-force like people do.

_____ 59. A STOMACH does not have a personality like a person has a personality.

_____ 60. If I were to get rid of a MICROWAVE, it would feel abandoned.

_____ 61. A COMPUTER has a spirit or life-force like people do.

_____ 62. An OCEAN does not have a personality like a person has a personality.

_____ 63. I would not apologize to a BACKPACK for neglecting it.

_____ 64. I do not act as if a CAR has a spirit or life-force like people do.

_____ 65. I treat a PET like a human.

_____ 66. I do not act as if a PET has a spirit or life-force like people do.

_____ 67. I would name a PET.

_____ 68. I treat a COMPUTER like a human.

_____ 69. I would talk to a PET.

	1	2	3	4	5
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

_____ 70. If I were to get rid of a STUFFED TOY, it would feel abandoned.

_____ 71. If I were to get rid of a PET, it would feel abandoned.

_____ 72. I treat a GOD OR HIGHER POWER like a human.

_____ 73. A MICROWAVE does <u>not</u> do things just to annoy me.

_____ 74. I do <u>not</u> act as if LUCK has a spirit or life-force like people do.

_____ 75. I would <u>not</u> buy a present for a BACKPACK.

_____ 76. If I were to get rid of a HOUSE PLANT, it would feel abandoned.

_____ 77. When I talk to a CAR, I do <u>not</u> believe it understands me.

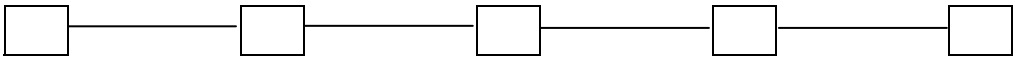
_____ 78. I treat a MICROWAVE like a human.

APPENDIX O: INTERPERSONAL TRUST SCALE

Interpersonal Trust Scale

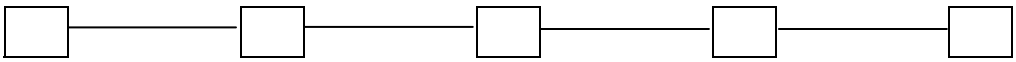
Please mark an 'X' in the box above the statement that best describes how you feel about that statement.

1. Hypocrisy is on the increase in our society.



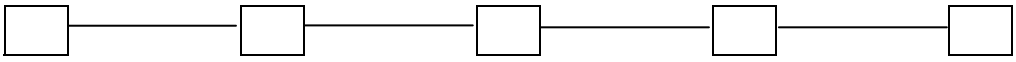
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

2. In dealing with strangers one is better off to be cautious until they have provided evidence that they are trustworthy.



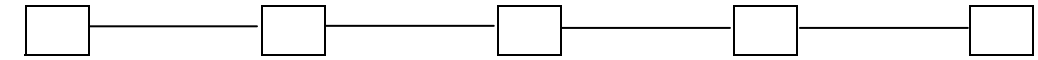
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

3. This country has a dark future unless we can attract better people into politics.



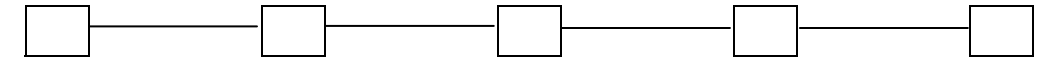
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

4. Fear and social disgrace or punishment rather than conscience prevents most people from breaking the law.



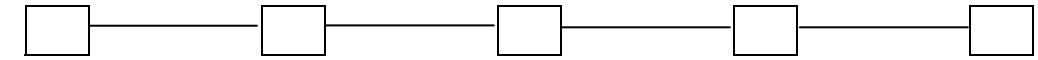
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

5. Using the honor system of *not* having a teacher present during exams would probably result in increased cheating.



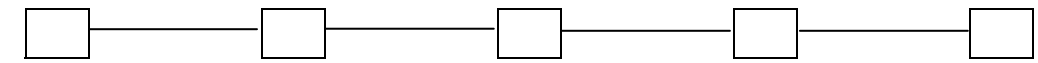
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

6. Parents usually can be relied on to keep their promises.



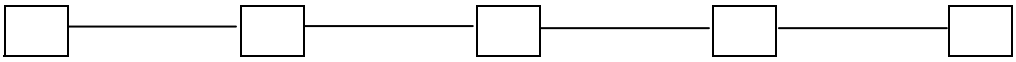
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

7. The United Nations will never be an effective force in keeping world peace.



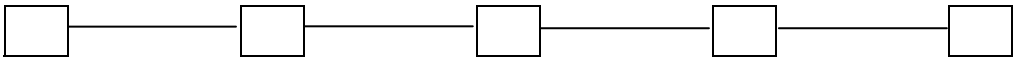
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

8. The judiciary is a place where we can all get unbiased treatment.



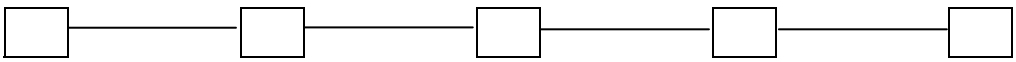
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

9. Most people would be horrified if they knew how much news that the public hears and sees is distorted.



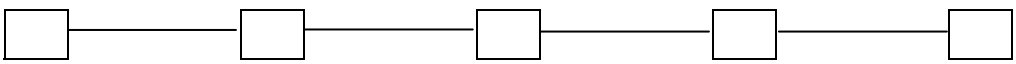
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

10. It is safe to believe that in spite of what people say most people are primarily interested in their own welfare.



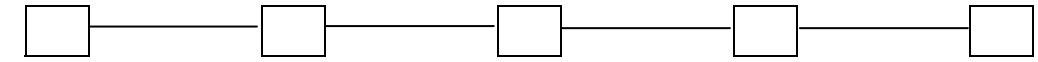
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

11. Even though we have reports in newspaper, radio, and T.V., it is hard to get objective accounts of public events.



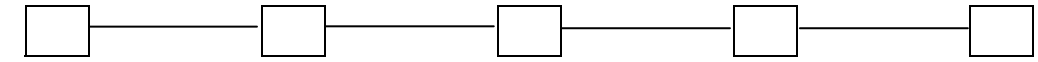
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

12. The future seems very promising.



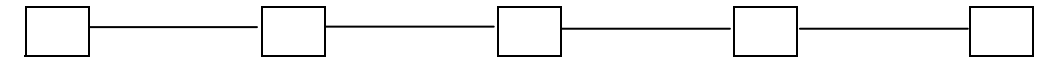
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

13. If we really knew what was going on in international politics, the public would have reason to be more frightened than they now seem to be.



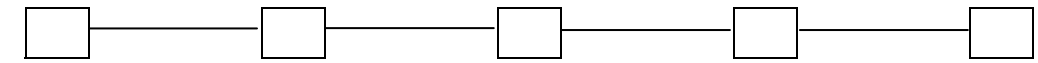
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

14. Most elected officials are really sincere in their campaign promises.



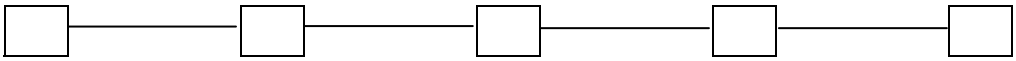
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

15. Many major national sports contests are fixed in one way or another.



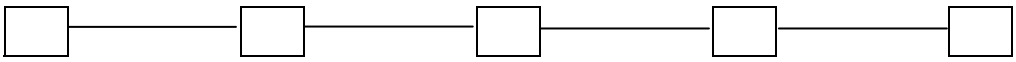
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

16. Most experts can be relied upon to tell the truth about the limits of their knowledge.



1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

17. Most parents can be relied upon to carry out their threats of punishment.

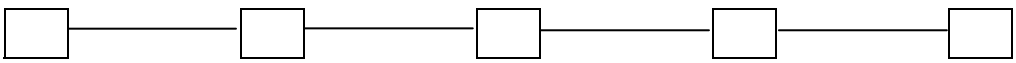


1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

18. Most people can be counted on to do what they say they will do.

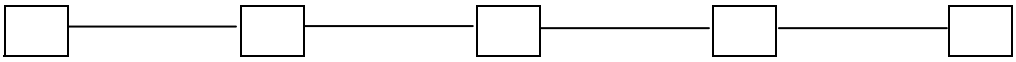
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

19. In these competitive times one has to be alert or someone is likely to take advantage of you.



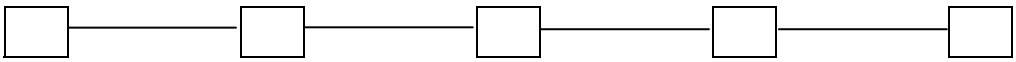
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

20. Most idealists are sincere and usually practice what they preach.



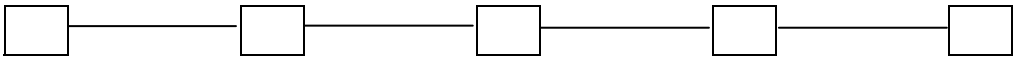
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

21. Most salesmen are honest in describing their products.



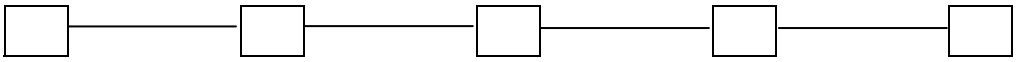
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

22. Most students in school would not cheat even if they were sure of getting away with it.



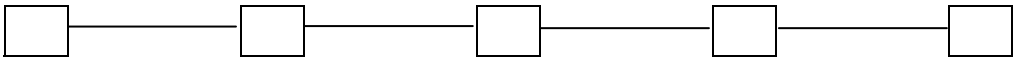
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

23. Most repairmen will not overcharge even if they think you are ignorant of their specialty.



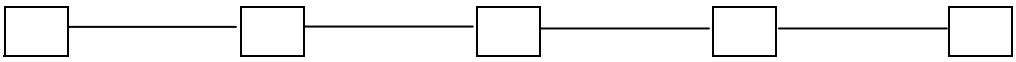
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

24. A large share of accident claims filed against insurance companies are phony.



1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

25. Most people answer public opinion polls honestly.



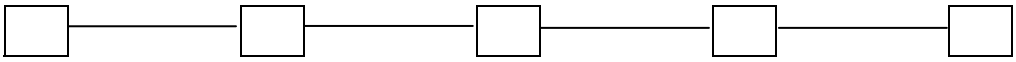
1	2	3	4	5
Strongly agree	Mildly agree	Agree and disagree equally	Mildly disagree	Strongly disagree

APPENDIX P: COMPLACENCY POTENTIAL RATING SCALE

Complacency Potential Rating Scale

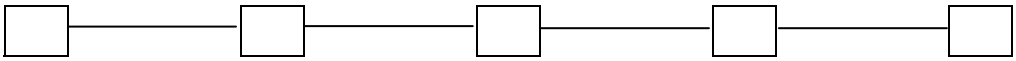
Please mark an 'X' in the box above the statement that best describes how you feel about that statement.

1. I think that automated devices used in medicine, such as CT scans and ultrasound, provide very reliable medical diagnosis.



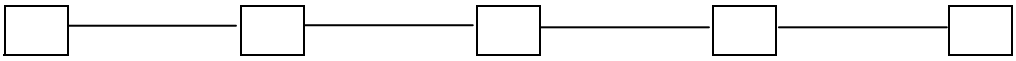
1	2	3	4	5
Strongly Disagree	Mildly Disagree	Disagree and agree equally	Mildly Agree	Strongly Agree

2. Automated devices in medicine save time and money in the diagnosis and treatment of disease.



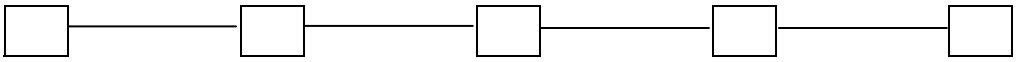
1	2	3	4	5
Strongly Disagree	Mildly Disagree	Disagree and agree equally	Mildly Agree	Strongly Agree

3. If I need to have a tumor in my body removed, I would choose to undergo computer-aided surgery using laser technology because it is more reliable and safer than manual surgery.



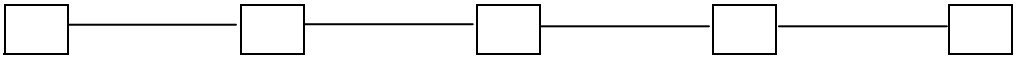
1	2	3	4	5
Strongly Disagree	Mildly Disagree	Disagree and agree equally	Mildly Agree	Strongly Agree

4. Automated systems used in modern aircraft, such as the automatic landing system, have made air journeys safer.



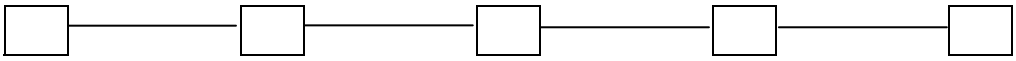
1	2	3	4	5
Strongly Disagree	Mildly Disagree	Disagree and agree equally	Mildly Agree	Strongly Agree

5. ATMs provide a safeguard against the inappropriate use of an individual's bank account by dishonest people.



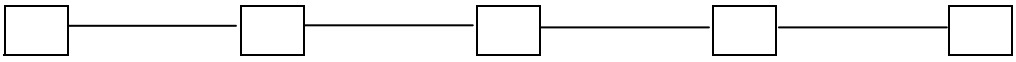
1	2	3	4	5
Strongly Disagree	Mildly Disagree	Disagree and agree equally	Mildly Agree	Strongly Agree

6. Automated devices used in aviation and banking have made work easier for both employees and customers.



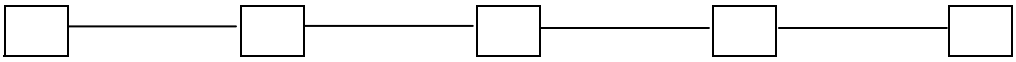
1	2	3	4	5
Strongly Disagree	Mildly Disagree	Disagree and agree equally	Mildly Agree	Strongly Agree

7. Even though the automatic cruise control in my car is set at a speed below the speed limit, I worry when I pass a police radar speed trap in case the automatic control is not working properly.



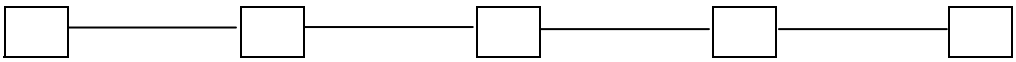
1	2	3	4	5
Strongly Disagree	Mildly Disagree	Disagree and agree equally	Mildly Agree	Strongly Agree

8. Manually sorting through card catalogues is more reliable than computer-aided searches for finding items in a library.



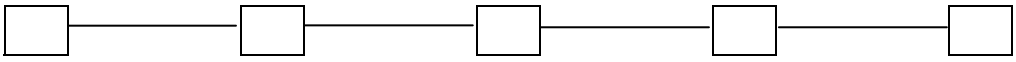
1	2	3	4	5
Strongly Disagree	Mildly Disagree	Disagree and agree equally	Mildly Agree	Strongly Agree

9. I would rather purchase an item using a computer than have to deal with a sales representative on the phone because my order is more likely to be correct using the computer.



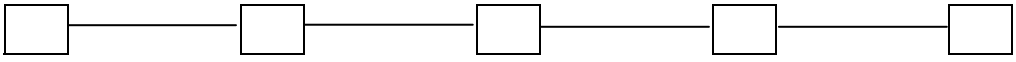
1	2	3	4	5
Strongly Disagree	Mildly Disagree	Disagree and agree equally	Mildly Agree	Strongly Agree

10. Bank transactions have become safer with the introduction of computer technology for the transfer of funds.



1	2	3	4	5
Strongly Disagree	Mildly Disagree	Disagree and agree equally	Mildly Agree	Strongly Agree

11. I feel safer depositing my money at an ATM than with a human teller.



1	2	3	4	5
Strongly Disagree	Mildly Disagree	Disagree and agree equally	Mildly Agree	Strongly Agree

12. I have to tape an important TV program for a class assignment. To ensure that the correct program is recorded, I would use the automatic programming facility on my VCR rather than manual taping.

1	2	3	4	5
Strongly Disagree	Mildly Disagree	Disagree and agree equally	Mildly Agree	Strongly Agree

APPENDIX Q: INSTRUCTIONS FOR EXPERIMENT 4

Experimenter Script – Please Read Italic Sections Aloud to Participant

The goal of this study is to examine how interface design impacts one's interaction with a distributed agent, in this case an unmanned ground vehicle (commonly referred to as a UGV – these are typically similar to remote control cars that are equipped with special equipment such as webcams). UGVs are frequently used when the environment is too dangerous for a human operator.

*In the following simulation you are operating under the premise that a group of terrorist have released a dangerous chemical into a commercial office building, and we are sending in two reconnaissance UGVs to ascertain the location of terrorists, improvised explosive devices (IEDs – basically a bomb), and unconscious civilians before reinforcements arrive. You can see examples of these objects at the top of your screen (*POINT OUT IED – PARTICIPANTS HAVE TROUBLE FINDING THIS*).*

We need you to monitor the video feeds from these two UGVs for each room. For each UGV please report whether you detect the presence of a terrorist, IED, civilian, or if that the room is clear. Due to time constraints the UGVs must move through different parts of the office building. So the videos that you see will be from two different rooms. The UGVs will automatically move through the building as quickly as possible. You will not be controlling the movement of the UGVs, thus you will have only one chance to view each of the rooms.

After the UGVs have sent each signal, it will conserve battery by turning off their video feeds while moving to the next room, during this short time period the response keys (POINT OUT RESPONSE KEYS – COMPARE ACTIVATED PRACTICE KEY TO DEACTIVATED RESPONSE KEYS) *will be activated and you can report your observation. You will not be able to change your answer after pressing a key. Additionally you may notice that after you respond it may take several seconds to move on to the next video, this is perfectly normal. Try to respond as quickly and accurately as possible.*

You will find a pair of headphones to the left on the monitor, please wear these during the experiment. No sound will come out of the headphones, they are merely meant to attenuate any extraneous noise.

*Do you have any questions or concerns at this point? (*HAVE PARTICIPANT DO FIRST 4 PRACTICE TRIALS).

Now you will have the option in this study to accept the recommendation of your teammates. Your teammates will report their recommendations (POINT OUT AID RESPONSE BOX ON SCREEN). *Following your teammates' recommendations is completely optional and the final decisions are your responsibility. You may choose to agree with your teammates or select your own response. However, be sure to respond to each trial, any trials you do not*

respond to will be counted as incorrect. Before you begin I'd like to tell you a little bit about your teammates…

SPECIAL INSTRUCTIONS (READ ONLY ONE TO EACH PARTICIPANT)

HUMAN CONDITION

Your teammates are two undergraduate students who have previously completed the experiment..

SIMILAR AUTOMATION

Your teammates are two "contrast detectors." They work by using a computer algorithm to analyze the visual scene for the target people and objects. These contrast detectors were developed to work with these specific UGV robots. The contrast detectors will not receive feedback on whether you have accepted or rejected their recommendations.

DISSIMILAR AUTOMATION

Your teammates are two different "contrast detectors." They work by using computer algorithms to analyze the visual scene for the target people and objects. A different computer algorithm was created for each of the UGV robots you will be using in this study. The contrast detectors will not receive feedback on whether you have accepted or rejected their recommendations.

*Do you have any questions or concerns at this point? (*HAVE PARTICIPANT DO LAST 4 PRACTICE TRIALS WITH AID). *Are you ready to begin the study?*

APPENDIX R: PRETRUST QUESTIONNAIRE TO EXPERIMENT 4

Participant #: _____

1. How well do you think **Teammate A** will perform during the 60 trials?

Not Very Well 1 2 3 4 5 6 7 8 9 Very Well

2. How well do you think **Teammate B** will perform the 60 trials?

Not Very Well 1 2 3 4 5 6 7 8 9 Very Well

3. How well do you think **You** will perform the 120 trials?

Not Very Well 1 2 3 4 5 6 7 8 9 Very Well

4. Who do you think will make more errors during the 120 trials? **I will make…**

Many More Errors 1 2 3 4 5 6 7 8 9 Far Fewer Errors

5. How many errors do you think **You** will make during the 120 trials?

I will make about _____ errors (numerical value b/n 0-120)

6. How many errors do you think **Teammate A** will make during the 60 trials?

Agent A will make about _____ errors (numerical value b/n 0-120)

7. How many errors do you think **Teammate B** will make during the 60 trials?

Agent B will make about _____ errors (numerical value b/n 0-120)

8. To what extent do you believe you can trust the decisions of **Teammate A**?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

9. To what extent do you believe you can trust the decisions of **Teammate B**?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

10. To what extent do you believe you can trust the decisions **You** will make?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

11. How would you rate the expected performance of **Teammate A** relative to your expected performance? Agent A will perform…

Better Than I Will Perform	1	2	3	4	5	6	7	8	9	Much Worse Than I Will Perform

12. How would you rate the expected performance of **Teammate B** relative to your expected performance? Agent A will perform…

Better Than I Will Perform	1	2	3	4	5	6	7	8	9	Much Worse Than I Will Perform

APPENDIX S: POST-SELF TRUST QUESTIONNAIRE TO EXPERIMENT 4

Participant # _____

Please answer the following questions regarding how you feel about YOUR performance only.

1. How high was your self-confidence in performing the search-and-rescue task?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

4. Please complete the computer-based questionnaire using the following definitions:

Mental Demand

How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?

Physical Demand

How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, restful or laborious?

Temporal Demand

How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?

Performance

How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

Effort

How hard did you have to work (mentally and physically) to accomplish your level of performance?

Frustration Level

How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

APPENDIX T: POST-TRUST QUESTIONNAIRES TO EXPERIMENT 4

Participant #: ____

Please answer the following questions regarding how you feel about Teammate A only.

1. To what extent does Teammate A perform this search-and-rescue task effectively?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

5. To what extent can you anticipate Teammate A's behavior with some degree of confidence?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

3. To what extent is the Teammate A free of errors?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

4. To what extent do you have a strong belief and trust in Teammate A to do the search-and-rescue task in the future without being monitored?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

5. How much did you trust the decisions of Teammate A overall?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

6. What percentage of responses by Teammate A do you think were correct?

_____ (enter a value between 0% to 100%)

7. How often did you notice an error made by Teammate A?

Not At All 1 2 3 4 5 6 7 8 9 Many Times

8. To what extent did you lose trust in Teammate A when you noticed it made an error?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

9. **Hypothetical Scenario:** Imagine that there are ten more video clips that need to be examined for terrorists, civilians, and IEDs. Also imagine that we were to offer you an additional compensation, of either $5.00 or an extra credit point for *each* of these ten additional video clips that is correctly identified. However, due to a software problem only you *or* Teammate A can make the decisions. Would you prefer that this additional compensation be based on the decisions made by the automated aid or the decisions made by you? (circle one)

Teammate A Decisions My Own Decisions

10. We would like to know what led to your decision to base your performance on either your decisions or on Teammate A's decisions. Please tell us everything you thought of in coming to this decision. Do not worry about spelling or grammatical errors. Please ask the experimenter for additional paper if necessary.

Participant #: ___

Please answer the following questions regarding how you feel about Teammate B only.

1. To what extent does Teammate B perform this search-and-rescue task effectively?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

6. To what extent can you anticipate Teammate B's behavior with some degree of confidence?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

3. To what extent is the Teammate B free of errors?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

4. To what extent do you have a strong belief and trust in Teammate B to do the search-and-rescue task in the future without being monitored?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

5. How much did you trust the decisions of Teammate B overall?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

6. What percentage of responses by Teammate B do you think were correct?

_____ (enter a value between 0% to 100%)

7. How often did you notice an error made by Teammate B?

Not At All 1 2 3 4 5 6 7 8 9 Many Times

8. To what extent did you lose trust in Teammate B when you noticed it made an error?

Very Little 1 2 3 4 5 6 7 8 9 A Great Amount

9. **Hypothetical Scenario:** Imagine that there are ten more video clips that need to be examined for terrorists, civilians, and IEDs. Also imagine that we were to offer you an additional compensation, of either $5.00 or an extra credit point for *each* of these ten additional video clips that is correctly identified. However, due to a software problem only you *or* Teammate B can make the decisions. Would you prefer that this additional compensation be based on the decisions made by the automated aid or the decisions made by you? (circle one)

Teammate B Decisions My Own Decisions

10. We would like to know what led to your decision to base your performance on either your decisions or on Teammate B's decisions. Please tell us everything you thought of in coming to this decision. Do not worry about spelling or grammatical errors. Please ask the experimenter for additional paper if necessary.

APPENDIX U: DEBRIEFING FORM TO EXPERIMENT 4

Debriefing Form Robot Search-and-Rescue Study

You have now completed the study, thank you for your participation! This form is for you to take with you and explains the purpose of our research. Please do not share this form with others who plan to participate in this study as it may bias their responses.

The purpose of this study is to examine the issue of trust in multiple teammates. That is, how one's trust in a teammate changes given the type of teammates they are interacting with and the reliability (i.e., accuracy) of the teammates. In the study you just completed, you were informed that two distributed human agents or two intelligent robotic agents provided you with recommendations after each video clip, as to what kind of signal was present in the clip. However, another human or robot did not actually provide you with recommendations in the preceding study. In order to standardize participant experience, that is to ensure that everyone had the same experience with their teammates, your teammate's responses (robotic and human) were predetermined upon the condition you were randomly assigned to. Depending on the condition you were assigned to your teammates may have both been very accurate, both very inaccurate, or a mixture (with one accurate and one inaccurate).

If you have any complaints, concerns, or questions about this research, or you would like any information about the results of the study once it is completed feel free to contact Ms. Jennifer Ross at jmross@mail.ucf.edu / 407-687-4435 or Dr. James Szalma at jszalma@mail.ucf.edu / 407-823-0920.

Your responses are confidential to the experimenters and will be published anonymously as group data.

If you are interested in obtaining more information on this topic we would recommend the following articles available through the UCF library:

- Muir, B. M., & Moray, N. Trust in automation: Part 2. Experimental studies of trust and human intervention in automated systems. *Ergonomics, 37*, (1996), 1905--1922.
- Parasuraman, R. & Riley, V. Humans and automation: Use, misuse, disuse, and abuse. *Human Factors, 39*, (1997), 230--253.

Finally, thank you again for helping us with this research.

APPENDIX V: OVERALL MEANS AND STANDARD DEVIATIONS ACROSS ALL CONDITIONS FOR STUDY 4

		Uniform-High	Mixed	Uniform-Low	Mixed-High	Mixed-Low
Human	Trust	6.65 (1.83)	6.76 (0.94)	5.41 (1.42)	7.21 (1.29)	6.3 (1.57)
	Reliance	83.89 (5.62)	78.51 (5.46)	75.45 (5.19)	81.91 (6.35)	75.03 (5.76)
	Performance	82.88 (4.82)	78.51 (4.87)	76.64 (3.32)	81.72 (6.00)	75.30 (5.50)
	Self-confidence	5.25 (1.76)	5.55 (2.12)	5.21 (1.82)	N/A	N/A
	Workload	68.99 (14.50)	70.79 (9.47)	73.61 (10.97)	N/A	N/A
Different-Type Robotic	Trust	7.20 (1.05)	6.39 (1.65)	5.71 (1.56)	6.64 (1.60)	6.15 (2.25)
	Reliance	85.34 (6.14)	80.96 (5.68)	76.11 (6.85)	84.05 (5.87)	77.89 (6.56)
	Performance	83.96 (5.31)	79.65 (3.49)	75.83 (6.04)	83.03 (5.16)	76.26 (4.45)
	Self-confidence	5.53 (1.78)	5.09 (2.11)	4.61 (1.89)	N/A	N/A
	Workload	68.24 (14.00)	73.02 (13.06)	72.09 (18.13)	N/A	N/A
Same-Type Robotic	Trust	7.21 (1.27)	6.05 (1.78)	5.53 (1.75)	6.66 (1.84)	5.64 (2.28)
	Reliance	84.90 (6.80)	79.80 (5.32)	77.90 (5.76)	82.91 (5.90)	76.67 (6.35)
	Performance	83.76 (6.06)	79.95 (3.42)	77.88 (3.66)	82.32 (4.86)	77.58 (4.79)
	Self-confidence	5.12 (2.15)	5.45 (2.17)	5.18 (1.70)	N/A	N/A
	Workload	67.70 (13.04)	70.72 (14.06)	69.39 (11.48)	N/A	N/A
Control	Performance	76.62 (4.36)				
	Self-confidence	4.52 (1.81)				
	Workload	71.40 (13.83)				

Note: Values for each of the mixed-reliability agents are presented individually on the right of the table and averaged in the fourth column.

APPENDIX W: ATS QUESTIONNAIRE FACTORS CORRELATIONS TO TRUST AND RELIANCE

Overall correlations for the four factors: Extreme, Pet, God or Deity, and Negative Anthropomorphism. No overall correlations were significant ($p > .05$ in all cases).

Type Anthropomorphism	Dependent Measure	Correlation	Significance	N
Extreme	Trust	-.07	.24	296
	Reliance	-.05	.07	296
Pet	Trust	.10	.52	295
	Reliance	.02	.77	295
God or Deity	Trust	.04	.37	296
	Reliance	-.07	.80	296
Negative	Trust	.02	.24	296
	Reliance	.05	.43	296

Next correlations were broken down by agent type: Human, Same-Type Robotic, or Different-Type Robotic.

Type Anthropomorphism	Agent	Dependent Measure	Correlation	Significance	N
Extreme	Human	Trust	.07	.50	99
		Reliance	.01	.89	99
	Same	Trust	.03	.75	99
		Reliance	-.07	.52	99
	Dif.	Trust	*-.29***	.004	98
		Reliance	-.09	.358	98
Pet	Human	Trust	.09	.37	98
		Reliance	-.03	.74	98
	Same	Trust	.17	.10	99
		Reliance	.04	.68	99
	Dif.	Trust	.08	.41	98
		Reliance	.05	.65	98
God or Deity	Human	Trust	-.03	.79	99
		Reliance	-.09	.38	99
	Same	Trust	.13	.19	99
		Reliance	-.04	.70	99
	Dif.	Trust	-.02	.86	98
		Reliance	-.13	.22	98
Negative	Human	Trust	.07	.50	99
		Reliance	-.08	.44	99
	Same	Trust	.04	.72	99
		Reliance	-.03	.81	99
	Dif.	Trust	-.05	.66	98
		Reliance	*.22**	.03	98

* Correlation is significant at the 0.05 level (2-tailed), ** Correlation is significant at the 0.01 level (2-tailed)

Next correlations were broken down by reliability condition: Both High, Both Low, and Mixed.

Type Anthropomorphism	Reliability	Dependent Measure	Correlation	Significance	N
Extreme	High	Trust	-.04	.71	98
		Reliance	-.06	.53	98
	Mixed	Trust	-.12	.26	99
		Reliance	-.08	.42	99
	Low	Trust	-.06	.53	99
		Reliance	-.02	.89	99
Pet	High	Trust	.14	.18	97
		Reliance	-.08	.45	97
	Mixed	Trust	.01	.93	99
		Reliance	*.22**	.03	99
	Low	Trust	.13	.19	99
		Reliance	-.13	.19	99
God or Deity	High	Trust	.11	.28	98
		Reliance	-.06	.55	98
	Mixed	Trust	-.07	.51	99
		Reliance	-.12	.23	99
	Low	Trust	.13	.20	99
		Reliance	.02	.81	99
Negative	High	Trust	-.01	.93	98
		Reliance	-.10	.34	98
	Mixed	Trust	-.07	.49	99
		Reliance	.02	.82	99
	Low	Trust	-.02	.86	99
		Reliance	.07	.50	99

* Correlation is significant at the 0.05 level (2-tailed), ** Correlation is significant at the 0.01 level (2-tailed)

The final breakdown looked at agent by reliability condition.

Agent	**Type Anthropomorphism**	**Reliability**	**Dependent Measure**	**Correlation**	**Significance**	**N**
Human	Extreme	High	Trust	.12	.49	33
			Reliance	.15	.37	33
		Mixed	Trust	-.06	.76	33
			Reliance	-.16	.37	33
		Low	Trust	.02	.92	33
			Reliance	.26	.15	33
	Pet	High	Trust	.33	.07	32
			Reliance	-.20	.29	32

Agent	Type Anthropomorphism	Reliability	Dependent Measure	Correlation	Significance	N
		Mixed	Trust	-.01	.96	33
			Reliance	.35	.05	33
		Low	Trust	-.12	.53	33
			Reliance	-.27	.13	33
	God or Deity	High	Trust	-.09	.62	33
			Reliance	-.07	.70	33
		Mixed	Trust	-.14	.44	33
			Reliance	-.28	.11	33
		Low	Trust	-.01	.97	33
			Reliance	.00	1.00	33
	Negative	High	Trust	.10	.59	33
			Reliance	-.07	.70	33
		Mixed	Trust	-.20	.27	33
			Reliance	-.31	.08	33
		Low	Trust	.04	.82	33
			Reliance	-.04	.83	33
Same-Type Robotic	Extreme	High	Trust	-.16	.37	33
			Reliance	-.31	.08	33
		Mixed	Trust	.17	.34	33
			Reliance	.17	.34	33
		Low	Trust	.06	.74	33
			Reliance	.01	.96	33
	Pet	High	Trust	.02	.91	33
			Reliance	-.01	.97	33
		Mixed	Trust	.02	.91	33
			Reliance	-.07	.69	33
		Low	Trust	.18	.33	33
			Reliance	-.11	.55	33
	God or Deity	High	Trust	*.49***	.004	33
			Reliance	-.02	.91	33
		Mixed	Trust	-.05	.79	33
			Reliance	-.08	.64	33

Agent	Type Anthropomorphism	Reliability	Dependent Measure	Correlation	Significance	N
		Low	Trust	.04	.82	33
			Reliance	-.07	.71	33
	Negative	High	Trust	.02	.92	33
			Reliance	-.35*	.05	33
		Mixed	Trust	-.01	.98	33
			Reliance	.10	.59	33
		Low	Trust	.02	.92	33
			Reliance	.18	.33	33
Different-Type Robotic	Extreme	High	Trust	-.18	.33	32
			Reliance	.01	.98	32
		Mixed	Trust	-.47**	.01	33
			Reliance	-.13	.46	33
		Low	Trust	-.27	.12	33
			Reliance	-.24	.19	33
	Pet	High	Trust	.00	1.00	32
			Reliance	-.03	.89	32
		Mixed	Trust	-.13	.87	33
			Reliance	.42*	.02	33
		Low	Trust	.32	.07	33
			Reliance	-.08	.67	33
	God or Deity	High	Trust	-.11	.56	32
			Reliance	-.14	.44	32
		Mixed	Trust	-.01	.98	33
			Reliance	.03	.89	33
		Low	Trust	.39*	.03	33
			Reliance	.16	.38	33
	Negative	High	Trust	-.21	.25	32
			Reliance	.11	.53	32
		Mixed	Trust	-.06	.78	33
			Reliance	.36*	.04	33
		Low	Trust	-.12	.50	33
			Reliance	.04	.81	33

* Correlation is significant at the 0.05 level (2-tailed), ** Correlation is significant at the 0.01 level (2-tailed)

APPENDIX X: ITS QUESTIONNAIRE CORRELATIONS TO TRUST AND RELIANCE

Overall ITS correlations, no overall correlations were significant ($p > .05$ in all cases).

Dependent Measure	Correlation	Significance	N
Trust	-.10	.07	296
Reliance	-.07	.24	296

Next ITS correlations were broken down by agent type: Human, Same-Type Robotic, or Different-Type Robotic.

Agent	Dependent Measure	Correlation	Significance	N
Human	Trust	-.12	.24	99
	Reliance	-.21*	.04	99
Same	Trust	-.28**	.01	99
	Reliance	-.17	.10	99
Dif.	Trust	.10	.33	98
	Reliance	.13	.20	98

* Correlation is significant at the 0.05 level (2-tailed), ** Correlation is significant at the 0.01 level (2-tailed)

Next ITS correlations were broken down by reliability condition: Both High, Both Low, and Mixed.

Reliability	Dependent Measure	Correlation	Significance	N
High	Trust	.07	.48	98
	Reliance	.00	.98	98
Mixed	Trust	-.08	.46	99
	Reliance	-.12	.24	99
Low	Trust	-.20*	.05	99
	Reliance	.01	.92	99

* Correlation is significant at the 0.05 level (2-tailed)

The final breakdown looked at agent by reliability condition.

Agent Type	Reliability	Dependent Measure	Correlation	Significance	N
Human	High	Trust	.04	.81	33
		Reliance	-.21	.24	33
	Mixed	Trust	-.10	.59	33
		Reliance	-.30	.09	33
	Low	Trust	-.35*	.05	33
		Reliance	-.25	.15	33
Different-Type Robotic	High	Trust	.31	.08	32
		Reliance	.32	.07	32
	Mixed	Trust	.07	.71	33
		Reliance	.07	.70	33
	Low	Trust	.16	.37	33
		Reliance	.27	.13	33
Same-Type Robotic	High	Trust	-.02	.90	33
		Reliance	-.09	.63	33
	Mixed	Trust	-.18	.31	33
		Reliance	-.13	.46	33
	Low	Trust	-.43*	.01	33
		Reliance	-.15	.42	33

* Correlation is significant at the 0.05 level (2-tailed), ** Correlation is significant at the 0.01 level (2-tailed)

APPENDIX Y: CPRS QUESTIONNAIRE OVERALL AND FACTOR CORRELATIONS TO TRUST AND RELIANCE

Overall correlations for the four factors: Extreme, Pet, God or Deity, and Negative Anthropomorphism. No overall correlations were significant ($p > .05$ in all cases).

Type CPRS	Dependent Measure	Correlation	Significance	N
Overall	Trust	.18**	.002	296
	Reliance	.14*	.014	296
Confidence	Trust	.09	.134	296
	Reliance	.08	.185	296
Reliance	Trust	.14*	.015	296
	Reliance	.06	.298	296
Trust	Trust	.18*	.002	296
	Reliance	.11	.065	296
Safety	Trust	.09	.109	296
	Reliance	.16**	.007	296

Next correlations were broken down by agent type: Human, Same-Type Robotic, or Different-Type Robotic.

Type CPRS	Agent	Dependent Measure	Correlation	Significance	N
Overall	Human	Trust	.13	.19	99
		Reliance	.14	.16	99
	Same	Trust	.25*	.01	99
		Reliance	.24*	.02	99
	Dif.	Trust	.17	.10	98
		Reliance	.05	.60	98
Confidence	Human	Trust	.09	.37	99
		Reliance	.03	.76	99
	Same	Trust	.05	.62	99
		Reliance	.20*	.05	99
	Dif.	Trust	.12	.24	98
		Reliance	.02	.86	98
Reliance	Human	Trust	.08	.42	99
		Reliance	.14	.17	99
	Same	Trust	.27**	.01	99
		Reliance	.09	.36	99
	Dif.	Trust	.07	.48	98
		Reliance	-.05	.65	98
Trust	Human	Trust	.17	.09	99
		Reliance	.11	.26	99
	Same	Trust	.19	.07	99
		Reliance	.12	.23	99

Type CPRS	Agent	Dependent Measure	Correlation	Significance	N
	Dif.	Trust	.17	.09	98
		Reliance	.07	.50	98
Safety	Human	Trust	.01	.94	99
		Reliance	.13	.19	99
	Same	Trust	.16	.11	99
		Reliance	.22*	.03	99
	Dif.	Trust	.09	.38	98
		Reliance	.12	.26	98

* Correlation is significant at the 0.05 level (2-tailed), ** Correlation is significant at the 0.01 level (2-tailed)

Next correlations were broken down by reliability condition: Both High, Both Low, and Mixed.

Type CPRS	Agent	Dependent Measure	Correlation	Significance	N
Overall	High	Trust	.01	.93	98
		Reliance	.05	.60	98
	Mixed	Trust	.28**	.01	99
		Reliance	.22*	.03	99
	Low	Trust	.09	.37	99
		Reliance	-.06	.59	99
Confidence	High	Trust	.02	.86	98
		Reliance	.05	.62	98
	Mixed	Trust	.17	.10	99
		Reliance	.19	.06	99
	Low	Trust	-.02	.85	99
		Reliance	-.12	.25	99
Reliance	High	Trust	-.00	.97	98
		Reliance	-.01	.89	98
	Mixed	Trust	.30**	.00	99
		Reliance	.15	.15	99
	Low	Trust	.01	.91	99
		Reliance	-.13	.21	99
Trust	High	Trust	.10	.32	98
		Reliance	-.01	.92	98
	Mixed	Trust	.12	.23	99
		Reliance	.15	.15	99
	Low	Trust	.20*	.05	99
		Reliance	.07	.53	99
Safety	High	Trust	-.12	.25	98
		Reliance	.13	.21	98

Type CPRS	Agent	Dependent Measure	Correlation	Significance	N
	Mixed	Trust	.15	.15	99
		Reliance	.08	.44	99
	Low	Trust	.06	.57	99
		Reliance	.03	.76	99

* Correlation is significant at the 0.05 level (2-tailed), ** Correlation is significant at the 0.01 level (2-tailed)

The final breakdown looked at agent by reliability condition.

Agent	**Type** CPRS	**Reliability**	**Dependent Measure**	**Correlation**	**Significance**	**N**
Human	Overall	High	Trust	.11	.56	33
			Reliance	-.01	.98	33
		Mixed	Trust	.16	.36	33
			Reliance	.28	.11	33
		Low	Trust	.10	.59	33
			Reliance	.09	.63	33
	Confidence	High	Trust	.09	.62	33
			Reliance	-.17	.33	33
		Mixed	Trust	.22	.21	33
			Reliance	.25	.17	33
		Low	Trust	.00	.98	33
			Reliance	.10	.58	33
	Reliance	High	Trust	.27	.13	33
			Reliance	-.02	.91	33
		Mixed	Trust	.14	.43	33
			Reliance	.15	.40	33
		Low	Trust	.09	.62	33
			Reliance	-.03	.89	33
	Trust	High	Trust	.33	.06	33
			Reliance	.00	1.00	33
		Mixed	Trust	-.03	.87	33
			Reliance	.18	.33	33
		Low	Trust	.20	.26	33
			Reliance	.15	.41	33

Agent	Type CPRS	Reliability	Dependent Measure	Correlation	Significance	N
	Safety	High	Trust	-.13	.47	33
			Reliance	-.10	.57	33
		Mixed	Trust	.03	.86	33
			Reliance	.18	.33	33
		Low	Trust	-.03	.88	33
			Reliance	.05	.79	33
Same-Type Robotic	Overall	High	Trust	-.13	.47	33
			Reliance	.10	.57	33
		Mixed	Trust	.12	.49	33
			Reliance	.10	.58	33
		Low	Trust	.35*	.05	33
			Reliance	.13	.48	33
	Confidence	High	Trust	-.08	.68	33
			Reliance	.17	.35	33
		Mixed	Trust	-.02	.90	33
			Reliance	.07	.69	33
		Low	Trust	-.02	.93	33
			Reliance	.16	.38	33
	Reliance	High	Trust	.05	.79	33
			Reliance	-.08	.66	33
		Mixed	Trust	.24	.19	33
			Reliance	.01	.97	33
		Low	Trust	.17	.34	33
			Reliance	-.09	.63	33
	Trust	High	Trust	-.24	.18	33
			Reliance	-.03	.86	33
		Mixed	Trust	-.01	.94	33
			Reliance	.13	.47	33
		Low	Trust	.52**	.002	33
			Reliance	.09	.62	33
	Safety	High	Trust	-.03	.87	33
			Reliance	.23	.19	33

Agent	Type CPRS	Reliability	Dependent Measure	Correlation	Significance	N
		Mixed	Trust	.16	.37	33
			Reliance	.04	.82	33
		Low	Trust	.09	.61	33
			Reliance	.15	.40	33
Different-Type Robotic	Overall	High	Trust	-.11	.57	32
			Reliance	.01	.95	32
		Mixed	Trust	.56**	.001	33
			Reliance	.26	.14	33
		Low	Trust	-.12	.49	33
			Reliance	-.26	.15	33
	Confidence	High	Trust	-.07	.70	32
			Reliance	.13	.49	32
		Mixed	Trust	.40*	.02	33
			Reliance	.36*	.04	33
		Low	Trust	-.04	.81	33
			Reliance	-.37*	.04	33
	Reliance	High	Trust	-.14	.46	32
			Reliance	-.25	.17	32
		Mixed	Trust	.43*	.01	33
			Reliance	.21	.24	33
		Low	Trust	-.28	.12	33
			Reliance	-.25	.17	33
	Trust	High	Trust	.16	.38	32
			Reliance	-.02	.90	32
		Mixed	Trust	.45**	.01	33
			Reliance	.07	.70	33
		Low	Trust	-.10	.58	33
			Reliance	-.03	.88	33
	Safety	High	Trust	-.28	.13	32
			Reliance	.18	.34	32
		Mixed	Trust	.21	.25	33
			Reliance	.02	.91	33

Agent	Type CPRS	Reliability	Dependent Measure	Correlation	Significance	N
		Low	Trust	.07	.68	33
			Reliance	-.05	.80	33

* Correlation is significant at the 0.05 level (2-tailed), ** Correlation is significant at the 0.01 level (2-tailed)

APPENDIX Z: IRB APPROVAL FORMS

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901, 407-882-2012 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Notice of Expedited Review and Approval of Requested Addendum/Modification Changes

From: **UCF Institutional Review Board**
FWA00000351, Exp. 5/07/10, IRB00001138

To: **Jennifer M Ross**

Date: **October 29, 2007**

IRB Number: **SBE-07-05111**

Study Title: **Effects of distributed communication on comprehension of task goals**

Dear Researcher:

Your requested addendum/modification changes to your study noted above which were submitted to the IRB on 10/26/2007 were approved by **expedited** review on 10/29/2007.

Per federal regulations, 45 CFR 46.110, the expeditable modifications were determined to be minor changes in previously approved research during the period for which approval was authorized.

Use of the approved, stamped consent document(s) is required. The new form supersedes all previous versions, which are now invalid for further use. Only approved investigators (or other approved key study personnel) may solicit consent for research participation. Subjects or their representatives must receive a copy of the consent form(s).

This addendum approval does NOT extend the IRB approval period or replace the Continuing Review form for renewal of the study.

On behalf of Tracy Dietz, Ph.D., IRB Chair, this letter is signed by:

Signature applied by Janice Turchin on 10/29/2007 11:26:12 AM EST

Janice M Turchin

IRB Coordinator

Internal IRB Submission Reference Number: 001439

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901, 407-882-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Notice of Expedited Initial Review and Approval

From : **UCF Institutional Review Board**
FWA00000351, Exp. 5/07/10, IRB00001138

To : **Jennifer M Ross**

Date : **January 31, 2008**

IRB Number: **SBE-08-05401**

Study Title: **Robot Search-and-Rescue Study Part II**

Dear Researcher:

Your research protocol noted above was approved by **expedited** review by the UCF IRB Vice-chair on 1/30/2008. **The expiration date is 1/29/2009.** Your study was determined to be minimal risk for human subjects and expeditable per federal regulations, 45 CFR 46.110. The category for which this study qualifies as expeditable research is as follows:

> 7. Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

The IRB has approved a **consent procedure which requires participants to sign consent forms.** Use of the approved, stamped consent document(s) is required. Only approved investigators (or other approved key study personnel) may solicit consent for research participation. Subjects or their representatives must receive a copy of the consent form(s).

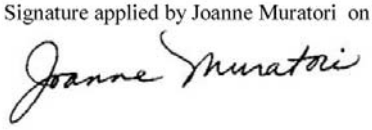
All data, which may include signed consent form documents, must be retained in a locked file cabinet for a minimum of three years (six if HIPAA applies) past the completion of this research. Any links to the identification of participants should be maintained on a password-protected computer if electronic information is used. Additional requirements may be imposed by your funding agency, your department, or other entities. Access to data is limited to authorized individuals listed as key study personnel.

To continue this research beyond the expiration date, a Continuing Review Form must be submitted 2 – 4 weeks prior to the expiration date. Advise the IRB if you receive a subpoena for the release of this information, or if a breach of confidentiality occurs. Also report any unanticipated problems or serious adverse events (within 5 working days). Do not make changes to the protocol methodology or consent form before obtaining IRB approval. Changes can be submitted for IRB review using the Addendum/Modification Request Form. An Addendum/Modification Request Form **cannot** be used to extend the approval period of a study. All forms may be completed and submitted online at http://iris.research.ucf.edu .

Failure to provide a continuing review report could lead to study suspension, a loss of funding and/or publication possibilities, or reporting of noncompliance to sponsors or funding agencies. The IRB maintains the authority under 45 CFR 46.110(e) to observe or have a third party observe the consent process and the research.

On behalf of Tracy Dietz, Ph.D., UCF IRB Chair, this letter is signed by:

Signature applied by Joanne Muratori on 01/31/2008 01:32:28 PM EST



IRB Coordinator

LIST OF REFERENCES

Bainbridge, L. (1983). Ironies of automation: Increasing levels of automation can increase, rather than decrease, the problems of supporting the human operator. *Automatica,* 19, 775-779.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W.H. Freeman.

Bass, E.J., & Pritchett, A.R. (2006). Human-automated judgment learning: Enhancing interaction with automated judgment systems. In A. Kirlik (Ed.). *Human-technology interaction: Methods and models for cognitive engineering and human-computer interaction,* (pp. 114-126). New York: Oxford University Press.

Barnett, J.S (2000). Affects of training on user confidence in automation. Unpublished Doctoral Dissertation, University of Central Florida, Orlando.

Beck, H.P., Dzindolet, M.T., & Pierce, L.G. (2002). Operators' automation usage decisions and the sources of misuse and disuse. *Advances in Human Performance and Cognitive Engineering Research,* 2, 37-78.

Bergeron, H.P., & Hinton, D.A. (1985). Aircraft automation : The problem of the pilot interface. *Aviation, Space, and Environmental Medicine,* 56(2), 144-148.

Bliss, J.P. (1993). *Alarm reaction patterns by pilots as a function of reaction modality.* Unpublished doctoral dissertation, University of Central Florida. Orlando, Florida.

Burdick, M.D., Skitka, L.J., & Mosier, K.L. (1997). The debiasing effects of accountability and feedback on automation bias. *Proceedings of the Human Factors Society 41st Annual Meeting.* Santa Monica, CA: The Human Factors Society.

Casey, S. (1993). *Set phasers on stun.* Santa Barbara, CA: Aegean.

Chappell, S.L. (1997). Cross-checked but not seen: The effects of automation and reliable systems. *Proceedings of the Human Factors Society 41st Annual Meeting.* Santa Monica, CA: The Human Factors Society.

Corritore, C.L., Kracher, B., & Wiedenbeck, S. (2001). Trust in online environment. In M.J. Smith, G. Salvendy, D. Harris, & R.J. Koubek (Eds.), *Usability evaluation and interface design: Cognitive engineering, intelligent agents and virtual reality*, (pp. 1548-1552). Mahwah, NJ: Lawrence Erlbaum Associates.

Deutsch, M. (1958). Trust and suspicion. *Journal of Conflict Resolution,* 2, 265-279.

Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. *Human Relations,* 13, 123-139.

De Vries, P., Midden, C., & Boushuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies,* 58, 719-735.

DiBello, L. (2001). Solving the problem of employee resistance to technology by reframing the problem as one of expertise and their tools. In: E. Salas & G. Klein (Eds.), *Linking expertise and naturalistic decision making* (pp. 71-93). Mahwah, NJ: Lawrence Erlbaum Associates.

Dzindolet, M.T., Beck, H., Pierce, L., & Dawe, L.A. (1998). Human decision making in an automated environment. *Proceedings of the Human Factors Society 42nd Annual Meeting.* Santa Monica, CA: The Human Factors Society.

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L., & Beck, H.P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies,* 58, 697-718.

Gempler, K.S., & Wickens, C.D. (1998). *Display of predictor reliability on a cockpit display traffic information,* Technical Report ARL-98-6/ROCKWELL-98-1 (Savoy: University of Illinois, Aviation Research Lab).

Global Security. (2005)). M2 and M3 Bradley Fighting Vehicle Systems (BFVS). Retrieved July 3, 2007, from http://www.globalsecurity.org/military/systems/ground/m2.htm

Halpin, S., Johnson, E., & Thornberry, J. (1973). Cognitive reliability in manned systems. *IEEE Transactions on Reliability*, R-22(3), 552-564.

Hancock, P.A., Mouloua, M., Gilson, R.D., Szalma, J., & Oron-Gilad, T. (2007). Provocation: Is the UAV control ratio the right question? Ergonomics in Design, 15(1), 7-30.

Hancock, P.A., & Parasuraman, R. (1992). Human factors and safety in the design of intelligent vehicle-highway systems (IVHS). *Journal of Safety Research,* 23(4), 181-198.

Hancock, P.A., Parasuraman, R., & Byrne, E.A. (1996). Driver-centered issues in advanced automation for motor vehicle. In R. Parasuraman and M. Mouloua (Eds.), Automation and human performance (pp337-364). Mahwah, NJ: Lawrence Erlbaum Associates.

Hart, S.G. & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock and N. Meshkati (Eds.), Human mental workload (pp. 139-183). Amsterdam: North-Holland.

Hotiu, A. The relationship between item difficulty and discrimination indices in multiple-choice tests in a physical science course. Unpublished masters thesis, Florida Atlantic University, Boca Raton, Fl, USA.

Inman, R.F. (2001). *Item response theory.* Unpublished masters thesis, University of Central Florida, Orlando, Fl, USA.

Jian, J., Bisantz, A.M., & Drury, C.G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics,* 4(1), 53-71.

Kantowitz, B.H., Campbell, J.L. (1996). Pilot workload and flightdeck automation. In R. Parasuraman and M. Mouloua (Eds.) *Automation and human performance: Theory and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Karvonen, K., & Parkkinen, J. (2001). Signs of trust: A semiotic study of trust formation in the web. In M.J. Smith, G. Salvendy, D. Harris, & R.J. Koubek (Eds.), *Usability evaluation and interface design: Cognitive engineering, intelligent agents and virtual reality*, (pp. 1076-1080). Mahwah, NJ: Lawrence Erlbaum Associates.

Lee, J.D. (1991). Trust, self-confidence and operators' adaptation to automation. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Lee, J.D., & Moray, N. (1992). Trust, control strategies and allocation of function for in-vehicle warning and sign information: Message style, location, and modality. *Transportation Human Factors,* 1, 347-377.

Lee, J.D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies,* 40, 153-184.

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

Lerch, F.J., & Prietula (1989). How do we trust machine advice? In M.J. Smith (Ed.), *Designing and using human-computer interface and knowledge based systems (*pp. 410-419). Amsterdam: Elsevier Science.

Lewandowsky, S., Mundy, M., & Tan, G.P.A. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied,* 6(2), 104-123.

Liu, C., & Hwang, S. (2000). Evaluating the effects of situation awareness and trust with robust design in automation. *International Journal of Cognitive Ergonomics,* 4(2), 125-144.

Masalonis, A.J., & Parasuraman, R. (2003). Fuzzy signal detection theory: Analysis of human and machine performance in air traffic control, and analytic considerations. *Ergonomics,* 46(11), 1045-1074.

May, P., Molloy, R., & Parasuraman, R. (1993). *Effects of automation reliability and failure rate on monitoring performance in a multitask environment.* Paper presented at the Annual Meeting of the Human Factors Society, Seattle, WA.

McClumpha, A.J., & James, M. (1994). Understanding automated aircraft. In M. Mouloua and R. Parasuraman (Eds.), *Human performance in automated systems: Current research and trends* (pp. 183-190). Hillsdale, NJ: Earlbaum.

McClumpha, A.J., James, M., Green, R.G., & Belyavin, A.J. (1991). Pilot's attitudes to cockpit automation. *Proceedings of the Human Factors Society 35th Annual Meeting.* Santa Monica, CA: The Human Factors Society.

McQuarrie, E.F. & Iwamoto, K. (1990). Public opinion toward computers as a function of exposure. *Social Science Computer Review*, 8, 221-233

Morgan, B.B. Jr., Herschler, D.A., Wiener, E.L., & Salas, E. (1993). Implications of automation technology for aircrew coordination and performance. *Human/Technology Interaction in Complex Systems,* 6, 105-136.

Mosier, K.L., & Skitka, L.J. (1998). Automation bias and errors: Are teams better than individuals? *Proceedings of the Human Factors Society 42nd Annual Meeting.* Santa Monica, CA: The Human Factors Society.

Mouloua, M., Gilson, R., & Hanock, P. (2003). Human-centered design of unmanned aerial vehicles. *Ergonomics in Design,* 11, 6-11.

Mouloua, M., Gilson, R.D., & Koonce, J. (1997). Automation, flight management and pilot training: Issues and considerations. In R.A. Telfer and P.J. Moore (Eds.), *Aviation training: Learners, instruction and organization* (pp. 78-86). Aldershot, United Kingdom: Avebury Aviation.

Muir, B.M. (1988). Trust between humans and machines, and the design of decision aids. In E. Hollnagel, G. Mancini, & D.D. Woods, Eds. *Cognitive engineering in complex dynamic worlds,* pp. 71-84. London: Academic Press.

Muir, B.M. (1989). *Operators' trust in and use of automatic controllers in a supervisory process control task.* Unpublished Doctoral Dissertation, University of Toronto.

Muir, B.M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics,* 37(11), 1905-1922.

Muir, B.M. & Moray, N. (1996). Trust in automation: Part II. Experimental studies of trust and human interaction in a process control simulation. *Ergonomics,* 39(3), 429-460.

National Transportation Safety Board. (1997). *Marine accident report – Grounding of the Panamanian passenger ship Royal Majesty on Rose and Crown Shoal near Nantucket, Massachusetts, June 10, 1995* (NTSB/MAR97/01). Washington, DC: Author.

Parasuraman, R., Masalonis, A.J., & Hancock, P.A. (2000). Fuzzy signal detection theory: Basic postulates and formulas for analyzing human and machine performance. *Human Factors,* 42(4), 636-659.

Parasuraman, R., Molloy, R., & Singh, I.L. (1993). Performance consequences of automation-induced "complacency." *International Journal of Aviation Psychology,* 3(1), 1-23.

Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors,* 39(2), 230-253.

Pritchett, A.R., & Bisantz, A.M. (2006). Measuring the fit between human judgments and alerting systems: A study of collision detection in aviation. In A. Kirlik (Ed.). *Human-technology interaction: Methods and models for cognitive engineering and human-computer interaction,* (pp. 91-104). New York: Oxford University Press.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* New York: Cambridge University Press.

Rempel, J.K., Holmes, J.G., & Zanna, M.P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology,* 49(1), 95-112.

Riley, V. (1994). A theory of operator reliance on automation. In. M. Mouloua and R. Parasuraman (Eds.), *Human performance in automated systems: Current research and trends* (pp. 8-14). Hillsdale, NJ: Erlbaum.

Ross, W. & LaCroix, J. (1996). Multiple meanings of trust in negotiation theory and research: A literature review and integrative model. *International Journal of Conflict Management,* 7, 314-360.

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

Rotter, J.B. (1971). Generalized expectancies for interpersonal trust. *American Psychologist,* 26*,* 443-452.

Rotter, J.B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist,* 35*,* 1-7.

Seong, Y., Bisantz, A.M., & Gattie, G.J. (2006). Trust, automation, and feedback: An integrated approach. In A. Kirlik (Ed.). *Human-technology interaction: Methods and models for cognitive engineering and human-computer interaction,* (pp. 105-113). New York: Oxford University Press.

Shavelson, R.J. (1996). Statistical reasoning for the behavioral sciences (3rd ed.). Boston: Allyn & Bacon.

Sheridan, T.B. (2002). Humans and automation: System design and research issues. Santa Monica, CA: John Wiley & Sons, Inc.

Sheridan, T.B., & Hennessy, R.T. (1984). *Research and Modeling of Supervisory Control Behavior.* Washington, DC: National Academy Press.

Sheridan, T.B., Vamos, T., & Aida, S. (1983). Adapting automation to man, culture and society, *Automatica,* 19, 605-612.

Singh, I.L., Molloy, R., & Parasuraman, R. (1993). Individual differences in monitoring failures of automation. *Journal of General Psychology,* 120(3), 357-373.

Squire, P., Trafton, G., & Parasuraman, R. (2006). Human control of multiple unmanned vehicles: Effects of interface type on execution and task switching time. *Proceedings of the SIGCHI/SIGART conference on Human-robot interaction,* USA, 1, 26-32.

Weigmann, D.A. (2001). Agreeing with automated diagnostic aids: A study of users' concurrence strategies. *Human Factors,* 44(1), 44-50.

Wiegmann, D.A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues of Ergonomic Science, 2*(4), 352-367.

Wiener, E.L. (1988). Cockpit automation. In E.L. Wiener and D.C. Nagel (Eds.), *Human factors in aviation* (pp. 433-461). San Diego, CA: Academic Press.

Wrightsman, L.S. (1991). Interpersonal trust and attitudes toward human nature. In: Robinson, J.P., Shaver, P.R., and Wrightsman, L.S. (Eds.). *Measures of personality and social psychological attitudes*, pp. 373-412. San Diego: Academic Press.

Young, M.S., & Stanton, N.A. (2001). I didn't do it: Accidents of automation. In M.J. Smith, G. Salvendy, D. Harris, and R.J. Koubek (Eds.), *Usability evaluation and interface design: Cognitive engineering, intelligent agents and virtual reality*, (pp. 1410-1414). Mahwah, NJ: Lawrence Erlbaum Associates.

Zuboff, S. (1988). *In the age of the smart machine: The future of work and power*. New York: Basic Books.