2013

The Effects Of Diagnostic Aiding On Situation Awareness Under Robot Unreliability

David Schuster
University of Central Florida

THE EFFECTS OF DIAGNOSTIC AIDING ON SITUATION AWARENESS UNDER ROBOT UNRELIABILITY

by

DAVID A. SCHUSTER
B.A. University of Tampa, 2006
M.A. University of Central Florida, 2010

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

Summer Term
2013

Major Professor: Florian G. Jentsch

ABSTRACT

In highly autonomous robotic systems, human operators are able to attend to their own, separate tasks, but robots still need occasional human intervention. In this scenario, it may be difficult for human operators to determine the status of the system and environment when called upon to aid the robot. The resulting lack of situation awareness (SA) is a problem common to other automated systems, and it can lead to poor performance and compromised safety. Existing research on this problem suggested that reliable automation of information processing, called diagnostic aiding, leads to better operator SA. The effects of unreliable diagnostic aiding, however, were not well understood. These effects are likely to depend on the ability of the operator to perform the task unaided. That is, under conditions in which the operator can reconcile their own sensing with that of the robot, the influence of unreliable diagnostic aiding may be more pronounced. When the robot is the only source of information for a task, these effects may be weaker or may reverse direction. The purpose of the current experiment was to determine if SA is differentially affected by unreliability at different levels of unaided human performance and at different stages of diagnostic aiding. This was accomplished by experimentally manipulating the stage of diagnostic aiding, robot reliability, and the ability of the operator to build SA unaided. Results indicated that while reliable diagnostic aiding is generally useful, unreliable diagnostic aiding has effects that depend on the amount of information available to operators in the environment. This research improves understanding of how robots can support operator SA and can guide the development of future robots so that humans are most likely to use them effectively.

To my grandparents, Helen and Michael Gross, and Helen and Edwin Schuster

ACKNOWLEDGMENTS

I am wholeheartedly grateful to the many people who made this dissertation possible. Thank you to the members of my committee: Mustapha Mouloua, James Szalma, and Randall Shumaker. My academic advisor, Florian Jentsch, has done more for my professional and educational development than I could have imagined; Florian, I cannot thank you enough.

I have had the honor of working with exceptional graduate and undergraduate students as a member of the Team Performance Laboratory. Thank you to my colleagues and research assistants, past and present, for providing a great place to research and learn. Without the determined work of Andrew Talone, Tony Baker, and Kate Kapalo, this study could not have been completed on time. Scott Ososky, Beth Phillips, Liz Lazzara, and Joe Keebler provided thoughtful sounding boards and support. Thomas Fincannon helped solve a number of statistical enigmas.

Finally, I would not be here without the support of my family. Thank you Mom and Dad, Natalie, Brandon, and Jonathan, for your love, encouragement, and patience. Last but not least, thank you, Philip, for sharing this journey with me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

Statement of the Problem

Robotic systems for war are evolving rapidly. One reason for this is a United States congressional mandate that at least one-third of military systems be unmanned by 2015 ("Building unmanned ground vehicles," 2003). Unmanned systems will be capable of a wider range of autonomous behavior while working in closer collaboration with people. On the battlefield, robots may support and execute missions, select tactics, and understand political contexts (Chen, Haas, & Barnes, 2007). Future robots will be “co-combatants with teams of Soldiers in complex tactical environments” (United States Army Research Laboratory, 2011, para. 3).

In highly autonomous robotic systems, human operators are able to attend to their own, separate tasks, rather than directly operating the robot to accomplish its primary task. Because this a major benefit of these systems, robots are being developed to function with as little human intervention as possible (Cosenzo, Parasuraman, & De Visser, 2010). Nevertheless, as robots grow in capability, they will continue to need occasional human intervention (Burke, Murphy, Coovert, & Riddle, 2004). For highly autonomous robot systems to succeed, imperfect robots must, therefore, handle complex tasks while still allowing humans to intervene when they fail.

The Out-of-the-Loop Performance Problem

Robot autonomy makes robot work possible by multiplying the effort of humans and freeing them from dangerous or undesirable tasks. Under conditions of high robot autonomy, human operators attend to their own tasks almost exclusively. In this scenario, it can become difficult for humans to determine the status of a system and of its environment when called upon to aid a robot. The goal-directed, high-level knowledge held by the human operator is known as situation awareness (SA; Rousseau et al., 2004). When a robot is highly autonomous and reliable, the details of its task are largely unimportant to the performance of tasks allocated to the human. When a robot is not fully reliable, however, the human must devote additional cognitive resources to correct, supplement, or mitigate its mistakes. Such demands can negatively affect both SA and performance (Endsley, 1995). A resulting lack of SA under conditions of automation failure has come to be known as the *out-of-the-loop performance problem* (Endsley, 1995) and the *out-of-the-loop unfamiliarity problem* (Wickens, 2002).

The Problem of Unreliable Robot Autonomy

The *out-of-the-loop performance problem* characterizes a common problem of capable, autonomous robots (Wickens, 2002). Robustness in robot capabilities remains a challenge to robotics (Stancliff, Dolan, & Trebi-Ollennu, 2005). As new capabilities are developed, robots may be able to perform new tasks, but reliability will be limited, especially initially. For human operators to take advantage of new robot capabilities, operators must be able to recover from robot failures.

When robots fail, they may do so in non-obvious ways. The need for a shift from full autonomy to operator intervention may not be pronounced. For example, a robot may navigate terrain independently but become stuck in the mud. Because its wheels are spinning normally, it does not sense that something is amiss and makes no notification to its human operator. This scenario is quite different from one in which there is an obvious failure of a complete subsystem. For example, a pilot may need to revert to an alternate method of navigation or control when an instrument in the cockpit fails or displays a value out of range. A combination of signals in the automation (the blank display or wrong value), combined with pilot knowledge and SA, indicate that a task previously handled by automation must be performed manually.

Because of the complexity and remoteness of environments in which robots work, robot failures may include instances in which a robot provides apparently valid information that is based on incorrect sensing. Although this type of failure may occur in other systems, and robots also fail in more easily detectable ways, a pressing problem is dealing with subtle sensing errors in a system separated from its operator by task assignment. That is, the operator and robot have different roles, making it difficult or unnecessary to monitor each action of the robot. These types of failures are not mechanical failures (such as when the robot is stuck and cannot complete the task), but rather failures in robot sensing and intelligence (i.e., the robot completes the task but does so incorrectly). At present, measuring a robot's confidence or meta-awareness of its sensors is a more difficult problem than sensing (for example, Eski, Erkaya, Savas, & Yildirim, 2010). Consequently, robot mistakes may be detectable only through cross-checking with other data, and failure at lower levels (i.e., sensing) may only have noticeable consequences at higher levels (i.e., decision making). From the operator's perspective, a shift takes place when a robot fails; what previously did not need to be known by the operator must now be attended. To maintain SA, interventions are needed to support the operator's information processing under conditions of robot unreliability.

Existing Approaches

An early response to the problem of operator SA was to investigate how automation, such as performed by a robot, or lack of automation, in a task may affect SA (Kaber & Endsley, 1997). As a form of automation, the robot's involvement in the task can be described as the *level of automation*. The general case of automation has been widely studied (see Parasuraman, Sheridan, & Wickens, 2000). An early taxonomy applicable to robots was developed by Sheridan and Verplank (1978) and expanded upon by Parasuraman, Sheridan, and Wickens (2000). Importantly, their taxonomy expanded upon prior models by including *what* task is automated in addition to *how much* automation is used. Under this model, the level of automation for a robot can be described as: (a) the levels of information processing in which the robot participates (i.e., *what*), and (b) the conditions under which the robot participates in each process (i.e., *when*).

The first two levels of this model are the focus of the current investigation because they map clearly onto Endsley's (1988) levels of SA (Horrey, Wickens, Strauss, Kirlik, & Stewart, 2009). Horrey et al. (2009) described a model in which information acquisition (stage 1 automation), leads to information analysis (stage 2 automation). Information acquisition (stage 1) is linked to level 1 SA, perception of elements, by sensation, perception, and attention. Information analysis (stage 2) is linked to level 2 SA, comprehension of the situation, by cognition, integration, and working memory. The first two levels of the model are the two stages of diagnostic aiding (Wickens & Dixon, 2007). Robots perform the *information acquisition* (stage 1) stage of diagnostic aiding when they gather relevant information through their sensors. Robots perform *information analysis* (stage 2), when they integrate multiple pieces of sensor data or when they integrate sensor data with previously stored or externally provided information. Thus, information acquisition (stage 1) is a precursor to information analysis (stage 2), and a robot that performs both stages operates at a higher level of automation than one that only performs information acquisition (stage 1). Automation that provides the later stage of diagnostic aiding leads to better decision making (Dexter, Willemsen-Dunlap, & Lee, 2007) and performance (Goodrich et al., 2007) in operators while lowering their workload (Manzey, Richenbach, & Onnasch, 2012).

Although robot diagnostic aiding may be beneficial to SA, the literature suggests that this relationship is highly sensitive to the presence of unreliability in the robot, and that the two stages of diagnostic aiding (i.e., information acquisition and information analysis) may be differentially affected. Performance decreases as the reliability of a diagnostic aid falls (Madhavan & Phillips, 2010). While unreliable information negatively impacts performance, the effect may be stronger for information analysis (stage 2) automation than for information acquisition (stage 1) automation (Rovira, McGarry, & Parasuraman, 2007; Sarter & Schroeder, 2001).

Mediators of the reliability-to-performance relationship have been investigated (e.g., trust and its impact on reliance; Madhavan & Wiegmann, 2007), but researchers have not addressed cognitive mediators. Specifically, there is a lack of understanding of the effects of reliability in information processing on an operator's ability to build and maintain SA. Theoretically, poor SA should explain the performance outcomes. Further, operators make use of other strategies to mitigate the problems caused by limited automation reliability (Johnson et al., 2009) and rely upon automation more than their own diagnoses when automation performs more reliably than the operator's unaided performance level (Madhavan & Wiegmann, 2007). One relevant case remains unexplored: what is the effect of unreliable diagnostic aiding on SA when the operator is unable to perform the task effectively when unaided?

In a recent meta-analysis, Wickens, Li, Santamaria, Sebok, and Sarter (2010) reported only five studies that investigated the effects of level of automation on performance and SA. They noted, "The situation awareness data are too few to create any well defined trend" (p. 391). This need was addressed by the current empirical study. Specifically, the purpose of this research was to investigate whether: (a) diagnostic aid reliability affects SA and performance in the same way, and (b) how this relationship may change as a function of different levels of unaided human performance.

Research Needs Addressed by the Current Study

A study addressing how the stage of diagnostic aiding, robot reliability, and unaided task performance interact to affect SA is important for several reasons. First, it may provide cognitive explanations for the effects of these constructs on performance. Past research has focused on either performance or non-cognitive mediators, such as misuse or disuse, while not addressing how these constructs may affect the knowledge held by the operator.

Second, because unaided human performance is an important, yet often overlooked, consideration for robot system design, this research will lead to understanding of how robots and humans can work interdependently to support the operator's SA. This will mean that any unreliability of the robot can be managed to minimize its negative impact, while the operator will receive support at the most appropriate time and to the degree needed.

Third, understanding of the determinants of SA within human-robot interaction will provide foundational knowledge needed to build the next generation of highly autonomous robotic systems. In order to implement autonomy effectively, we need to know how the operator's cognition may be affected. Of course, it would be ideal to make robots more autonomous and reliable while having human operators only perform tasks at which they excel (Sheridan, 2000). In reality, however, each of these variables is often one half of a tradeoff that must be made in order to satisfy other requirements for mission success. Reliability, especially, may be limited as early robots implement new forms of technology. The current study is also useful for understanding human cognition under various conditions of robot performance while simultaneously providing a foundation for applied researchers in the implementation of more effective future robots.

Purpose of the Current Study

Theoretical Perspective

The purpose of this study was to determine the conditions under which diagnostic aiding would contribute to operator SA, given limitations of robot reliability, on the one hand, and unaided human task performance, on the other. Because the stages of diagnostic aiding (Wickens & Dixon, 2007) map cleanly onto Endsley's first two levels of SA (1988), diagnostic aids that perform information analysis (stage 2) as well as information acquisition (stage 1) should lead to higher levels of SA than information acquisition (stage 1) alone. However, this relationship has been observed only under cases of perfect robot reliability.

Under imperfect reliability, two outcomes are possible, depending on the unaided performance of the operator. In the first case, the operator has a moderate or better capability to perform the task and build SA. In the presence of an unreliable robot, the challenge of integrating incorrect robot information should lower operator SA. Further, increasing the stage of diagnostic aiding (that is, adding automation of information analysis [stage 2]) should lower operator SA to a greater degree, because the operator must reconcile multiple, potentially unreliable or incorrect, pieces of information. The literature suggests that reconciling robot errors is more difficult in an integrated form (Rovira, McGarry, & Parasuraman, 2007; Sarter & Schroeder, 2001).

In the second case, consider an operator who has little or no ability to perform the task unaided. Even in the presence of an unreliable robot, I hypothesized that the operator would rely upon the robot completely. While the robot's reliability may have been poor, it would still offer a benefit to operators, and its use would be adaptive. These operators would have higher SA under higher stages of diagnostic aiding because they were dependent on the robot.

Study Variables

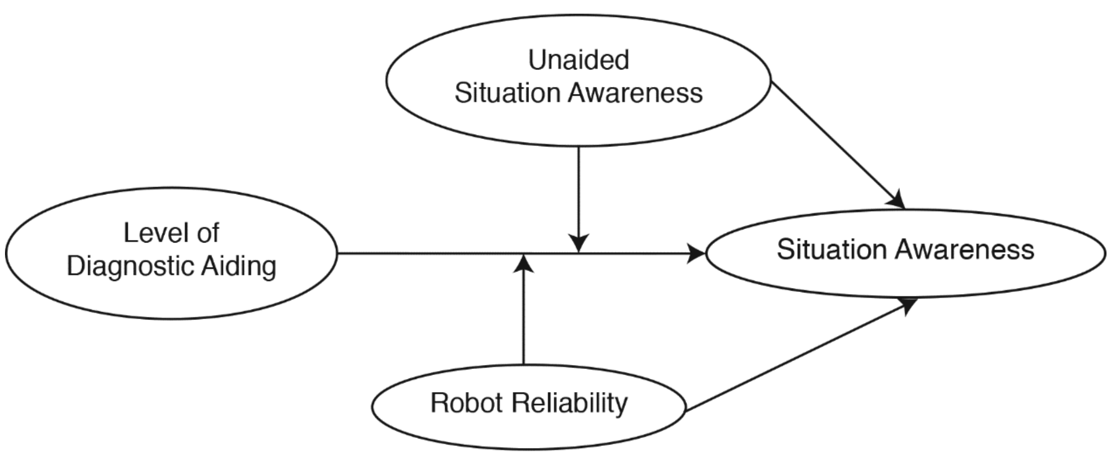


Figure 1. Model of relationships among study constructs.

The current study tested this theory by investigating the three-way interaction of stage of diagnostic aiding, robot reliability, and potential unaided SA on operator SA, while controlling for the amount of information provided by the robot (see Figure 1). The three independent variables were defined as follows. First, diagnostic aiding was defined as a form of automation operating at one of two levels. *Stage 1 diagnostic aiding* was defined as *information acquisition*. *Stage 2 diagnostic aiding* was specified as *information acquisition with information analysis*. Second, *robot reliability* was manipulated at three levels (60%, 80%, and 100%) and defined as the percentage of the time that sensing performed by the robot is correct. The three levels of reliability were selected to span across the range at which diagnostic aiding may be implemented and could be useful. Specifically, the lowest level (60%) was selected because it was at the bottom end of the 95% confidence interval for the minimum reliability level identified by Wickens and Dixon (2007) at which diagnostic aiding is still useful. The frequency of the robot's presentation of information was controlled across conditions. The robot's presentation of information was unidirectional. That is, there was no feedback loop to the robot, and the robot did not learn from its mistakes within a mission.

Third, *potential unaided SA* was defined as the percentage of relevant elements within the mission environment that the operator could reasonably access without the robot's assistance. It was manipulated at three levels (good [90%], moderate [50%], poor [10%]) that corresponded to the proportion of relevant mission elements that the operator could obtain information about without use of the robot. Put simply, it was how much information the operators could access and use on their own.

The dependent variable, *operator SA*, was defined as the percentage of relevant mission elements known by the operator during the mission. A measure of SA using objective questions about the mission elements was modeled after the SA Global Assessment Technique (SAGAT; Endsley, 2000a). SAGAT is an *objective* measure of SA whereby the task is paused mid-mission, and participants are prompted to answer objective questions. SA was also measured through participant self-assessment using the Situation Awareness Rating Technique (SART; following the methodology of Endsley et al., 2000a). SART is a *subjective* self-report measure of SA that was administered after each mission. SART consists of ten questions within three subscales: demand on attentional resources, supply of attentional resources, and understanding of the situation.

CHAPTER TWO: LITERATURE REVIEW

Situation Awareness

Situation awareness (SA) describes the relevant knowledge held by an operator while performing a task. Endsley (1988) created a model in which SA is a high-level, goal-directed information-processing function as part of a sensation-decision-action cycle. Within this model, SA has three levels: (a) perception of elements in the current situation, (b) comprehension of the current situation, and (c) projection of future status. Although these are not intended to be strictly hierarchical, higher-level comprehension develops through integration of lower level perception (Endsley, 2000b). There is an ongoing debate regarding whether SA should represent a state of knowledge or a process (Rousseau, Tremblay, & Breton, 2004), but the present study focused on SA as an *outcome*. Generally, SA is goal-directed, high-level knowledge that results from an individual's information processing within an environment (Rousseau et al., 2004). This definition implies that the *process* aspect of SA is human information processing. SA has been demonstrated to be a determinant of performance across systems (Durso & Sethumadhavan, 2008; Wickens, Li, Santamaria, Sebok, & Sarter, 2010). Because it is the knowledge needed to perform a task, SA, by definition, supports performance in tasks requiring maintenance of dynamic, complex knowledge of task states.

Theoretical Issues Surrounding SA

As a construct, SA arose from observations of fighter pilots, who considered it an intuitive skill long before it was investigated scientifically (Harwood, Barnett, & Wickens, 1988). As a consequence, the science of SA has required some time to catch up to its use in the field. This mismatch has resulted in a construct that is challenging to measure and inconsistently defined. These issues deserve brief discussion to better frame the use of the construct in this research.

Process versus outcome. Endsley's model of SA borrows from information processing theory (Tenney & Pew, 2006) and blends state and process by considering SA at three levels. Outcomes at these three levels (perception, comprehension, and projection) are intertwined with process. Some theorists have proposed that SA is better described as two constructs, the process (situation assessment) and the outcome (SA; Endsley, 1995; Pew 1994; Salas, Cannon-Bowers, Fiore, & Stout, 2001), while others claim the two are theoretically and practically intertwined (Vidulich, Dominguez, Vogel, & McMillan, 1994). A lack of comparability across measurement techniques (Endsley, Sollenberger, Nakata, & Stein, 2000) complicates both theory building and measurement. Note that current models and prior research blend process and outcome as taking place at the individual cognitive level. This research, therefore, defined SA as the resultant knowledge from the process of situation assessment. Generally, it can be said that SA is goal-directed, high-level knowledge that comes as a result of an individual's information processing within an environment (Rousseau et al., 2004).

What is a Robot?

It is important to operationally define the term *robot* and distinguish it from related constructs, because there is no widespread agreement on the meaning of the term. When asked his definition, Engelberger said, "I can't define a robot, but I know one when I see one" (as cited in "Your view: How would," 2007). The term originally came from Karel Čapek's 1920 play, R.

U. R. (Čapek, 1920). In the play, robots were assembled creatures made of lab-grown organs designed to perform mechanical labor. Modern definitions of a robot tend to envision electromechanical systems, and they vary in how a robot is distinguished from other forms of agents.

Agents

At minimum, a robot is an agent. An agent is an "an instantiation of an object together with an associated goal or set of goals" (Luck & d'Inverno, 1995, p. 55). Common to the definitions of agency are a distinguishable entity capable of performing a task or goal. I will use this broad construct to refer to either people or robots as goal-oriented, task-performing entities with the caveat that agent refers to little more than the ability to execute a task and is not an implication of similarity to people.

Operational Definition of Robots

The capabilities that distinguish an agent from a robot are needed to qualify the definition of a robot. For purposes of this research, a robot is an *electromechanical device or system that performs a task or goal, is physically embodied, senses the world, and acts upon the world*. Embodiment distinguishes a postal truck from an e-mail server; the latter does not have "a body" nor interacts with the world by using its body for sensation and action (Kiesler, Powers, Fussell, & Torrey, 2008). Being physically embodied implies that the agent acts upon the world. Acting upon the world distinguishes the postal truck from a postal scale. Moving, traversing, carrying, building, and attacking are all actions upon the physical world. The world is further distinguished from a closed, controlled area by its complexity; robots must perform in unpredictable environments. This distinguishes a robot from a dishwasher, which acts upon a limited and highly controlled environment. Finally, robots are closed-loop systems in that their behavior is affected by the result of their own sensing.

Applications of the Operational Definition

Animatronics, animated machines, are not robots when their actions are preprogrammed and do not change based on their own sensing. Being a robot implies that the agent is capable of sensing the world and incorporating the sensor input (even minimally or combined with the actions of a human operator) into future behavior. A hammer acts upon the physical world but does not sense the world. Similarly, an industrial "robot" on a manufacturing line may not be considered a robot by this definition. The ISO defines a robot as an "automatically controlled, reprogrammable, multipurpose manipulator programmable in three or more axes" (International Organization for Standardization, 1994). If the industrial "robot" simply parrots a preprogramed task in an open loop system, it falls outside the operational definition for the present work.

Level of Automation as a Metric of Robot Autonomy

Definition and Taxonomies

The degree to which a robot is involved in the task has been called the *level of robot autonomy*. Autonomy has origins in the Greek word *autonomia*, which means independence ("Autonomy," 2011). The United States Department of Defense defines an autonomous battlefield entity as one "that does not require the presence of another battlefield entity in order to conduct its own simulation in the battlefield environment" (United States Department of

Defense, 1998, p. 91). Robot autonomy is the extent to which the behavior of a robot results from integration of its own sensing (Franklin & Graesser, 1997) and the extent to which it makes decisions not mediated by other entities (Luck & D'Inverno, 1995). Robot autonomy is generally discussed as a quality integrating the robot's capabilities and authority across a series of tasks (Johnson et al., 2010). Although autonomy is an intuitive quality of robots, there are few developed quantitative models that describe robot autonomy (an exception is Goodrich, McLain, Anderson, Sun, & Crandall, 2007), and fewer still that include operational definitions at a level required for the current empirical investigation.

Robot behavior can be considered to belong to the broader class of automation. Automation is any "device or system that accomplishes (partially or fully) a goal that was previously, or conceivably could be, carried out (partially or fully) by a human operator" (Parasuraman, Sheridan, & Wickens, 2000, p. 287). By considering robot behavior as automation, one can describe the robot's involvement in a particular task as the *level of automation*.

This study considered the level of automation to be a quantifiable measure of robot autonomy. Thus, the two terms are very similar; the level of automation is a way to quantify the level of robot autonomy.

The level of automation (LOA) can be described as how fully a system carries out the task; the taxonomy was that of Sheridan and Verplank (1978), who described levels from full operator control to complete task execution by automation. The current study used the level of automation as an operational definition for robot autonomy.

Models of Level of Automation

The effects of automation on performance may depend on how automation is defined (i.e., what is automated) in relation to how performance is defined (i.e., what must be accomplished). Parasuraman, Sheridan, and Wickens (2000) offered, separately, a text revision (see Table 1) and an expansion of Sheridan and Verplank's (1978) levels of automation, which added a second dimension, information processing stage, to describe what tasks are performed. Thus, information acquisition (stage 1), information analysis (stage 2), decision selection (stage 3), and action implementation (stage 4) each have an independent level of automation that can vary from low to high.

Table 1
Parasuraman, Sheridan, and Wickens (2000, p. 287, Table 1) levels of 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 to veto before automatic execution, or
5	Executes that suggestion if the human approves, or
4	Suggests one alternative
3	Narrows the selection 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 take all decisions and actions

The first two levels of this model are known as diagnostic aiding (Wickens & Dixon, 2007). Diagnostic aiding encompasses automation of information processing as a precursor to (and excluding) decision selection (stage 3) and action implementation (stage 4). The two stages of diagnostic aiding support the first two levels of Endsley's model of SA, respectively (Horrey, Wickens, Strauss, Kirlik, & Stewart, 2009). Horrey et al. (2009) described a model in which information acquisition (stage 1) leads to information analysis (stage 2). Information acquisition (stage 1) is linked to level 1 SA, perception of elements, by sensation, perception, and attention. Information analysis (stage 2) is linked to level 2 SA, comprehension of the situation, by cognition, integration, and working memory.

Robots perform the information acquisition (stage 1) when they gather relevant information through their sensors. Robots perform information analysis (stage 2), when they integrate multiple pieces of sensor data or when they integrate sensor data with previously stored or externally provided information. Thus, information acquisition (stage 1) is a precursor to information analysis (stage 2), and a robot that performs both stages operates at a higher level of automation than one that only performs information acquisition (stage 1). In other words, diagnostic aiding is the automation of information processing, and the two stages of diagnostic aiding can be expected to support the corresponding first two levels of SA.

Effects of Automation on Performance

Moderated effects of automation. Examination of the research on diagnostic aiding, and on the general case of automation, revealed that while diagnostic aiding is generally effective in both traditional systems such as aviation (Rudisill, 2000), as well as robot-specific applications (Goodrich et al., 2007), diagnostic aiding can be detrimental (Kaber & Endsley, 2004; Ruff, Narayanan, and Draper, 2002; Yeh, Wickens, & Seagull, 1999). The conditions under which diagnostic aiding is detrimental are poorly understood, however. This suggests that, although automation is a helpful technology, the relationship is moderated by other constructs. Evidence for the existence of moderators is discussed next.

Effects on SA and performance. In contrast to workload, studies do not show a unilateral benefit for higher levels of automation on performance and SA. Taking a broad view across automation research, the results are inconsistent. The literature's equivocal findings describing the effects of level of automation on SA (Kaber & Endsley, 2004) may have been due to an operationalization of level of automation that confounded the level (that is, the amount of automation involvement) with what is automated.

An illustration of this is the difference between *management by consent* and *management by exception*. In management by consent, operators are required to respond before the robot takes action. Conversely, in management by exception, the robot will perform its programmed action unless the operator intervenes. Under the Parasuraman et al. (2000) 10-level taxonomy of levels of automation (see Table 1), this is the difference between automation at level 5 and automation somewhere between levels 6 and 7. Ruff, Narayanan, and Draper (2002) found performance of remotely operated vehicles to be better in a management-by-consent scenario, in which the operator was required to approve a robot's decision. If automation was unilaterally better, performance should have been higher in the management-by-exception condition. However, participants in the management-by-exception condition had lower performance. Jentsch et al. (2012) found the opposite: performance was better at higher levels of reliable automation. Rehfeld (2006) found higher performance and SA at lower levels of automation under Endsley and Kaber's level of automation taxonomy (1999). Chen, Barnes, and Harper-Sciarini (2011) highlighted this discrepancy and suggested the existence of mediators. By taking a more nuanced approach to level of automation (by manipulating *what* as well as *when*) and by restricting the investigation to automation of the information processing that occurs before decisions and actions are made, these mediators will be explored.

As previously discussed, the *what* of automation has been modeled by Parasuraman et al. (2000). However, much of the applied literature has taken a *how much* approach to measuring and manipulating the level of automation. Although it is understandably easier to manipulate the presence or functionality of an entire system, research needs to specify the stage as well as the amount of automation. Horrey and Wickens (2001) adapted this approach and found that both information acquisition (stage 1 diagnostic aiding) and information analysis (stage 2 diagnostic aiding) led to better performance than an unaided condition on a battlefield simulation task, with the information analysis (stage 2) aid leading to a greater reduction in errors compared with the information acquisition (stage 1) aid. The authors suggested that the removal of cognitive integration by the information analysis (stage 2) aid reduced the cognitive demands on the operator, leading to superior performance. However, memory probe questions suggested that relevant items were processed more deeply with information acquisition (stage 1) diagnostic aiding (Horrey & Wickens, 2001).

The addition of information analysis (stage 2) automation to information acquisition (stage 1) automation has been shown to have a greater effect on decisions than information acquisition (stage 1) alone (Dexter, Willemsen-Dunlap, & Lee, 2007). In a study of anesthesiologists, nurses, and hospital housekeepers, operating room management information was presented as either a command display, which provided recommendations (stage 2), or a status display, which made decision-relevant information available (stage 1). When making decisions in subsequent scenarios, participants without either type of aid performed less accurately than random chance. Decision making, both a cognitive outcome and a performance measure, was improved only by the command display (status displays did not have a significant effect on decision accuracy). Further, incorrect command displays had greater costs associated with them for trust, and users were more likely to follow erroneous recommendations that did not affect safety. From this, Dexter et al. (2007) concluded that command displays are preferable but carry additional costs when their recommendations are incorrect.

The literature offers support for diagnostic aiding as an effective intervention for reducing workload, increasing SA, and supporting performance. Further, I have provided support for the stages of diagnostic aiding (Wickens & Dixon, 2007) to be used to specify what task is automated, with information analysis (stage 2) being a higher level of automation than information acquisition (stage 1). However, inconsistent findings suggest the presence of additional complexity in this relationship.

Hypothesis 1: Under perfect reliability, diagnostic aiding that performs acquisition and analysis (stage 2) will lead to better operator SA than one that performs acquisition alone (stage 1; a simple effect).

Hypothesis 1a: Under perfect reliability, diagnostic aiding that performs acquisition only (stage 1) will lead to better level 1 SA but not level 2 SA (a simple effect).

Hypothesis 1b: Under perfect reliability, diagnostic aiding that performs acquisition and analysis (stage 2) will lead to better level 1 SA and level 2 SA (simple effects).

Two hypothesized moderators of this relationship, reliability and task complexity, will be explored next.

Reliability

Global Effects of Reliability

Reliability is the effective performance of an automated system. Unreliability, then, is the inability of an automated system to perform a task as intended by the designer of the system. The present research aimed to see if SA and performance are affected by reliability in the same way, and differently at different levels of unaided human performance. As the reliability of automation falls, human performance declines; operators may have difficulty compensating for degraded information (Wickens & Dixon, 2007). Dixon, Wickens, & McCarley (2006) found that performance was better in a non-automated condition than in either miss-prone or false alarm prone automation. Madhavan and Phillips (2010) found that participants using a 90% reliable decision aid in an X-ray baggage-screening task achieved more hits and fewer false alarms than those using a 70% reliable decision aid.

Wickens and Dixon (2007) found evidence across studies that the relationship between reliability and reliance is affected by task demand. Even if operators are aware of the true reliability of the system, they may continue to rely upon imperfect automation in order to preserve cognitive resources. The focus of the present study was not on the highest levels of task demand in which the operator can do little but blindly follow the automation. Instead, unreliability may affect global performance through the automation's inability to contribute accurate information.

Trust and reliance. Trust and reliance are two mediators of the reliability-performance relationship that have been examined extensively. These two constructs are related, as trust has been found to affect performance through reliance (Parasuraman & Riley, 1997). Trust is an affective state (Lee & Moray, 1992) that has been defined as "the reliance by one agent that actions prejudicial to the well being of that agent will not be undertaken by influential others" (Oleson, Billings, Kocsis, Chen, & Hancock, 2011, p. 176). Several studies examining trust (e.g., May, 1993; Oakley, Mouloua, & Hancock, 2003) have found that operator detection of automation failures worsens as reliability improved. In these studies, operators were responsible for monitoring automation to determine if it was functioning properly. The goal of the research was to determine how operators' perceptions of the reliability of the system drive their behavior, which has been supported in subsequent research (Chen & Barnes, 2012; Madhavan & Phillips, 2010; Wickens & Dixon, 2007). While findings may appear to suggest that increased reliability reduces an operator's ability to detect automation failure states, they share a critical difference from the present research; participants were not provided with the specific reliability of the automation. In May's study (1993), participants were not told of the specific reliability of the system so as to maximize their trust in the system. Additionally, the automation-monitoring task was presented to the participants as a secondary task. If participants did not understand the true reliability of the system, they may have inappropriately relied upon (or completely disregarded) the automation, which was the effect the researchers aimed to find. Thus, these outcomes may be best explained as inappropriate trust, which led to inappropriate reliance. This is an important, but different, issue from operator strategy selection given automation with known unreliability.

While understanding of these mediators has guided engineering and provided insight into performance outcomes, little work has been done to examine the effects of information processing unreliability on the operator's ability to build and maintain SA, especially when the operator is aware of the true reliability of the system *a priori*. In other words, the performance outcomes have been studied more than the information processing mediators. Consequently, there is a lack of understanding of the cognitive mediators of the relationship between automation reliability and performance. In their meta-analysis, Wickens, Li, Santamaria, Sebok, and Sarter (2010) found only five studies that investigated the level of automation on performance and SA under conditions of automation. They noted, “The situation awareness data are too few to create any well defined trend” (p. 391). This need was addressed by the current empirical study, which studied three levels of system reliability and predicted:

Hypothesis 2: Operator SA will be higher at higher levels of robot reliability (a main effect).

Reliability at Each Level of Decision Aiding

In a study examining the impact of incorrect information on mission-critical decisions under time pressure, Ehrlich et al. (2011) found no effects on performance of recommendations (analogous to information analysis [stage 2]) versus recommendations along with justification (analogous to information analysis and acquisition [stage 1 and stage 2]). Both forms of decision aiding were helpful when accurate and detrimental when inaccurate. For inaccurate recommendations and justifications, the authors concluded that participants could have been biased towards the recommendations even when justifications were included. In Ehrlich et al., the effects of the justifications (a form of level 1 SA) alone were not investigated. This left the question of whether or not operators made better decisions when the incorrect data was provided directly (as was the case with justifications) or provided in an integrated form (as was the case with recommendations).

The literature suggests that while unreliable information negatively impacts performance, the effect is much stronger for information analysis (stage 2) automation than for information acquisition (stage 1) automation (Rovira, McGarry, & Parasuraman, 2007). Sarter and Schroeder (2001) found that a diagnostic aid that provided recommendations (information analysis [stage 2]), rather than status information (information acquisition [stage 1]), had a greater performance cost when the automation was not reliable. Rovira and colleagues (2007) found that unreliability degraded operator accuracy at three levels of increasingly automated information analysis (stage 2). Unreliability did not have a significant effect on accuracy in the information acquisition (stage 1) condition, however.

Crocoll and Coury (1990) found a similar pattern of results with an airplane identification task. In a study manipulating status (information acquisition [stage 1]) and recommendation (information analysis [stage 2]) information, the group receiving only status information was the least affected by inaccuracy in the automation. In line with this finding, Skitka, Mosier, and Burdick (1999) found that introduction of imperfect automation that monitored system state led to an increase in missed events.

One explanation of this effect is automation bias (Cummings, 2004); it is a form of misuse in which erroneous, automated recommendations are trusted, and conflicting information is disregarded. While this explains an operator's decision to rely upon automation or do without, it does not explain differences in attention or cognitive processing. In other words, it explains affect and behavior, but not SA.

Parasuraman and Wickens provided a cognitive explanation for why lower stages of diagnostic aiding may lead to better SA: "The user must continue to generate the values for the different courses of action. As a result, users may be more aware of the consequences of the choice and of the possibility that the choice may be incorrect because of a faulty automated diagnosis” (2008, p. 514). This may explain the empirical findings of Horrey and Wickens (2001). When operators perform information analysis (stage 2), they perform additional processing that may keep them “in the loop”. Consequently, the operators’ information analysis (stage 2) should lead to better SA during robot unreliability.

Galster, Bolia, and Parasuraman (2002) found that performance on a target detection task improved when an information status cue was added, even though this cue was not perfectly reliable. When a higher level of aiding was added in the form of decision suggestion, performance was not improved, unless the information status cue was also included. This suggests that under conditions of unreliability, operators may be able to recover from erroneous information provided by information acquisition (stage 1) automation more easily than from information analysis (stage 2) automation. In summary, when reliability is limited, access to lower-level data can help an operator to remain in the loop (Johnson, Saboe, Prewett, Coovert, & Elliott, 2009).

Hypothesis 3: Under imperfect reliability, automation of information analysis (stage 2) will lead to lower SA unless the operator would otherwise have poor SA without the aid (an interaction effect).

Hypothesis 3a: Under imperfect reliability, automation of information analysis (stage 2) will lead to lower SA when the operator would otherwise have moderate (50%) SA without the aid (a simple effect).

Hypothesis 3b: Under imperfect reliability, automation of information analysis (stage 2) will lead to lower SA when the operator would otherwise have good (90%) SA without the aid (a simple effect).

Potential Unaided Situation Awareness

Definition

In a recent review of levels of automation and automation reliability literature, Johnson et al. (2009) found evidence for mediators in the relationship between automation reliability and performance. In some tasks, imperfect automation had a severe, negative impact on performance. In others, the impact was minimal. Johnson et al. concluded that the differences were due to the availability of other, non-automation strategies for completing the task.

The availability of these strategies should be considered from an operator-centric perspective. That is, it is less important whether or not an alternative task completion strategy is available than whether the operator is aware of, and able to utilize, the strategy. Thus, an operator's SA in an unaided task captures individual cognitive performance. Potential unaided SA is the relevant knowledge that is available and held by an operator in the absence of diagnostic aiding. It can be thought of as performance on the cognitive aspects of the task independent of any automation. It is not paradoxical to distinguish potential unaided SA from the SA of an operator using a diagnostic aid. By manipulating this construct, I evaluated the impact of three scenarios spanning the realistic range of potential unaided SA: one in which the operator can independently obtain nearly all information needed to perform the task ("good [90%] potential unaided SA"), one in which the operator cannot obtain any substantial amount of the information needed to perform the task (“poor [10%] potential unaided SA”), and one in which half the information is available to the operator (“moderate [50%] potential unaided SA”), which reflects an even distribution of information available to operator and automation.

Diagnostic Aiding and Potential Unaided SA

One case is unexplored; what are the costs and benefits of diagnostic aiding when operators would otherwise have poor SA? Based on the literature, this would occur when task complexity (and thus workload) is very high. By manipulating characteristics of the task, unaided situation awareness can be manipulated. If operators demonstrate exceptionally low SA in the absence of a diagnostic aid, then the presence of even a fairly unreliable aid should be beneficial. Although studies have included difficult tasks, no research has been conducted investigating operator use of unreliable automation in tasks while manipulating potential unaided SA. Johnson et al. (2009) pointed out that providing operators with lower stages of diagnostic aiding becomes problematic as workload increases with the amount of data to be managed. Empirical data supports this claim (Rovira et al., 2007). Thus, a non-linear relationship may exist.

Madhavan and Wiegmann (2007) suggested that operators behave differently in the face of automation unreliability, depending on their own ability to perform the task unaided. Specifically, their easy-errors hypothesis says that errors on tasks that operators could perform themselves undermine trust and lead to disuse. Conversely, when the automation performs more reliably than the operators’ unaided performance level, operators will tend to rely upon the automation more than their own diagnoses. This behavior is adaptive (Dzindolet et al., 2003), as long as the operator is aware of the true reliability of a system.

In line with this hypothesis, Lee and Moray's (1992) results suggest that operators' automation use is affected by their ability to perform the task unaided. Both performance and trust were measured as outcomes from automation failure a task simulating a juice pasteurization factory. The researchers found that both trust and performance were negatively affected by automation failure, with performance recovering faster than trust. At the same time, participants tended to increase their use of automation as operators dealt with the fault. Lee and Moray suggested that operators' confidence in their own abilities affected automation use more than trust in the automation.

Additionally, there is evidence that operators adjust their behavior based on their perceptions of system reliability (Chen & Barnes, 2012; Madhavan & Phillips, 2010). This leads to two implications: (a) operators' knowledge of how automation fails will affect compliance with the automation, and (b) in difficult tasks, operators may (appropriately) rely on imperfect automation. Again, this behavior may be an appropriate strategy to dealing with unreliability across a system.

Hypothesis 4: When the operator would otherwise have poor SA, automation of information analysis (stage 2) will lead to better SA (an interaction effect).

Summary and Hypotheses

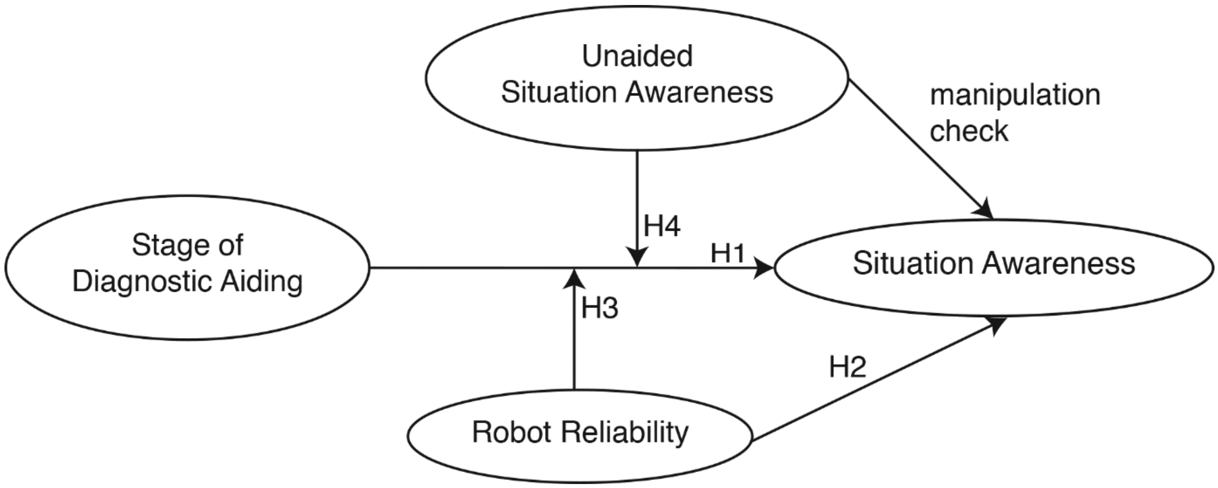


Figure 2. Research model with hypotheses.

As previously stated, this research tested for three effects: reliability of robot sensing, stage of diagnostic aiding, and potential unaided SA. The literature suggests that high robot reliability is beneficial. However, few robots are perfectly reliable, and even unreliable robots may be useful under some circumstances. Through this research, I aimed to find the conditions under which unreliable robots can still contribute to the SA of an operator. SA is an important determinant of performance across complex systems. However, under high levels of autonomy and low levels of reliability, operators may lose SA as they become disconnected from critical elements in the environment. A list of expected confounding variables, and the strategies that will be used to measure or exclude their effects is presented in Appendix A.

To review, the following hypotheses were tested as part of my research model (see Figure 2):

Hypothesis 1: Under perfect reliability, diagnostic aiding that performs acquisition and analysis (stage 2) will lead to better operator SA than one that performs acquisition alone (stage 1; a simple effect).

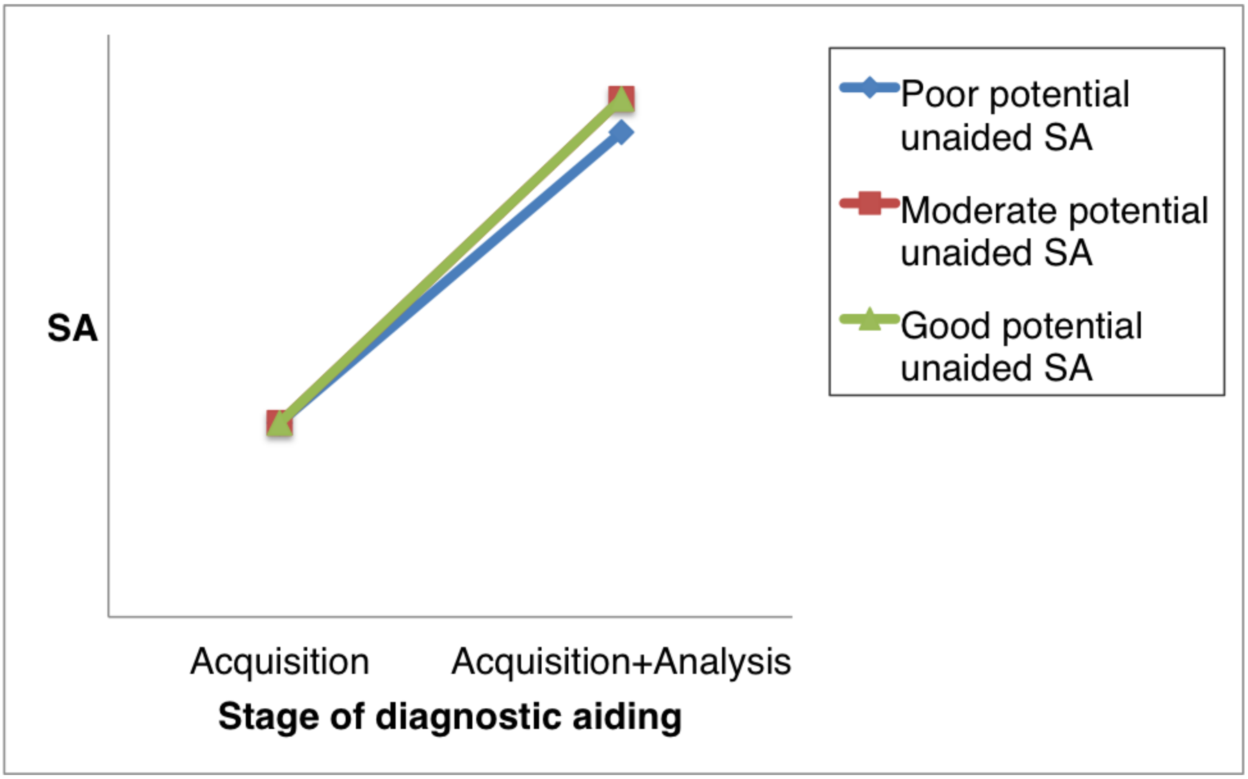


Figure 3. Hypothesized effects of level of automation on SA under perfect reliability (Hypothesis 1).

Hypothesis 1a: Under perfect reliability, diagnostic aiding that performs acquisition only (stage 1) will lead to better level 1 SA but not level 2 SA (a simple effect).

Hypothesis 1b: Under perfect reliability, diagnostic aiding that performs acquisition and analysis (stage 2) will lead to better level 1 SA and level 2 SA (simple effects).

Hypothesis 2: Operator SA will be higher at higher levels of robot reliability (a main effect).

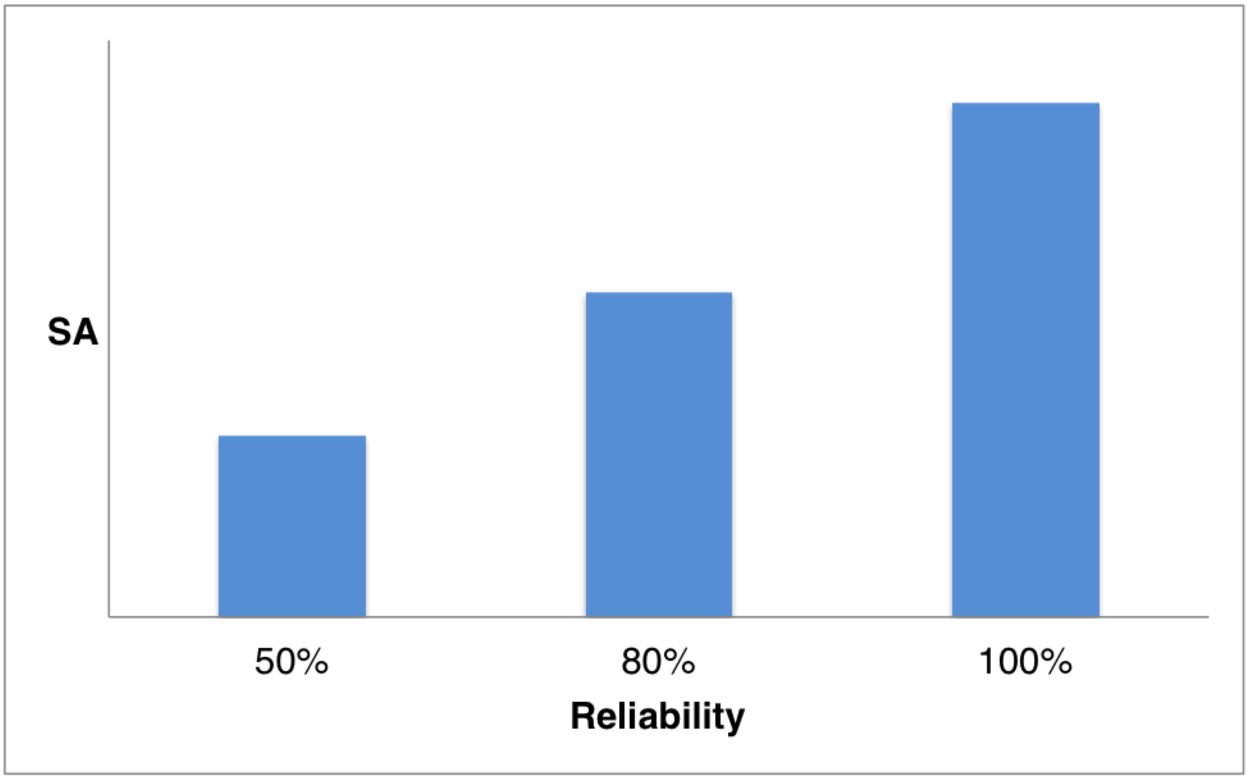


Figure 4. Hypothesized effect of reliability on SA (Hypothesis 2).

Hypothesis 3: Under imperfect reliability, automation of information analysis (stage 2) will lead to lower SA unless the operator would otherwise have poor (10%) SA without the aid (an interaction effect)

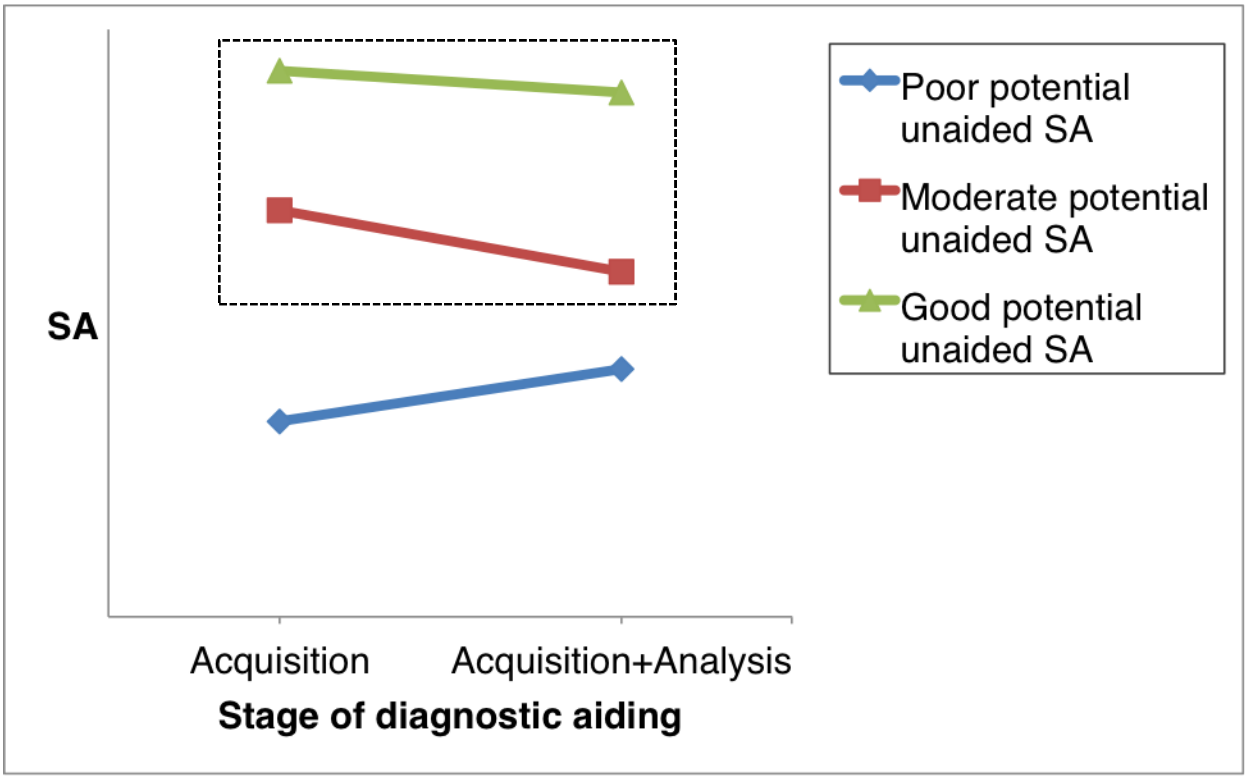


Figure 5. Hypothesized effects of level of diagnostic aiding under good (90%) and moderate (50%) potential unaided SA (dashed area) and imperfect reliability (Hypothesis 3).

Hypothesis 3a: Under imperfect reliability, automation of information analysis (stage 2) will lead to lower SA when the operator would otherwise have moderate (50%) SA without the aid (a simple effect).

Hypothesis 3b: Under imperfect reliability, automation of information analysis (stage 2) will lead to lower SA when the operator would otherwise have good (90%) SA without the aid (a simple effect).

Hypothesis 4: When the operator would otherwise have poor (10%) SA, automation of information analysis (stage 2) will lead to better SA, even when the diagnostic aid is unreliable (an interaction effect).

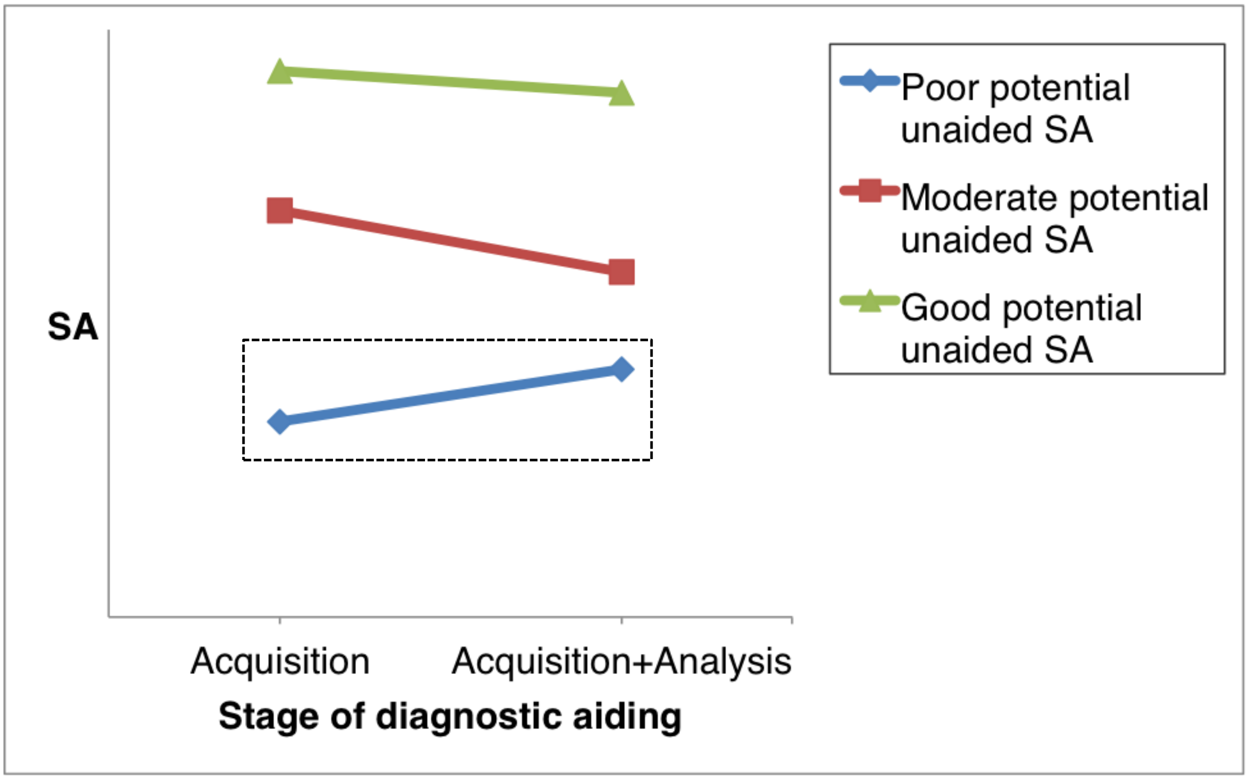


Figure 6. Hypothesized effects of level of diagnostic aiding under poor (10%) potential unaided SA (dashed area) and imperfect reliability (Hypothesis 4).

CHAPTER THREE: METHODOLOGY

Task Setting

The experimental scenario was a cordon and search task. In military operations, cordon and search is "conducted to seal (cordon) off an area in order to search it for persons or things such as items, intelligence data, or answers to PIR (primary intelligence requirements)" (United States Army, 2009, pp. 5-8). The management of relevant mission information is a critical part of operational safety and mission effectiveness in cordon and search, as well as in other military operations (United States Army, 2006).

According to the Army, "every Soldier is a sensor" (United States Army, 2008, pp. 9-1). This concept appropriately extends to robots when mission knowledge may be distributed across human and robot agents, with each agent having unique and complementary information. The current experimental scenario was used as an example of military operations requiring management of dynamic information distributed across agents.

SA in Cordon and Search

Because the definition and content of SA, and ultimately its measurement, are inherently tied to the task (Schuster, Keebler, Zuniga, & Jentsch, 2012), the mission goal must carry clearly definable knowledge requirements. In cordon and search, the elements in the environment include the potential targets of the search (for example, friendly or hostile individuals) and their relevance to mission goals (for example, identification of hostile individuals in a room clearing task). Based on this, in a simple collaborative cordon and search mission with two entities (human and robot), SA can be operationalized as follows:

Level 1 SA (perception): Knowledge of the existence of individuals, their locations, and/or their identifying characteristics.

Level 2 SA (comprehension): Knowledge of whether the individuals perceived at level 1 are friendly or hostile.

Level 3 SA (projection): Knowledge of the future states (for example, future locations) of individuals understood at level 2.

Because all individuals in a scenario are relevant to the participant's mission goal, whether the robot or participant can sense them directly, this operationalization of SA included all individuals in the building. The difficulty of maintaining SA in complex, yet highly automated, environments like aviation is a demonstrated example of the out-of-the-loop performance problem (Wickens, 2002). To examine this problem, this study applied information management requirements to a military operation. This task is relevant because object detection and biometric identification (e.g., facial recognition) are tasks robots perform with imperfect reliability in real-world environments. However, over time, automation performance in this task in the field can be expected to improve. In the current study, cordon and search served as a mission where a robot could be implemented at various levels of decision aiding while allowing for clear manipulation of reliability and potential unaided SA.

Design

The current study used a 2 (information acquisition [stage 1] vs. information acquisition with analysis [stage 2]) x 3 (60% reliable, 80% reliable, 100% reliable) x 3 (poor [10%] potential unaided SA, moderate [50%] potential unaided SA, good [90%] potential unaided SA) mixed factorial design (see Table 2). Stage of diagnostic aiding and potential unaided SA were within-subjects independent variables to increase sensitivity. A literature review suggested a low potential for interactions between individual differences and these two IVs (see Appendix C). Because of the potential for carry-over effects (Chen & Barnes, 2012; Madhavan & Phillips, 2010), robot reliability was a between-subjects independent variable. The dependent variable was SA of the participant.

Table 2
Experimental design with independent variables

IV 1: Stage of diagnostic aiding (within subjects)	IV 2: Robot reliability (between subjects)	IV 3: Potential unaided SA (within subjects)
Information acquisition (stage 1)	100%	Poor (10%) Moderate (50%) Good (90%)
	80%	Poor (10%) Moderate (50%) Good (90%)
	60%	Poor (10%) Moderate (50%) Good (90%)
Information acquisition with analysis (stage 2)	100%	Poor (10%) Moderate (50%) Good (90%)
	60%	Poor (10%) Moderate (50%) Good (90%)
	60%	Poor (10%) Moderate (50%) Good (90%)

Participants

Participants were 64 students recruited from the University of Central Florida's psychology undergraduate participant pool using the SONA Systems web site. Participants received course credit in exchange for participation. The research protocol was submitted to the University of Central Florida Institutional Review Board (IRB) for approval and to the United States Army for headquarters-level administrative review prior to the start of data collection.

Of the 64 participants, 17 were excluded from the analysis. Of these, two participants chose to end the study early, one participant was dismissed due to technical problems with the apparatus, and one participant was dismissed after falling asleep. The remaining 13 were excluded because they were not presented with one entire SART or objective SA assessment questionnaire due to a software malfunction. All subsequent analyses were performed following the removal of these participants, resulting in a sample of 47 participants.

A power analysis was conducted (Faul, Erdfelder, Buchner, & Lang, 2009) to determine the number of participants needed to detect a medium effect (f = .2) at an alpha level of .05 with a power level of .8. With each participant completing 12 trials (two instances of each combination of the within-subjects variables), the estimated sample size was 36 participants, suggesting that the final sample provided sufficient power.

The sample included 20 males and 27 females ranging in age from 18 to 47 years (M = 20.89, SD = 4.32). Refer to Table 8 for the frequencies of each gender by between-subjects condition. All participants in the sample reported that they were native speakers of English, did not have color-deficient vision, and did not have prior military experience.

Materials

Mission Environment

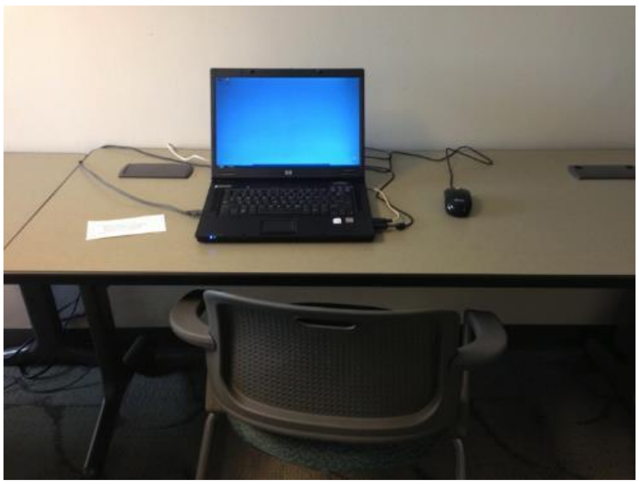


Figure 7. Experimental apparatus consisting of a laptop computer, mouse, and participant rule set reference card.

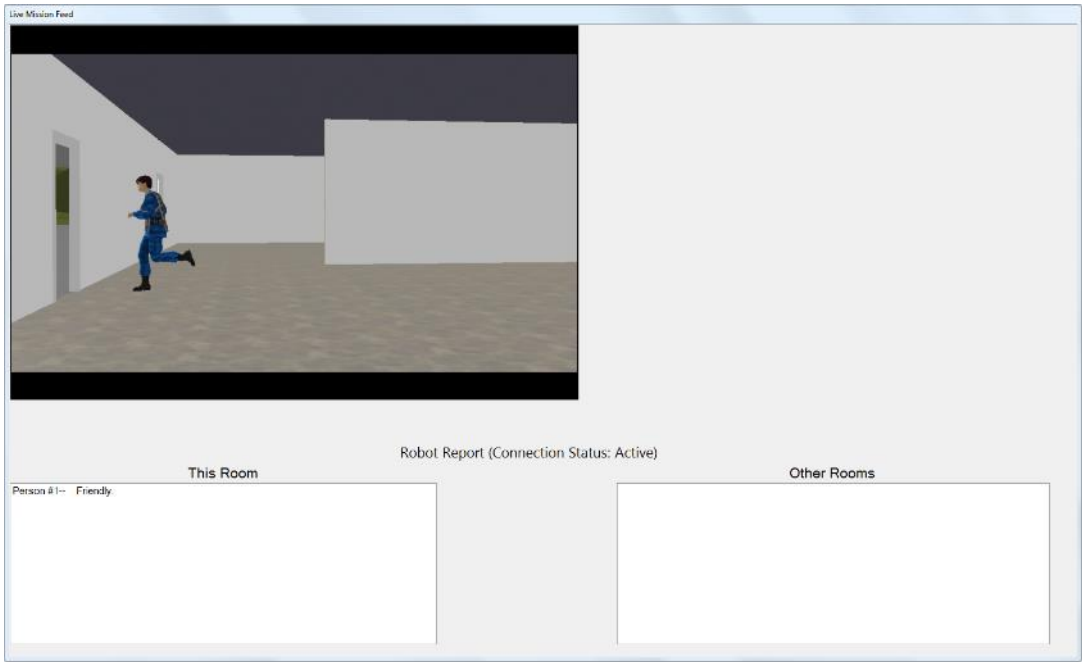


Figure 8. Screen capture showing the mission display with first-person view and robot-provided information.

A non-stereoscopic, three-dimensional virtual environment was used to present a first-person view of the mission environment on a computer monitor. A discrete event simulation of each mission was developed *a priori*. In each mission, a number of friendly and hostile individuals entered and left a building through multiple doorways. The robot and participant both monitored the building. The robot had the added ability of being able to see through interior, but not exterior, walls. Thus, the robot could aid the participant by seeing individuals that were outside of the participant's view. The participant and robot's positions were fixed. Within the simulation, a pre-rendered mission video 16.51 cm wide by 10.88 cm tall (approximately 15.4° by 10.19° degrees of visual angle) was displayed. The video showed the participant's perspective and showed individuals exiting and entering the building. During-mission and post-mission questionnaires were administered on the same computer.

The participants' goal was to determine how many people were inside the building at the present time and whether they were friendly or hostile individuals. To accomplish this goal, participants had to gather and integrate information from the robot.

Participants were trained on a rule set used to determine hostility. The four characteristics are listed in Table 3. The conditions under which an individual was hostile are listed in Table 4.

Table 3
Characteristics of individuals in the building

Characteristic	Possible values
Uniform color	Green, Red, Blue
Armed	Armed, Unarmed
Wearing a helmet	Helmet, No helmet
Running	Running, Walking

Table 4
Conditions under which individuals were determined to be hostile

Hostile (if and only if)
Green uniform and armed, or
Armed and wearing a helmet, or
Red uniform and running

Robot Team Member

The participant worked with a simulated robot (see Figure 9) to complete the shared goal of identification of friendly and hostile individuals. Although simulated, the robot could be thought of as performing an equivalent task to the human but with the added ability to see through the interior walls of the current building. The robot was stationary, and it had one or two features, depending on the condition: person recognition with feature identification (information acquisition [stage 1]) and person identification (information analysis [stage 2]).

Figure 9. Photo of the robot team member presented to the participant. The robot monitored for the presence of individuals and could see through walls.

Stage of Diagnostic Aiding Manipulation

Information acquisition (stage 1) condition. In the information acquisition (stage 1) condition, the robot monitored the door and updated its status display whenever an individual was detected. The robot was able to sense physical characteristics about the individual (e.g., uniform color, wearing a helmet, armed or not) and report these characteristics to the participant in a continuously updated list of individuals. The communication channel between the robot and the human was perfectly reliable, and the robot would always report the information it sensed. That is, when the robot sensed an individual, it always communicated that information to the participant, and the communication was accurate with respect to the robot's sensing.

The robot was not always accurate in its sensing of the physical characteristics, depending on the condition. The only error committed by the robot was in its sensing. Specifically, the errors committed by the robot were errors of classification only. That is, the robot never missed an individual who was present nor falsely reported the presence of an individual who was not there. For an individual who was present, however, the robot could report one of their characteristics, uniform color, incorrectly.

The reliability of the participants' robot was explained in a tutorial video (see Appendix I for the video content and Appendix J for the script). This video provided the exact error rate of the robot, explained the kinds of mistakes the robot could make, and offered examples to illustrate the impact of the error rate. It was explained that mistakes were independent; a series of mistakes did not make a future mistake less likely, and vice-versa.

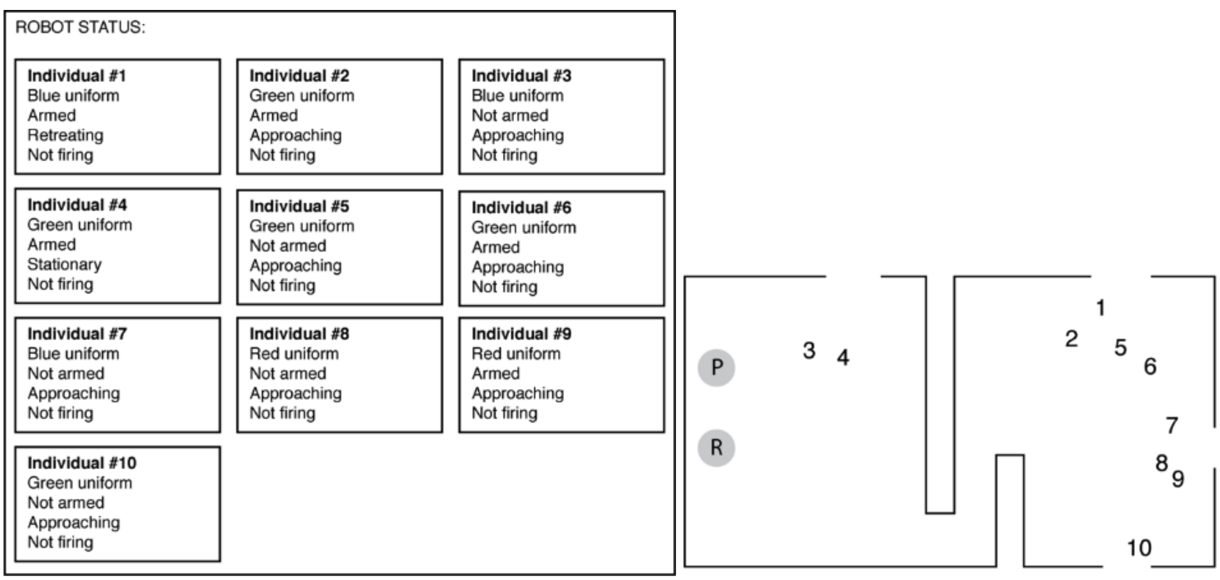


Figure 10. Demonstration of output from the information acquisition (stage 1) robot. In the experiment, this data was updated live and presented in a list. Note, P = participant, R = robot.

Information acquisition with analysis (stage 2) condition. In the information acquisition with analysis (stage 2), the robot performed information acquisition (stage 1) as described in the previous step. However, before reporting to the participant, the robot used the physical characteristics to make a determination about the hostility of the individual. In addition to their own first-person viewpoint, participants were only shown the friendly or hostile status as reported by the robot (see Figure 11). The information analysis (stage 2) performed by the robot was deterministic (i.e., any combination of physical characteristics can be used to classify the individual as friendly or hostile) and perfectly reliable.

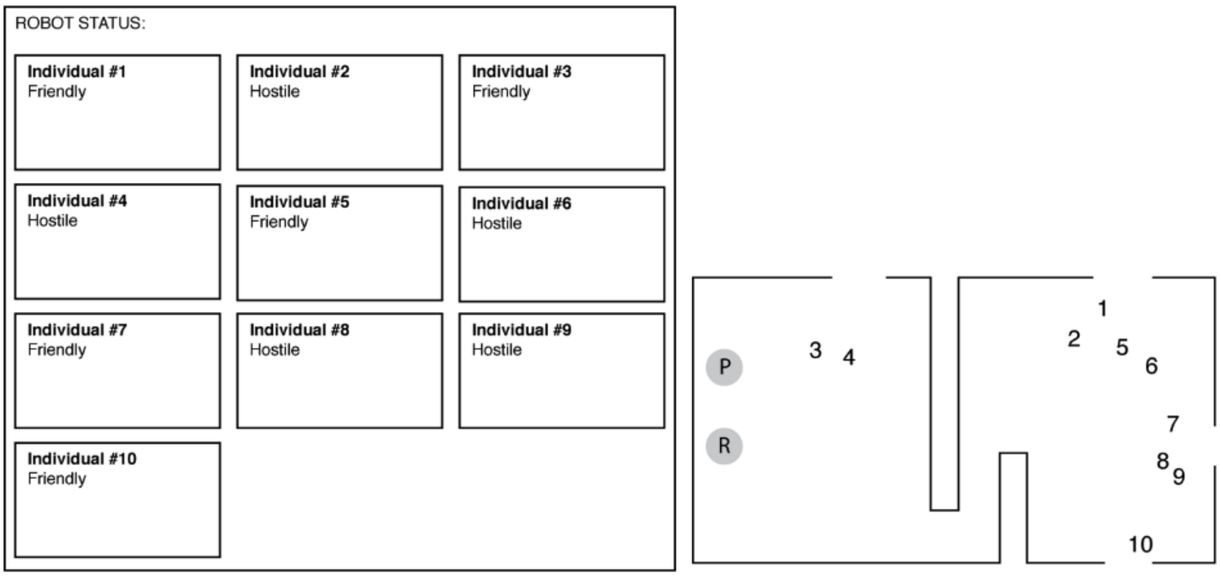


Figure 11. Example of output from the information acquisition with analysis (stage 2) robot as viewed by the participant. It was updated in real time. Note, P = participant, R = robot.

Robot Reliability Measure

The reliability of the robot's information acquisition (stage 1) was manipulated. Reliability was defined as the likelihood that the robot would provide information that would lead to a correct conclusion. In order to provide a consistent backstory to participants across the stages of diagnostic aiding, reliability was manipulated at the feature level; errors took the form of incorrect sensing of a single feature. However, the only errors that were made by the robot were errors that led to an incorrect conclusion (i.e., the incorrect determination of hostility). In this way, the consequence of each robot error was held constant across conditions. An annotated example showing how the robot's errors translated into the participant's display is shown in Appendix E.

Errors were presented to the participant as independent events. Each time the robot sensed an individual, there was a 40%, 20%, or 0% chance of the robot sensing the wrong uniform color, leading it to misinterpret the individual's hostility. For example, in the 60% reliability condition, if the robot had just made two errors, the probability that it would make an error on the next individual remained at 60%. However, to control for the number of errors experienced by participants, errors were predetermined so that the participant experienced the correct percentage of robot errors across the mission.

Three levels of reliability were selected to span across the range at which diagnostic aiding may be implemented and could be useful. The lowest level (60%) was at the bottom end of the 95% confidence interval for the minimum reliability level identified by Wickens and Dixon (2007) at which diagnostic aiding is still useful. Wickens and Dixon's finding was based on a meta-analysis, which showed a drop-off of performance to below chance near 71% reliability. Importantly, the lowest level of reliability used in this study (60%) was above chance, so using the robot should have provided a benefit if the participant could not otherwise obtain the same information.

Low (60%) reliability. In the low reliability condition, the robot made errors in its perception of uniform color. Incorrect feature perception led to incorrect determination of hostility in 60% of the individuals reported on by the robot.

Moderate (80%) reliability. In the moderate reliability condition, the robot made errors in its perception of uniform color. Incorrect feature perception led to incorrect determination of hostility in 20% of the individuals reported on by the robot.

High (100%) reliability. In the high reliability condition, the robot was consistently accurate.

Potential Unaided SA Manipulation

The potential unaided SA of the participant without the robot was also manipulated. The participant's SA was quantified as the completeness and correctness of their knowledge of individuals in the building at a particular point in time. Thus, potential unaided SA was operationally defined as the proportion of people in the building that the participant could see and identify if the robot were not present. Because participants could only see within their current room in the building, the participant's unaided SA was limited to the proportion of people entering and existing within the current room. As a manipulation check, the potential unaided SA manipulation was tested to see if it predicted the participant's SA.

In the current study, one explanation for any effects of "good (90%)" potential unaided SA might have been that the robot provided confirmation of information already known by the participant. In the "moderate (50%)" and "poor (10%)" potential unaided SA conditions, the diagnostic aiding provided unique information not otherwise known by the human. This was done to mirror a real world scenario in which the robot's primary purpose is to act as a uniquely contributing sensor. Any observed effects may or may not apply to robots that exist primarily to offer confirmation of known information.

To ensure that the robot acted as a uniquely contributing sensor while preventing complete separation of the robot and participant's tasks, individuals that were directly visible to participants moved in one of three ways. In the first (see Figure 12), individuals entered in view of the participant, remained in the visible room, and then exited in view of the participant. In the second (see Figure 13), individuals entered in view of the participant but walked into another room before leaving the building. Finally, in the third (see Figure 14), individuals entered in another room but walked into the visible room before leaving the building. Visible individuals were randomly assigned one of these three movement paths, and the frequency of each movement path was approximately equal within each mission.

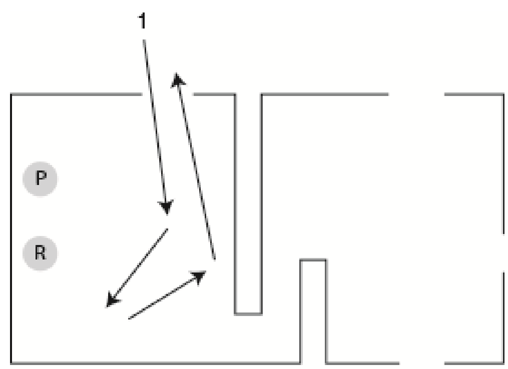


Figure 12. Arrows represent the movement of individual 1. Approximately one-third of visible individuals were always visible to the participant.

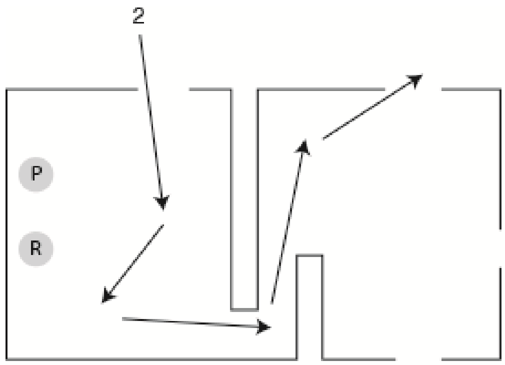


Figure 13. Arrows represent the movement of individual 2. Other individuals entered the building in view of the participant but exited out of the participant's view. The paths are examples only; each individual had a unique path.

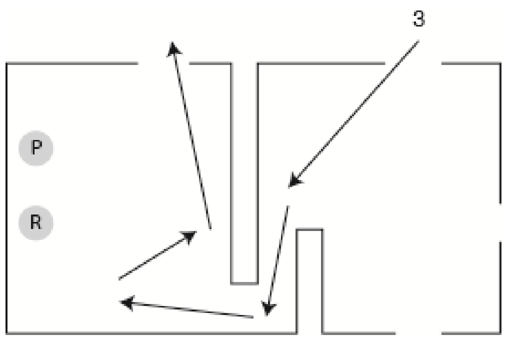


Figure 14. Arrows represent the movement of individual 3. Approximately one-third of visible individuals entered the building out of the participant’s view but moved into the visible room before exiting.

Poor (10%) potential unaided SA. In the “poor (10%)” potential unaided SA condition, the participant was able to view 2 of the 20 people entering and exiting due to the participant’s position and the layout of the building. That is, 10% of the relevant events (individuals entering or leaving) occurred in the participant’s field of view (see examples in Appendix F).

Moderate (50%) potential unaided SA. In the “moderate (50%)” potential unaided SA condition, the participant was able to view 10 of the 20 people entering and exiting due to the participant's position and the layout of the building. That is, half (50%) of the relevant events (individuals entering or leaving) occurred in the participant's field of view.

Good (90%) potential unaided SA. In the "good (90%)" potential unaided SA condition, the participant was able to see 18 of the 20 people in the building as they entered or exited, or 90% of the individuals.

Biographical Data Form

Participants were asked their age, gender, visual acuity, ability to detect color, and prior military experience.

Informed Consent and Debrief Form

Participants participated in an informed consent process as required by the IRB (see Appendix G). At the conclusion of the study, participants were provided with a debrief form describing the purpose of the study and the manipulations (see Appendix H).

Measures

Objective Situation Awareness Assessment

An SA measure was developed based on the SAGAT method described by Endsley (2000a). The questions used in the measure were based on the current state of the mission and were objective in that they had a single correct answer. Two questions were asked, each classified as measuring one of Endsley's three levels of SA (perception or comprehension). Because the mission goals did not involve prediction of future states, no questions were asked at Endsley's third level of SA, projection.

To assess level 1 SA, participants were asked for the uniform color of the last individual who entered the building at either entrance. To assess level 2 SA, participants were asked for the friendly or hostile status of the last individual who entered the building at either entrance. Answers to these questions were scored as correct or incorrect. The percentage of correct responses for each item resulted in two measures, uniform color accuracy and status accuracy. Participants were told that these questions would be asked.

Situation Awareness Rating Technique (SART)

The Situation Awareness Rating Technique (SART) is a self-report measure of SA to be administered after each mission. The SART consists of ten questions within three subscales: demand on attentional resources, supply of attentional resources, and understanding of the situation. Participants rated their SA on each of the ten dimensions using Likert-type items from 1 ("Low") to 7 ("High"). The mean response of each subscale was calculated, and these numbers were summed using the following formula: supply of attentional resources + understanding of the situation - demand on attentional resources = SART score. This followed the methodology of Endsley et al. (2000a).

Performance Measure

Participants were also asked to report, separately, the number of friendly and hostile individuals currently in the building. Because the goal of the participants was to maintain accurate counts throughout the mission, count accuracy was used as a performance measure. This measure scored as 1 for having both friendly and hostile counts correct and 0 if either were incorrect. The mean of four measurements during each mission was computed and used as the performance score for the mission.

Metacognitive Measures

At the conclusion of each mission, participants were asked to rate their own performance, the performance of the robot, and the performance of the human-robot team. Three items asked participants to rate their ability to identify individuals without the robot's help, their ability to identify individuals with the help of the robot, and the robot's ability to identify individuals. Each was measured using Likert-type items ranging from 1 ("Low") to 7 ("High").

Spatial Ability

Spatial ability was measured using the Guilford-Zimmerman Spatial Orientation test and scored as the number of items answered correctly.

Procedure

Table 5
Experiment timeline

Time (hours:minutes from start)	Event
0:00	Participant arrival; informed consent process and completion of biographical data
0:10	Spatial ability measure
0:20	Video introduction
0:32	Practice missions
0:38	Missions 1-12
1:52	Debrief; participant dismissal

Upon arrival, the experimenter presented the informed consent process to the participant. Next, the participant was seated at a computer workstation to complete the pre-task measures of biographical data and spatial ability. Following this, the participant was shown a video presentation explaining the study procedures, introducing the purpose and goals of the task, demonstrating the interface, and describing the robot. This video was the same for all participants with the exception of the portion that described the robot; it differed across conditions of robot reliability. The participant was told the percentage of the time that the robot would make errors, that the robot's errors would occur randomly, and that the errors would be limited to the classifications of uniform color. The participants were told that when they could not see individuals, they should rely upon the information provided by the robot. The participants were also presented with preferred strategies for accomplishing their mission goals (see Appendix J). The SAGAT procedure was explained.

Next, participants completed two two-minute minute practice missions plus 12 "live" missions lasting four and a half minutes each. In the first practice mission, the robot was not used so that participants could first gain familiarity with the task. For each mission, the participant was seated in front of a computer workstation. On this computer, the mission video was played and surveys were administered. After being asked if they were ready to begin, the participant was directed to click a button that started the mission timer and began playing the mission video. The mission video displayed the participant's view of a doorway and showed individuals entering and exiting the building through the doorway. Meanwhile, as explained to the participant beforehand, the robot team member was watching both the front and back doors and reporting people as they were detected. Below the mission video, on the same monitor, a chat window showed two lists of individuals identified by the robot. The leftmost list showed individuals in the room visible to the participant. The rightmost list showed individuals within the building but outside the room visible to the participant. When the robot detected an individual, it added the individual to the appropriate list. When an individual left the room, they were removed from the list. The lists included the characteristics of the individual (i.e., the individual's uniform color, whether or not they were armed, whether or not they wore a helmet, and if they were running or walking) in the information acquisition (stage 1) condition. In the information acquisition with analysis (stage 2) condition, the robot only reported the presence of a "friendly" or "hostile" individual.

Meanwhile, at four points during each mission, the mission window was minimized and the objective SA questions were displayed. Questions were asked at a randomly selected, but predetermined, time within 10 seconds after the first, second, third, and fourth minute of the mission. No time limit was provided for responses to the SA questions. After the participant responded to an SA question, they were presented with a button that allowed them to continue the mission. The participant's video feed ended after four and a half minutes, which did not include the time taken to respond to the objective SA questions. The SART was administered at the conclusion of each mission along with the metacognitive measures.

The mission procedure continued until all missions were completed. After the last mission finished, the study concluded. Participants were debriefed and dismissed.

CHAPTER FOUR: RESULTS

Statistical analyses were performed using SPSS version 20 with an alpha level of .05, unless otherwise stated.

Demographic Variables

Means, standard deviations, and intercorrelations for study variables are presented in Table 6. Two significant correlations indicated that more senior students tended to be older and that males tended to have higher spatial ability.

Table 6
Descriptive statistics and intercorrelations among study variables

Variable	*M*	SD	1	2	3	4	5
1. Age	20.89	4.32	–				
2. Year in school	2.51	1.04	.50**	–			
3. GPA	3.17	0.44	.51	-.05	–		
4. Spatial orientation	18.40	10.04	-.21	-.30*	.01	–	
5. Gender	0.57	0.50	-.08	.18	.07	-.49**	–

* $p < .05$. ** $p < .01$. Gender was coded as 0 = male, 1 = female.

Check of Random Assignment

To check the effectiveness of the random assignment, a series of one-way analyses of variance (ANOVAs) were conducted using demographic information as the dependent measure. These means are listed in

Table 7; differences in means across conditions were small and did not reach statistical significance, with the exception of the number of years in school; participants in the 80% reliability condition had, on average, been at the university less (M = 2.00 years, SE = 0.18) than those in the 60% reliability (M = 2.73 years, SE = 0.28, p = 0.05, d = 0.79) and the 100% reliability (M = 2.81 years, SE = 0.28, p = 0.03, d = 0.86) conditions. Although significant, these differences were comparatively small, and no further relationships with years in school were found through additional analyses.

Additionally, a two-way Pearson χ^2 test was computed to determine whether there was an association between the question, "Do you wear prescription glasses or contact lenses" and experimental condition. This test was not significant, p = .07. In all, these checks suggested that the random assignment was successful at distributing participants across conditions.

To avoid the effects of gender confounding the manipulation of reliability, the last 11 participants were randomly assigned to conditions based on their gender. A two-way Pearson χ^2 test was conducted to determine whether there was an association between gender and experimental group. This test was not statistically significant, p = .97. Table 8 lists sample sizes for gender by condition.

Table 7
Group means and standard deviation for demographic variables by condition

Variable	Overall *M* (*SD*)	60% Reliability *M* (*SD*)	80% Reliability *M* (*SD*)	100% Reliability *M* (*SD*)	*df*	*F*	*p*
Age	20.89 (4.32)	20.53 (2.20)	19.50 (0.97)	22.63 (6.80)	2, 44	2.30	.11
Year in school	2.51 (1.04)	2.73 (1.10)	2.00 (0.73)	2.81 (1.10)	2, 44	3.23	.049
GPA	3.17 (0.44)	3.23 (0.48)	3.13 (0.48)	3.16 (0.36)	2, 42	0.21	.81
Spatial orientation	18.40 (10.04)	21.60 (9.13)	14.94 (7.59)	18.88 (12.25)	2, 44	1.79	.18

Table 8
Sample size listed by gender and condition

Gender	Overall	60% Reliability	80% Reliability	100% Reliability
Male	20	6	7	7
Female	27	9	9	9
Total	47	15	16	16

Tests of Hypotheses

The effects of the study manipulations were tested using a series of 3-reliability x (2-level of decision aiding x 3-potential unaided SA [x 2-trial]) mixed model ANOVAs, one for each measure of SA.

Intercorrelations

Intercorrelations among the dependent measures were calculated at each level of the within-subjects manipulations. These are presented in Tables 9-14. The pattern of correlations suggested a strong positive relationship between the objective level 1 (i.e., uniform color accuracy) and level 2 (i.e., status accuracy) SA measures. This relationship appeared to diminish as potential unaided SA increased. A relationship between the SART and the objective SA measures was not observed, but both the SART and the objective measures were correlated with performance. These results suggested that each of the SA measures captured elements of SA that were useful for task performance. At the same time, the SART may measure different aspects of SA than are captured by the objective measures.

Table 9
Intercorrelations among dependent measures for poor (10%) potential unaided SA with acquisition (stage 1) aiding

Variable	*M*	*SD*	1	2	3	4
1. Uniform color accuracy SA	.45	.26	–			
2. Status accuracy SA	.55	.22	.83**	–		
3. SART	19.22	7.70	.22	.16	–	
4. Performance	.35	.28	.30*	.36*	.29*	–

* $p < .05$. ** $p < .01$. $N = 47$ for all measures.

Table 10

Intercorrelations among dependent measures for moderate (50%) potential unaided SA with acquisition (stage 1) aiding

Variable	*M*	*SD*	1	2	3	4
1. Uniform color accuracy SA	.69	.22	–			
2. Status accuracy SA	.74	.16	.69**	–		
3. SART	20.01	8.33	.06	.03	–	
4. Performance	.47	.31	.46**	.28	.20	–

* $p < .05$. ** $p < .01$. $N = 47$ for all measures.

Table 11

Intercorrelations among dependent measures for good (90%) potential unaided SA with acquisition (stage 1) aiding

Variable	*M*	*SD*	1	2	3	4
1. Uniform color accuracy SA	.72	.13	–			
2. Status accuracy SA	.73	.09	.52**	–		
3. SART	17.82	7.67	-.19	-.08	–	
4. Performance	.36	.28	.26	.48**	.19	–

* $p < .05$. ** $p < .01$. $N = 47$ for all measures.

Table 12

Intercorrelations among dependent measures for poor (10%) potential unaided SA with analysis (stage 2) aiding

Variable	*M*	*SD*	1	2	3	4
1. Uniform color accuracy SA	.41	.18	–			
2. Status accuracy SA	.66	.15	.05	–		
3. SART	20.50	7.44	.03	.19	–	
4. Performance	.35	.25	.02	.28	.25	–

* $p < .05$. ** $p < .01$. $N = 47$ for all measures.

Table 13

Intercorrelations among dependent measures for moderate (50%) potential unaided SA with analysis (stage 2) aiding

Variable	*M*	*SD*	1	2	3	4
1. Uniform color accuracy SA	.36	.16	–			
2. Status accuracy SA	.72	.17	.07	–		
3. SART	20.10	7.07	-.10	.22	–	
4. Performance	.53	.33	.32*	.31*	.30*	–

* $p < .05$. ** $p < .01$. $N = 47$ for all measures.

Table 14

Intercorrelations among dependent measures for good (90%) potential unaided SA with analysis (stage 2) aiding

Variable	*M*	*SD*	1	2	3	4
1. Uniform color accuracy SA	.71	.17	–			
2. Status accuracy SA	.72	.20	.61**	–		
3. SART	18.35	7.99	-.07	.02	–	
4. Performance	.37	.24	-.06	.07	.19	–

* $p < .05$. ** $p < .01$. $N = 47$ for all measures.

Tests of Normality

Normality was assessed for each of the dependent variables. Significant skewness was defined as having a ratio of skewness to standard error of skewness greater than 1.96. This cutoff was a test of the null hypothesis that the data was normally distributed at an alpha level of .05. Significantly skewed levels of the interaction of the within-subjects variables are listed in Table 15. A number of levels of each measure were skewed, but the only measure with a consistent pattern of skewness was the SART, which was negatively skewed. The robustness of ANOVA to violations of normality in the population is a topic of debate but is improved with larger sample sizes (Tabachnick & Fidel, 2007). Given cell sizes of 15-16 and an overall sample size of 47, and to preserve interpretability of the findings, the dependent measures were not transformed prior to analysis.

Table 15
List of significantly non-normal dependent measures by levels of the within-subjects variables

Measure	Level of Potential Unaided SA	Stage of Diagnostic Aiding	Skew / Std. Error of Skew
SART	Low	Acquisition	-3.79
SART	Moderate	Acquisition	-3.26
SART	Good	Acquisition	-3.68
SART	Good	Analysis	-2.37
Status accuracy	Moderate	Acquisition	-2.74
Status accuracy	Good	Analysis	-2.93
Uniform color accuracy	Moderate	Acquisition	-3.18
Uniform color accuracy	Moderate	Analysis	2.03
Uniform color accuracy	Good	Analysis	-4.46

The assumption of homogeneity of variance was assessed for each ANOVA. Because the sample sizes were relatively equal, the significance test of Box's *M* was ignored (Tabachnick & Fidel, 2007). Because of its sensitivity, the recommended probability level for this test was .001 (Pallant, 2007). This test was significant only for the SART.

Finally, multivariate ANOVAs were used to avoid the assumption of sphericity on the within-subjects factors.

Manipulation Check

Uniform color accuracy. There was a statistically significant main effect for potential unaided SA on the level 1 SA measure, uniform color accuracy, $F(2, 43) = 108.93$, $p < .001$, partial $\eta^2 = .84$. There were differences across all three conditions, and these differences were in the hypothesized direction (see Figure 15). Level 1 SA was highest at good (90%) potential unaided SA ($M = .72$, $SE = .02$) compared to both moderate (50%; $M = .52$, $SE = .02$, $p < .001$, $d = 1.47$), and poor (10%; $M = .43$, $SE = .02$, $p < .001$, $d = 2.42$) potential unaided SA. Moderate (50%) potential unaided SA lead to higher level 1 SA than poor ($p = .002$, $d = 0.64$). This result supported the manipulation check.

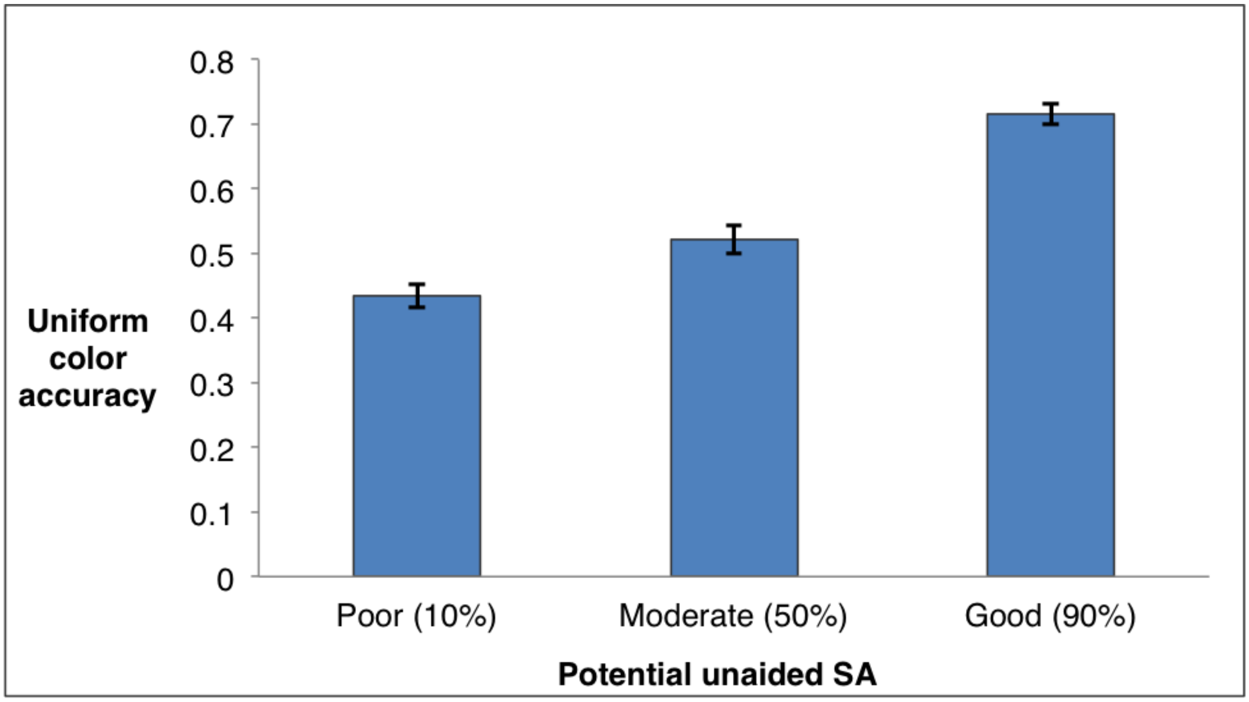


Figure 15. Uniform color accuracy as a function of levels of potential unaided SA. Error bars show standard errors.

Status accuracy. The manipulation of potential unaided SA was expected to affect SA across the SA measures. There was a significant main effect for potential unaided SA on the level 2 measure, status accuracy, $F(2, 43) = 34.63$, $p < .001$, partial $\eta^2 = .62$ (see Figure 16). Participants had a lower percentage of correct responses when potential unaided SA was poor (10%; $M = .60$, $SE = .02$) than when it was moderate (50%; $M = .73$, $SE = .02$, $p < .001$, $d = 1.15$) or good (90%; $M = .73$, $SE = .02$, $p < 0.001$, $d = 1.11$).

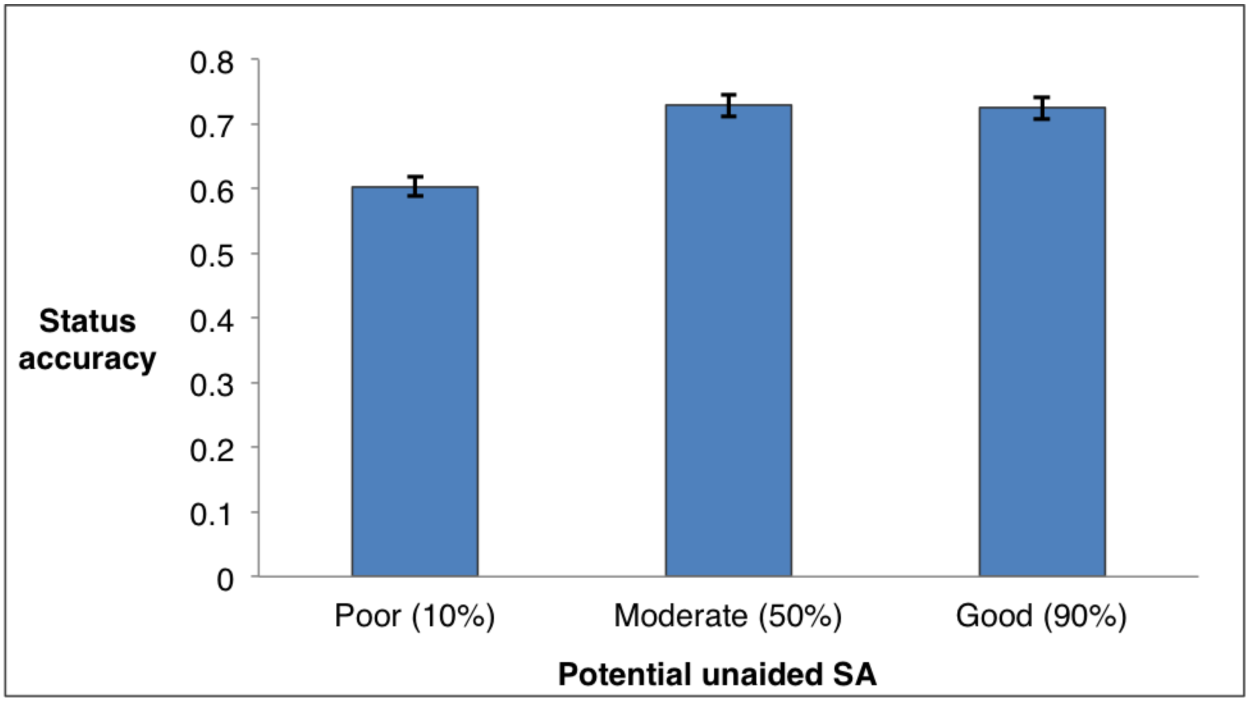


Figure 16. Status accuracy as a function of potential unaided SA. Error bars show standard errors.

SART. There was a significant main effect for potential unaided SA on the SART, $F(2, 43) = 8.97$, $p = .001$, partial $\eta^2 = .29$. Participants rated their SA as lower when they had good (90%; $M = 18.06$, $SE = 0.98$) potential unaided SA than either moderate (50%; $M = 20.03$, $SE = 0.94$, $p = .001$, $d = .30$) or poor (10%; $M = 19.85$, $SE = 0.98$, $p = .008$, $d = .30$) potential unaided SA (see Figure 17). This was unexpected, as the good (90%) potential unaided SA missions should have had the highest ratings of SA.

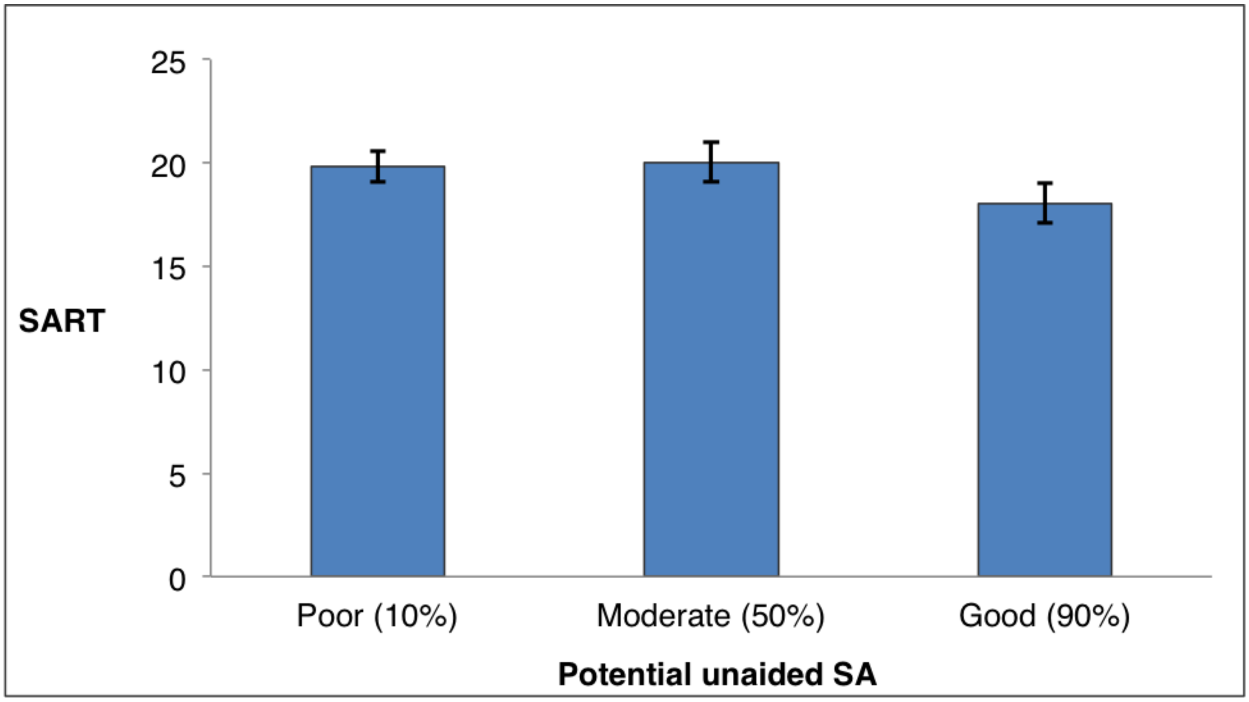


Figure 17. SART scores as function of potential unaided SA. Error bars show standard errors.

Hypothesis 1

Hypothesis 1 was tested to confirm that reliable diagnostic aiding improved SA and that the stages of diagnostic aiding, information acquisition (stage 1) and information analysis (stage 2), corresponded to the levels of SA, perception and comprehension, respectively.

Uniform color accuracy. *Hypothesis 1a: Under perfect reliability, diagnostic aiding that performs acquisition only (stage 1) will lead to better level 1 SA but not level 2 SA (a simple effect).*

For the measure of level 1 SA, an interaction between reliability and the stage of diagnostic aiding, showing greater SA for information acquisition (stage 1) diagnostic aiding than for analysis (stage 2) diagnostic aiding at perfect reliability, was expected to support this hypothesis. This would support the conclusion that information acquisition (stage 1) aiding supports level 1 SA.

The three-way interaction was examined first to determine whether the anticipated relationship depended on the level of potential unaided SA, but it was not statistically significant, $F(4, 86) = 2.20$, $p = .076$, partial $\eta^2 = .09$. However, a significant two-way interaction between reliability and the stage of diagnostic aiding was observed, $F(2, 44) = 6.34$, $p = .004$, partial $\eta^2 = .22$. Supporting this hypothesis, accuracy was higher with information acquisition (stage 1) at 100% reliability ($M = .71$, $SE = .03$) than with information analysis (stage 2; $M = .52$, $SE = .02$, $p < .001$, $d = 1.71$).

Status accuracy. *Hypothesis 1b: Under perfect reliability, diagnostic aiding that performs acquisition and analysis (stage 2) will lead to better level 1 SA and level 2 SA (simple effects).*

An interaction between reliability and the stage of diagnostic aiding, showing greater SA for information analysis (stage 2) diagnostic aiding than for acquisition (stage 1) diagnostic aiding on a measure of level 2 SA was expected to support this hypothesis. Because this relationship may have depended on the level of potential unaided SA, the three-way interaction was examined first.

The three-way interaction was significant, $F(4, 86) = 5.86$, $p < .001$, partial $\eta^2 = .21$. However, mean differences at 100% reliability were not significant. Means for moderate (50%; $p = .16$, $d = 0.47$) and good (90%; $p = .65$, $d = 0.15$) potential unaided SA were in the anticipated direction, but the means for poor (10%; $p = .53$, $d = 0.22$) potential unaided SA showed a small, albeit non-significant, advantage for information acquisition (stage 1). This interaction is further explained and illustrated below.

SART. Because the SART did not distinguish between level 1 and level 2 SA, it was not expected to be sensitive to these effects. This was indeed the case, with an interaction between reliability and stage of diagnostic aiding not reaching significance, $F(2, 44) = 2.45$, $p = .10$, partial $\eta^2 = .10$ (see Figure 18).

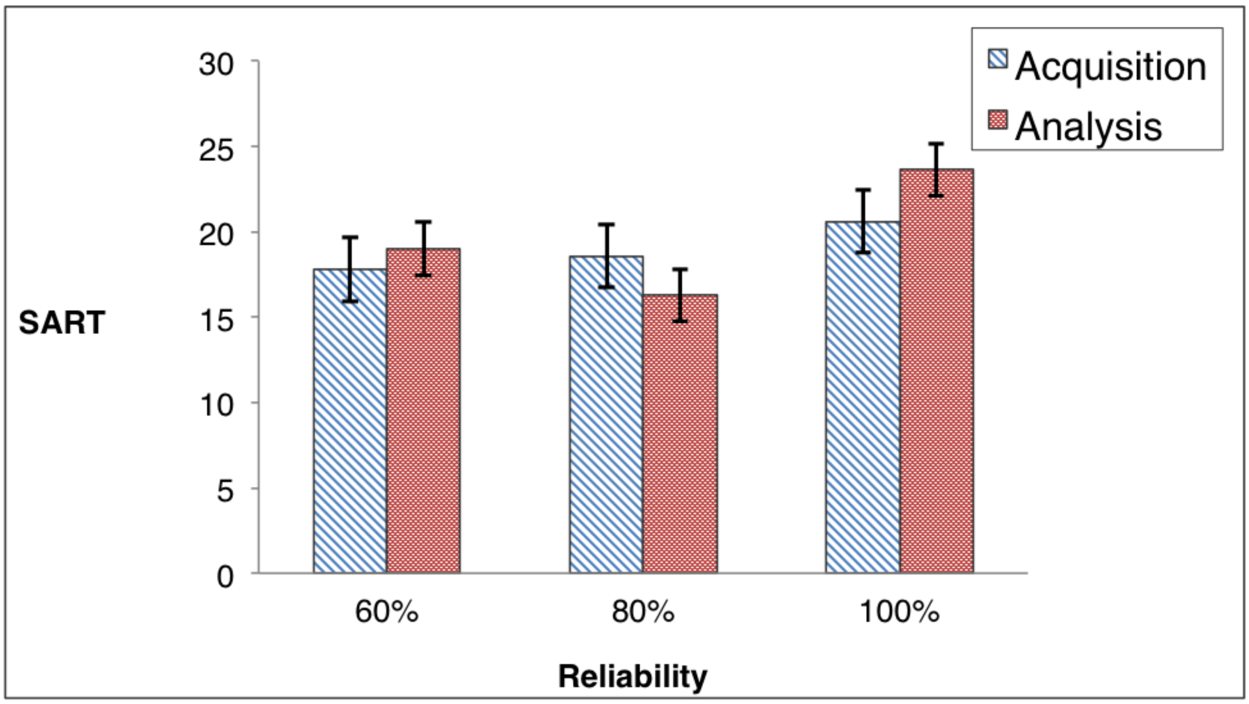


Figure 18. SART as a function of reliability condition and stage of diagnostic aiding. The interaction did not reach significance. Error bars show standard errors.

Hypothesis 2

Given support for the manipulation of potential unaided SA and the utility of perfectly reliable diagnostic aiding, I next examined the effects of unreliability. Hypothesis 2 addressed the overall effect of reliability:

Hypothesis 2: Operator SA will be higher at higher levels of robot reliability (a main effect).

While this hypothesis was supported, the relationship was more complex than what was predicted. It was expected that higher reliability would be associated with higher SA in a relatively linear fashion, especially for measures of level 2 SA, where the effect was expected to be strongest. This single linear effect was not observed (see Figure 18).

Uniform color accuracy. A two-way interaction of potential unaided SA and reliability was found $F(4, 86) = 8.74$, $p < .001$, partial $\eta^2 = .29$. Accuracy was higher in the 100% reliability condition ($M = .57$, $SE = .03$) than in either the 80% ($M = .33$, $SE = .03$, $p < .001$, $d = 2.05$) or 60% ($M = .40$, $SE = .03$, $p < .001$, $d = 1.45$) conditions at poor (10%) potential unaided SA (see Figure 19). Differences were not significant at moderate (50%) or good (90%) potential unaided SA ($p > .05$ in each case).

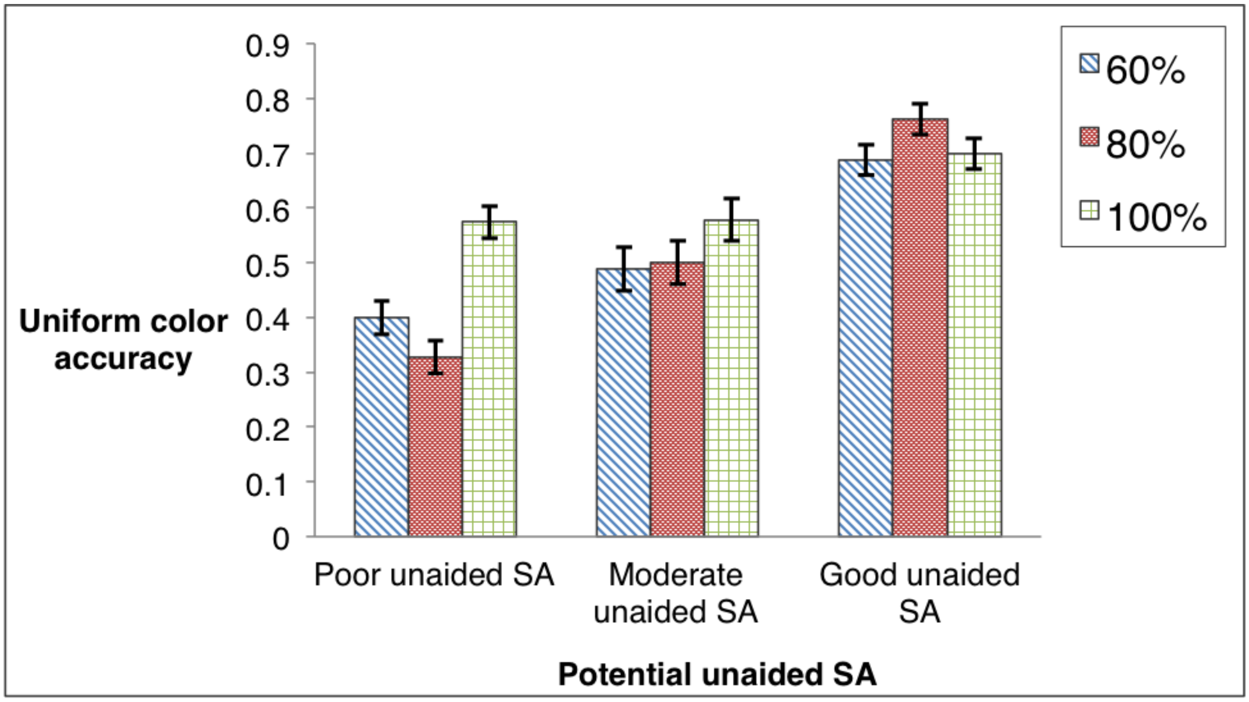


Figure 19. Uniform color accuracy as a function of reliability condition and potential unaided SA. Error bars show standard errors.

Status accuracy. For the level 2 SA measure, uniform color accuracy, examination of the three-way interaction suggested that the effect of reliability depended on the level potential unaided SA, so the two-way interaction of potential unaided SA and reliability was examined. A two-way interaction between these factors was significant, $F(4, 86) = 13.69$, $p < .001$, partial $\eta^2 = .25$ (see Figure 20). Significant differences were observed at poor (10%) potential unaided SA; accuracy was higher at 100% reliability ($M = .73$, $SE = .03$) than 80% ($M = .54$, $SE = .03$, $p < .001$, $d = 1.84$) and 60% ($M = .54$, $SE = .03$, $p < .001$, $d = 1.84$) reliability. These differences were also observed at moderate (50%) potential unaided SA; accuracy was higher at 100% reliability ($M = .82$, $SE = .03$) than 80% ($M = .68$, $SE = .03$, $p = .001$, $d = 1.21$) and 60% ($M = .68$, $SE = .03$, $p < .001$, $d = 1.25$) reliability. At good (90%) potential unaided SA, SA was higher at 80% reliability (*M* = .76, *SE* = .03) than at 60% reliability (*M* = .67, *SE* = .03, *p* = .05, *d* = 0.10).

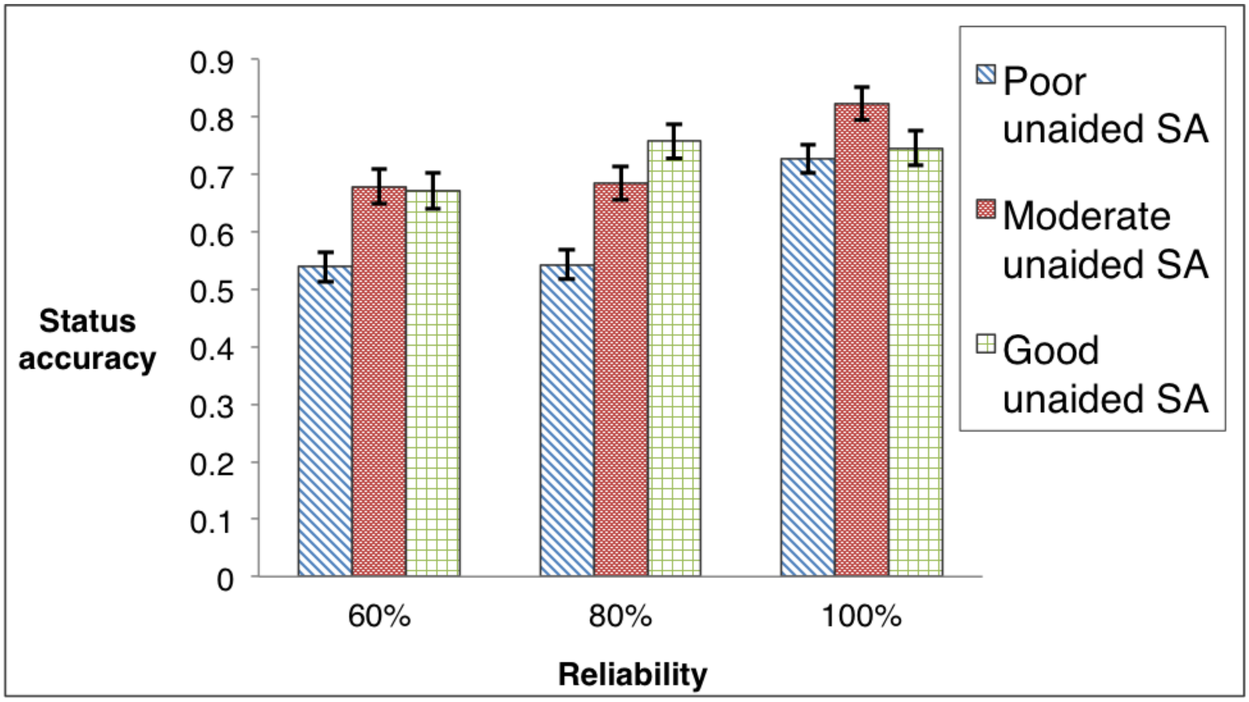


Figure 20. Status accuracy as a function of reliability and potential unaided SA. Error bars show standard errors.

SART. No significant main effects of reliability, nor any interaction effects, were found for the SART. There was a non-significant trend of higher SART scores in the 100% reliability condition, $F(2, 44) = 2.95$, p = .06, partial $\eta^2 = .19$ (see Figure 21). Because none of the SART effects reached significance, with the exception of a main effect for potential unaided SA previously discussed, the remaining hypotheses were tested using the two objective measures of SA, status accuracy and uniform color accuracy.

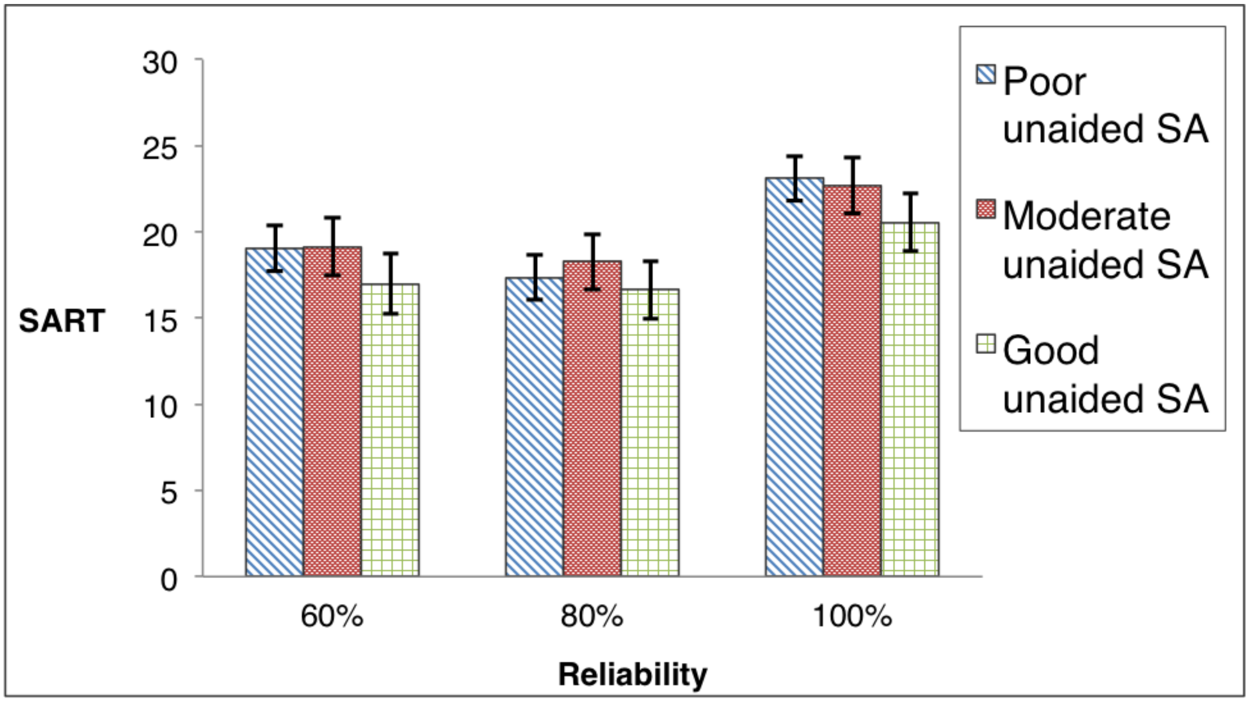


Figure 21. SART as a function of reliability condition and potential unaided SA. Error bars show standard errors.

Hypothesis 3

While tests of Hypothesis 2 showed higher SA at perfect reliability than at imperfect reliability, this relationship was dependent on potential unaided SA. To see how the two stages of diagnostic aiding were differentially affected by this relationship, Hypothesis 3 was tested.

Hypothesis 3a. *Under imperfect reliability, automation of information analysis (stage 2) will lead to lower SA when the operator would otherwise have moderate (50%) SA without the aid (a simple effect).*

It was expected that for both level 1 and level 2 SA, unreliable information analysis (stage 2) would lead to lower SA than information acquisition (stage 1). This effect was expected to be strongest for level 1 SA.

Uniform color accuracy. For the level 1 SA measure, uniform color accuracy, the three-way interaction was not significant, $F(4, 86) = 2.20$, $p = .08$, partial $\eta^2 = .09$. However, a significant interaction was observed between potential unaided SA and stage of diagnostic aiding, $F(2, 43) = 37.71$, $p < .001$, partial $\eta^2 = .64$ (see Figure 22). This suggested that the stage of diagnostic aiding affected SA differently at different levels of potential unaided SA, but this effect was similar across all levels of reliability. When potential unaided SA was moderate (50%), accuracy was higher with information acquisition (level 1; $M = .69$, $SE = .03$) than information analysis (stage 2; $M = .36$, $SE = .02$, $p < .001$, $d = 1.72$). This finding supports this hypothesis and, additionally, demonstrates the effect at perfect reliability.

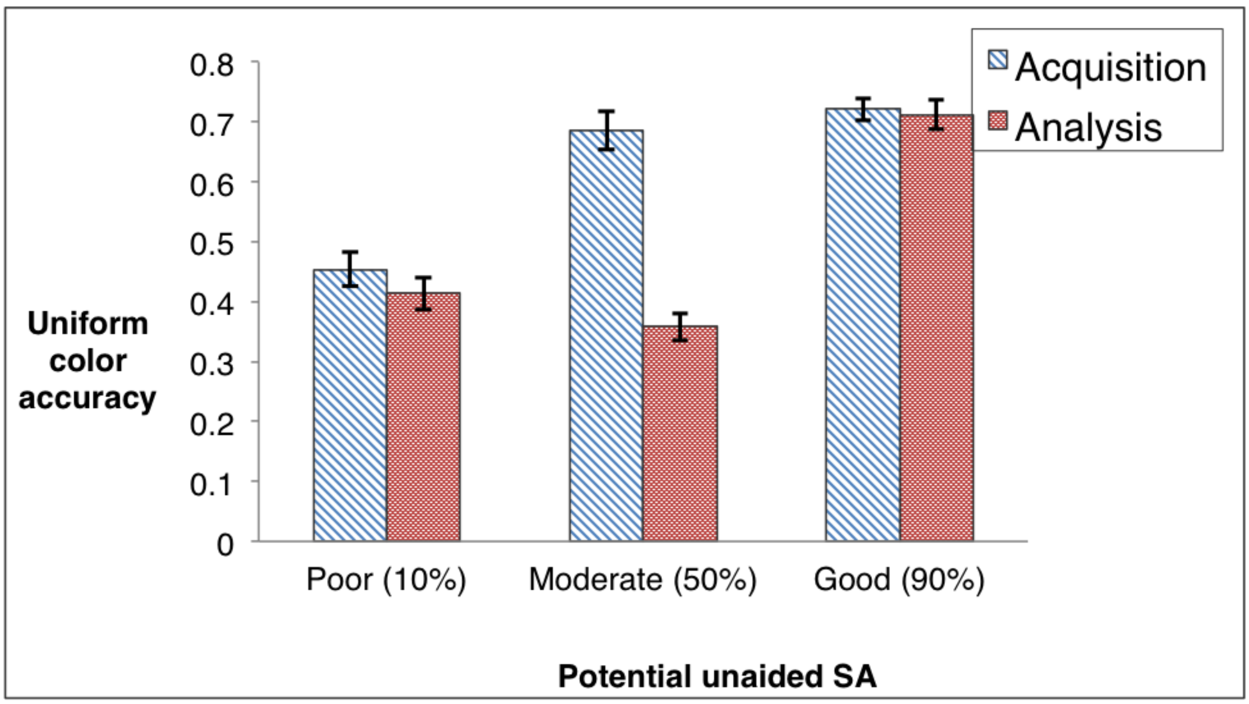


Figure 22. Uniform color accuracy as a function of stage of diagnostic aiding and potential unaided SA. Error bars show standard errors.

Status accuracy. A similar pattern was observed in the level 2 SA measure, status accuracy. Based on the three-way interaction already reported, post-hoc testing revealed a significant difference between information acquisition (stage 1; *M* = .74, *SE* = .04) and information analysis (stage 2; *M* = .62, *SE* = .04, *p* = .02, *d* = 0.83) diagnostic aiding at 60% reliability when potential unaided SA was moderate (50%; see Figure 23). This finding supports the hypothesis. Unlike the level 1 measure, which was found across levels of reliability, this effect was restricted to the lowest level of reliability, 60%. While the difference was not significant at 80% (*p* = 0.63, *d* = 0.15), the means were in the hypothesized direction.

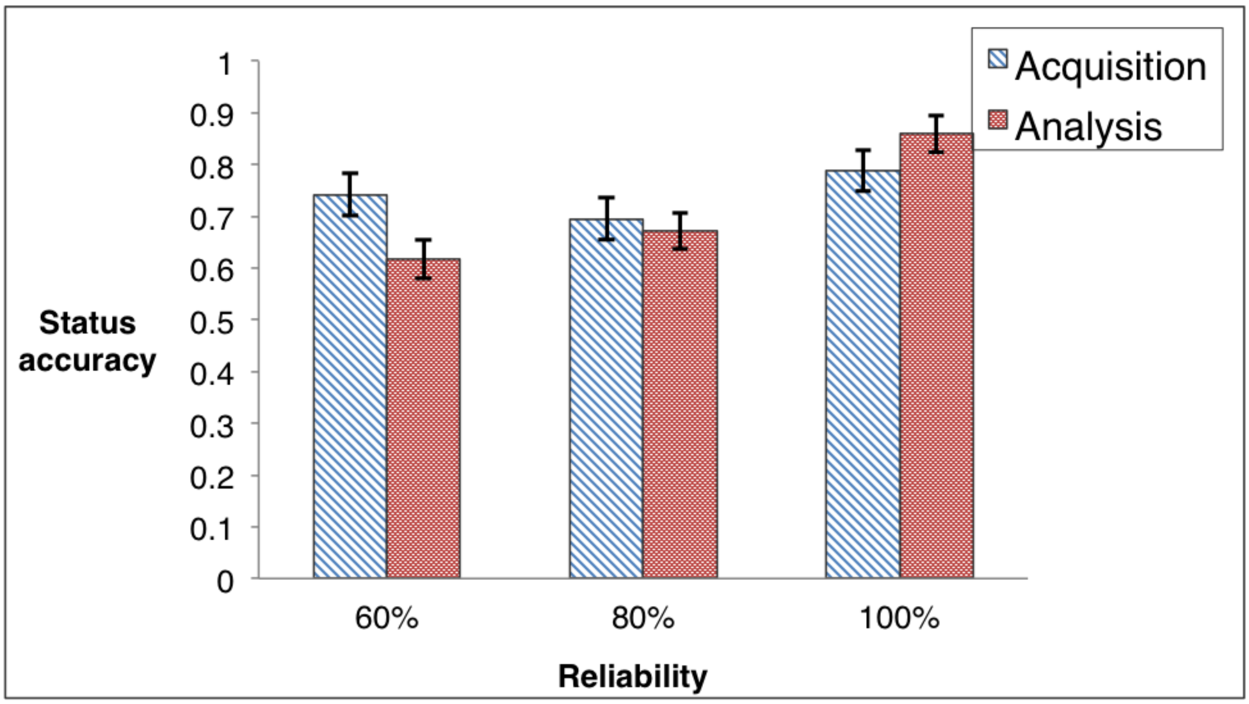


Figure 23. Status accuracy as a function of reliability and stage of diagnostic aiding *at moderate (50%) potential unaided SA*. Error bars show standard errors.

Hypothesis 3b. *Under imperfect reliability, automation of information analysis (stage 2) will lead to lower SA when the operator would otherwise have good (90%) SA without the aid (a simple effect).*

Based on the same two- and three-way interactions found for Hypothesis 3a, additional post-hoc testing was conducted at the good (90%) level of potential unaided SA. It was expected that this hypothesis would be most evident at level 2 SA. No significant differences were found across the two stages of diagnostic aiding for status accuracy (see Figure 24; $p = .27$, $d = 0.37$ at 60% reliability; $p > .99$, $d = 0$ at 80% reliability) or uniform color accuracy ($p = 0.75$, $d = 0.06$; see Figure 25). While the differences due to experimental effects cannot be distinguished from chance, the means were in the hypothesized direction at 60% reliability, with any observed differences disappearing at 80% reliability. In all, the results neither disconfirm, nor provide support for, this part of the hypothesis.

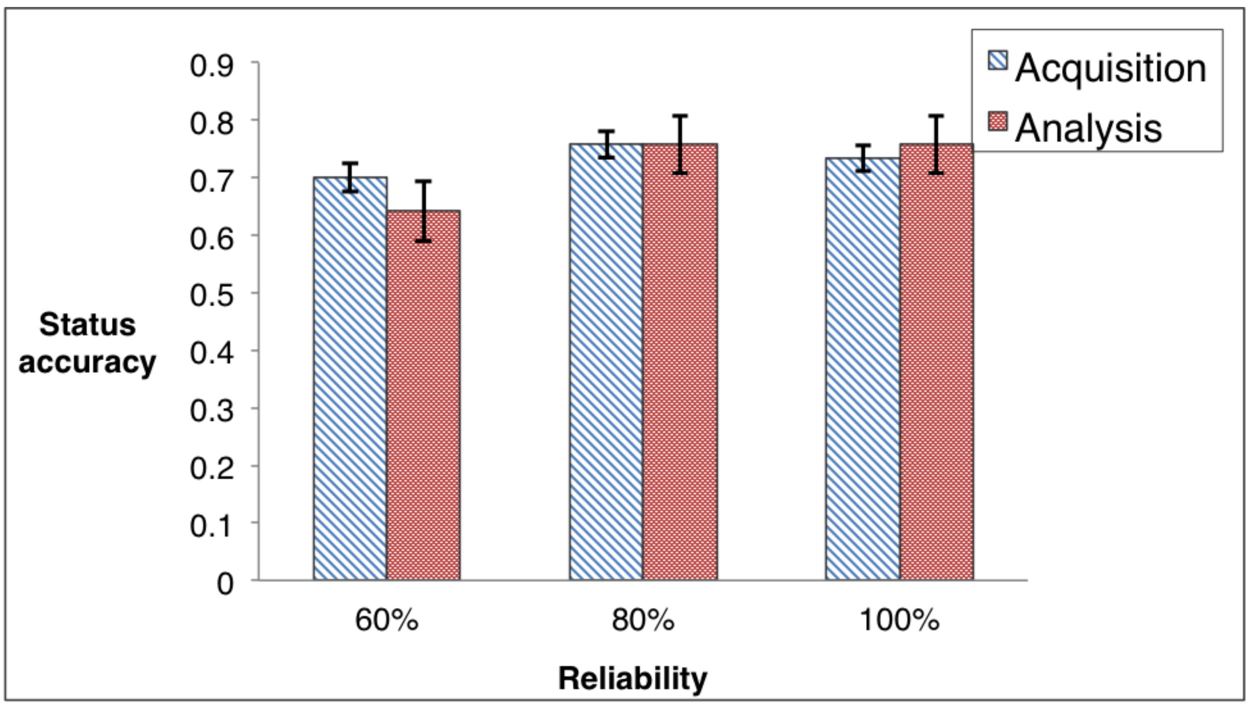


Figure 24. Status accuracy as a function of reliability and stage of diagnostic aiding *at good (90%) potential unaided SA*. Error bars show standard errors.

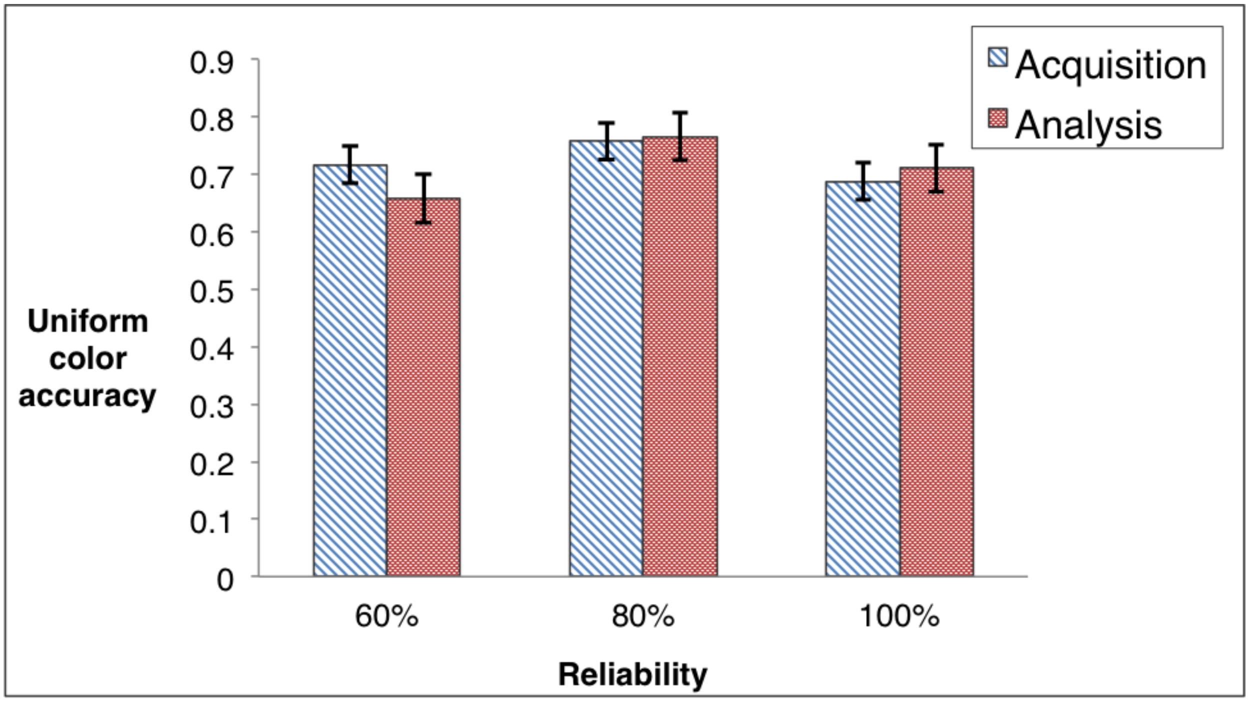


Figure 25. Uniform color accuracy as a function of reliability and stage of diagnostic aiding *at good (90%) potential unaided SA*. Error bars show standard errors.

Hypothesis 4

The final tests examine the case in which an operator could not effectively build SA without the help of the robot. This effect was anticipated to be strongest at level 2 SA.

Hypothesis 4: When the operator would otherwise have poor (10%) SA, automation of information analysis (stage 2) will lead to better SA (an interaction effect).

Uniform color accuracy. For the measure of level 1 SA, uniform color accuracy, the significant interaction between potential unaided SA and stage of diagnostic aiding was examined at the poor (10%) level of potential unaided SA. No significant differences were found (p = .35, d = 0.21; see Figure 22).

Status accuracy. To test Hypothesis 4 for the measure of level 2 SA, status accuracy, post hoc tests were conducted for the three-way interaction to examine differences across stage of diagnostic aiding when potential unaided SA was poor (10%). In support of the hypothesis, a significant difference was observed at 80% reliability (see Figure 26); information analyses led to better status accuracy ($M = .73$, $SE = .03$) than information acquisition (level 1; $M = .36$, $SE = .04$, $p < .001$, $d = 2.65$).

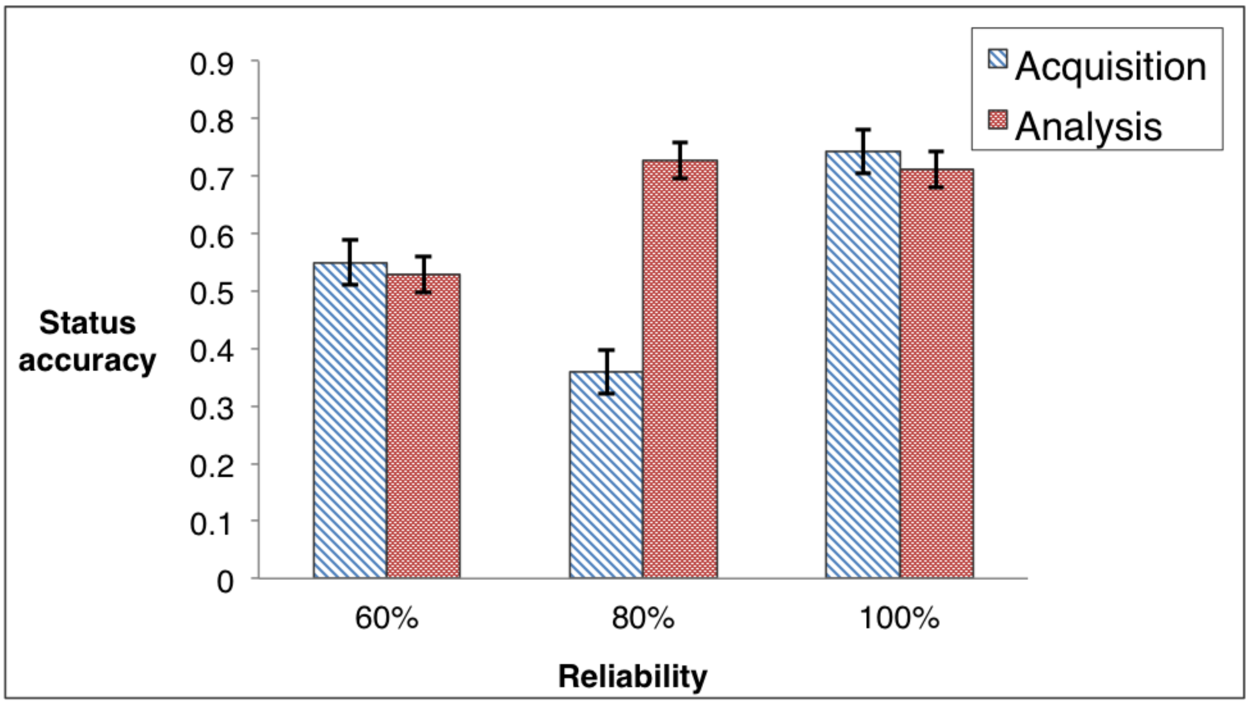


Figure 26. Status accuracy as a function of reliability and stage of diagnostic aiding *at poor (10%) potential unaided SA*. Error bars show standard errors.

Performance

Because accuracy was an additional effect of interest, the hypotheses were tested for the performance metric, count accuracy.

Manipulation Check

There was a significant main effect for potential unaided SA on performance, $F(2, 43) = 43.47$, $p < .001$, partial $\eta^2 = .67$ (see Figure 27). Participants had the highest accuracy when potential unaided SA was moderate (50%; $M = .50$, $SE = .04$) and lower accuracy when potential unaided SA was poor (10%; $M = .35$, $SE = .03$, $p < .001$, $d = 17.55$) or good (90%; $M = .37$, $SE = .03$, $p < .001$, $d = 0.50$). Although this was not in line with the manipulation, it can be explained by the presence of interaction effects discussed further below.

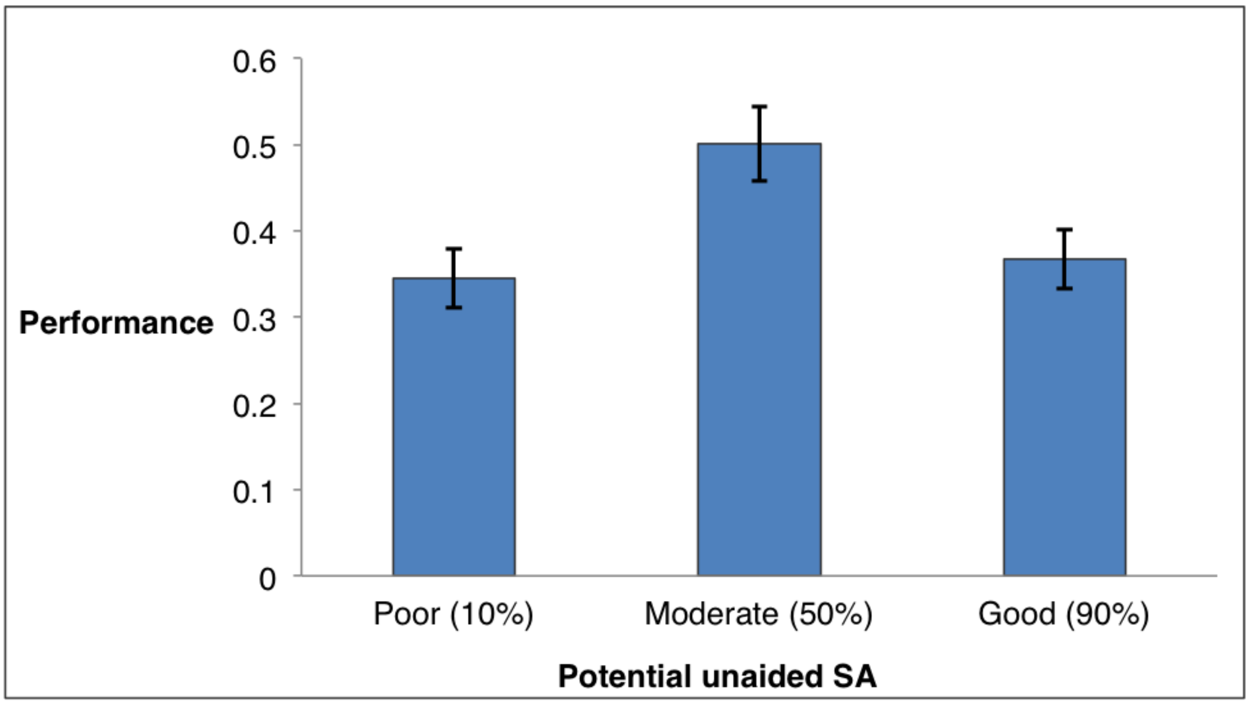


Figure 27. Performance as a function of potential unaided SA. Error bars show standard errors.

Hypothesis 1

A three-way interaction between reliability, potential unaided SA, and stage of diagnostic aiding was found for the performance measure, $F(4, 86) = 9.55$, $p < .001$, partial $\eta^2 = .31$. At 100% reliability, a significant difference was observed between the levels of automation at the moderate (50%) level of potential unaided SA (see Figure 28). Accuracy was higher with information analysis (stage 2; *M* = .65, *SE* = .82) than information acquisition (stage 1; *M* = .47, *SE* = .07, *p* < .001, *d* = 0.59).

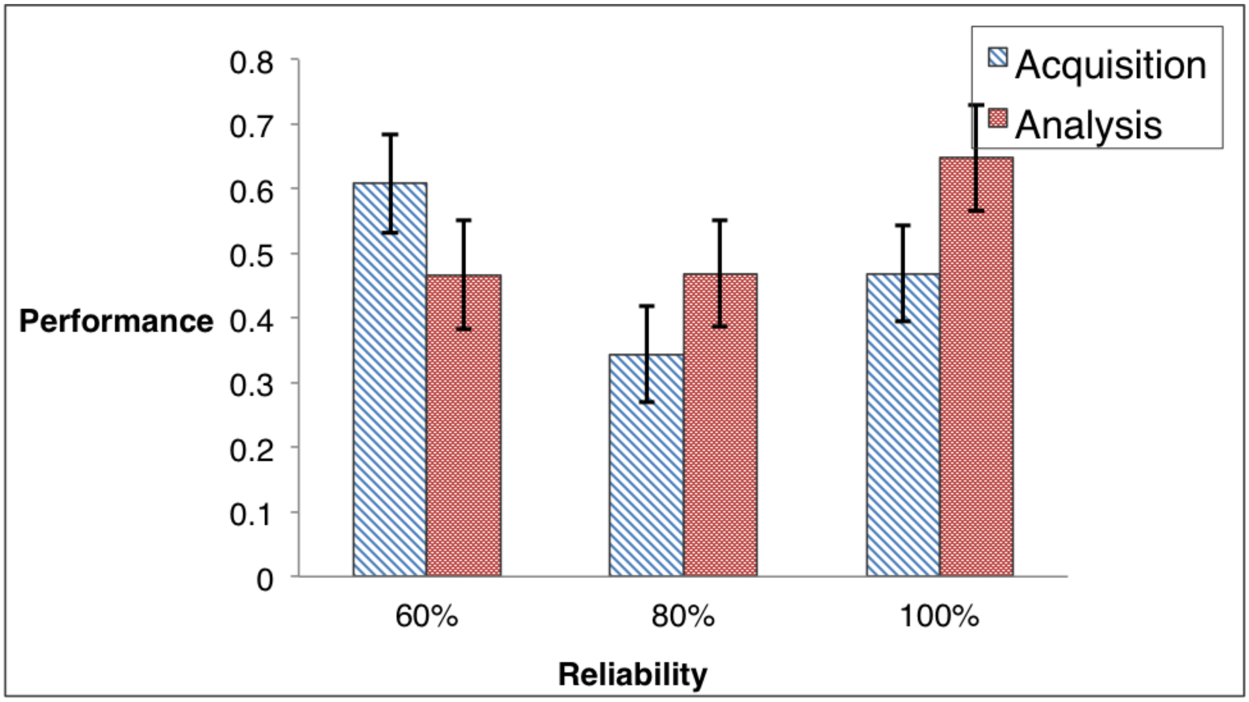


Figure 28. Performance as a function of stage of diagnostic aiding and reliability *at moderate (50%) potential unaided SA*. Error bars show standard errors.

Hypothesis 2

Examination of the three-way interaction suggested that the effect of reliability depended on the level of potential unaided SA, so the two-way interaction of potential unaided SA and reliability was examined. This interaction was significant, $F(4, 86) = 13.69$, $p < .001$, partial $\eta^2 = .39$. At poor (10%) potential unaided SA, accuracy was higher at 100% reliability (*M* = .41, *SE* = .06) versus 80% (*M* = .25, *SE* = .06, *p* = .002, *d* = 1.16) and 60% (*M* = .27, *SE* = .06, *p* = .005, *d* = 1.05) reliability. At good (90%) potential unaided SA, no significant differences were observed across reliability (*p* > .10 in each case), but the pattern of means was in the anticipated direction, with increasing accuracy as reliability increased. At moderate (50%) potential unaided SA, while there were again no significant differences ($p > .10$ in each case), the pattern of results was more like that of poor (10%) potential unaided SA (see Figure 29).

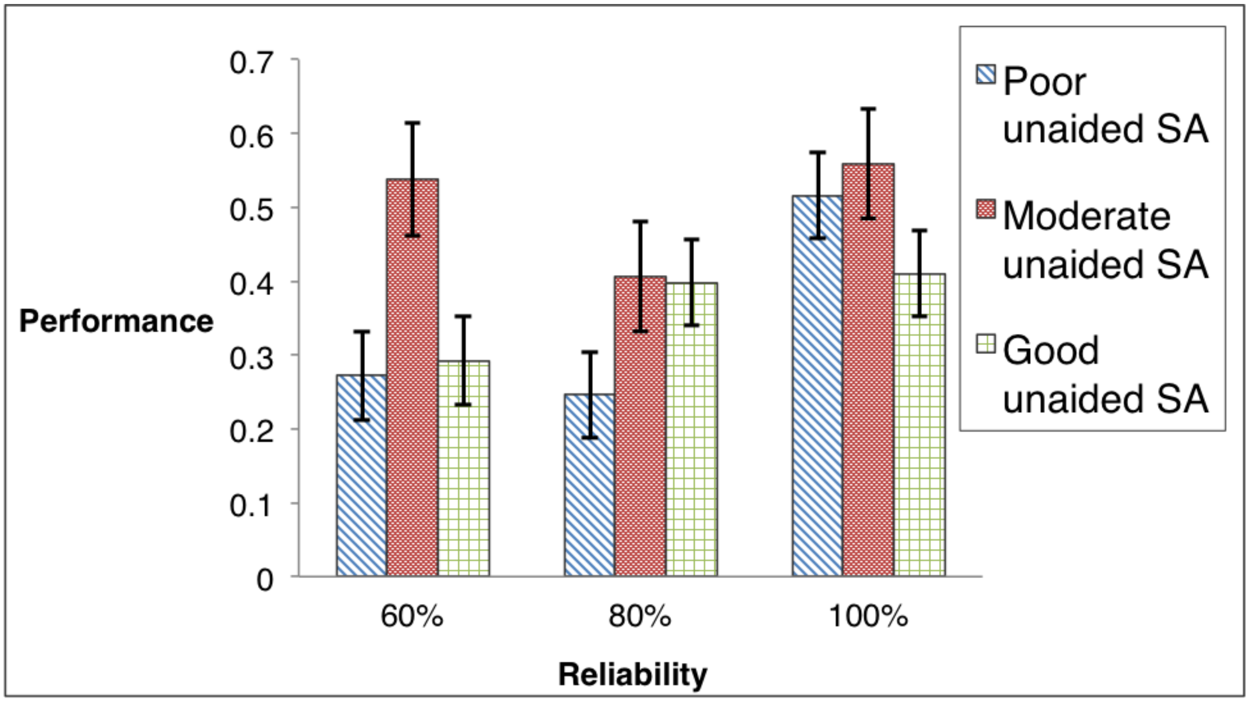


Figure 29. Performance as a function of reliability level and potential unaided SA. Error bars show standard errors.

Hypothesis 3

Post-hoc testing of the previously reported three-way interaction of potential unaided SA, reliability, and stage of diagnostic aiding on count accuracy revealed a significant difference between information acquisition (stage 1) and information analysis (stage 2) diagnostic aiding at all three levels of reliability, although not the in the same direction (see Figure 28). In support of the hypothesis, participants at 60% were more accurate with information acquisition (stage 1) aiding ($M = .61$, $SE = .08$) than they were with information analysis (stage 2) aiding ($M = .47$, $SE = .09$, $p = .005$, $d = 0.45$). At 80% reliability, however, the direction of the relationship was reversed, with participants being less accurate with information acquisition (stage 1; $M = .34$, $SE = .07$) than with information analysis (stage 2; $M = .47$, $SE = .08$, $p = .01$, $d = 0.40$). This trend continued at 100% reliability, with information acquisition (stage 1; $M = .47$, $SE = .07$) leading to worse SA than information analysis (stage 2; $M = .65$, $SE = .08$, $p < .001$, $d = 0.57$).

Additional post-hoc testing was conducted at the good (90%) level of potential unaided SA (see Figure 30). No significant differences were found across the two stages of diagnostic aiding for counting accuracy at 60% ($p = 0.28$, $d = 0.26$), 80% ($p = 0.19$, $d = 0.45$), or 100% ($p = 0.51$, $d = 0.15$) reliability.

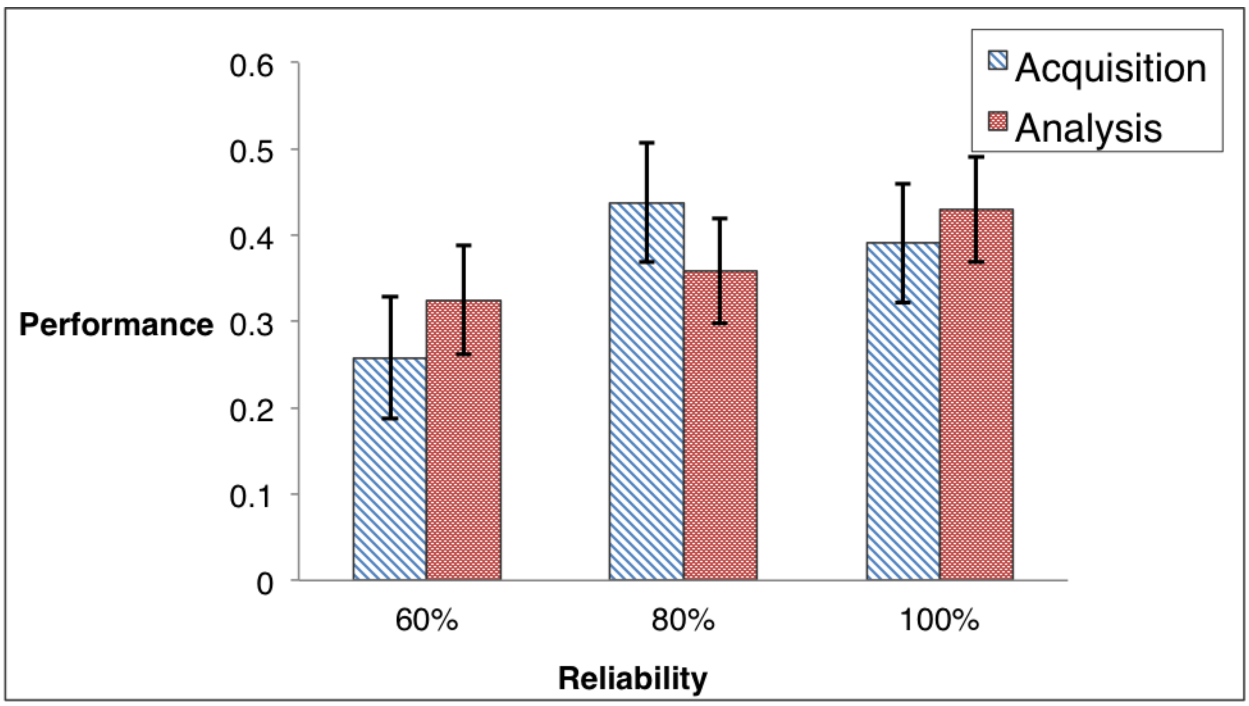


Figure 30. Performance as a function of reliability and stage of diagnostic aiding *at good (90%) potential unaided SA*. Error bars show standard errors.

Hypothesis 4

The effects anticipated under Hypothesis 4 did not significantly affect performance. At 80%, the effect was in the expected direction ($p = 0.12$, $d = 0.22$; see Figure 31).

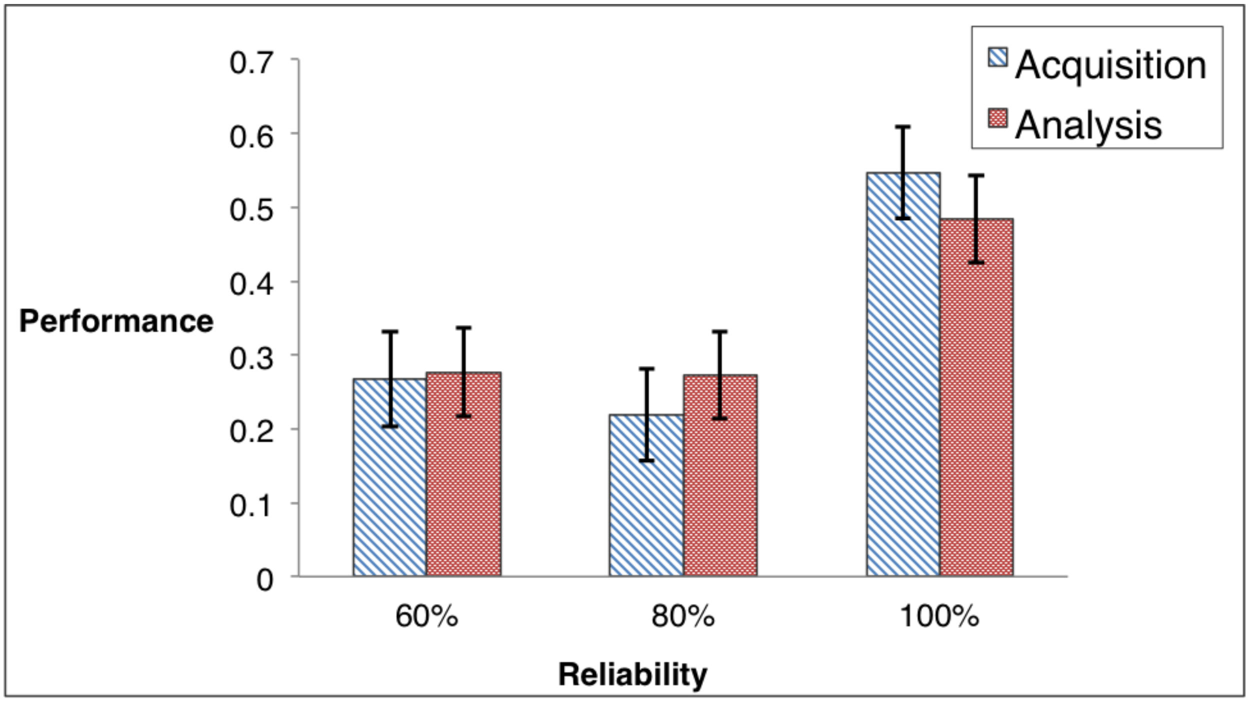


Figure 31. Performance as a function of reliability and stage of diagnostic aiding *at poor (10%) potential unaided SA*. Error bars show standard errors.

Metacognitive Measures

Only potential unaided SA significantly affected participants' ratings of their own performance, $F(2, 43) = 3.99$, $p = .03$, partial $\eta^2 = .16$ (see Figure 32). Participants rated their own performance worst after poor (10%) potential unaided SA missions ($M = 4.36$, $SE = 0.23$) compared to both good (90%; $M = 4.77$, $SE = 0.22$, $p = .02$, $d = 0.27$) and moderate (50%; $M = 4.77$, $SE = 0.22$, $p = .007$, $d = 0.27$) ones.

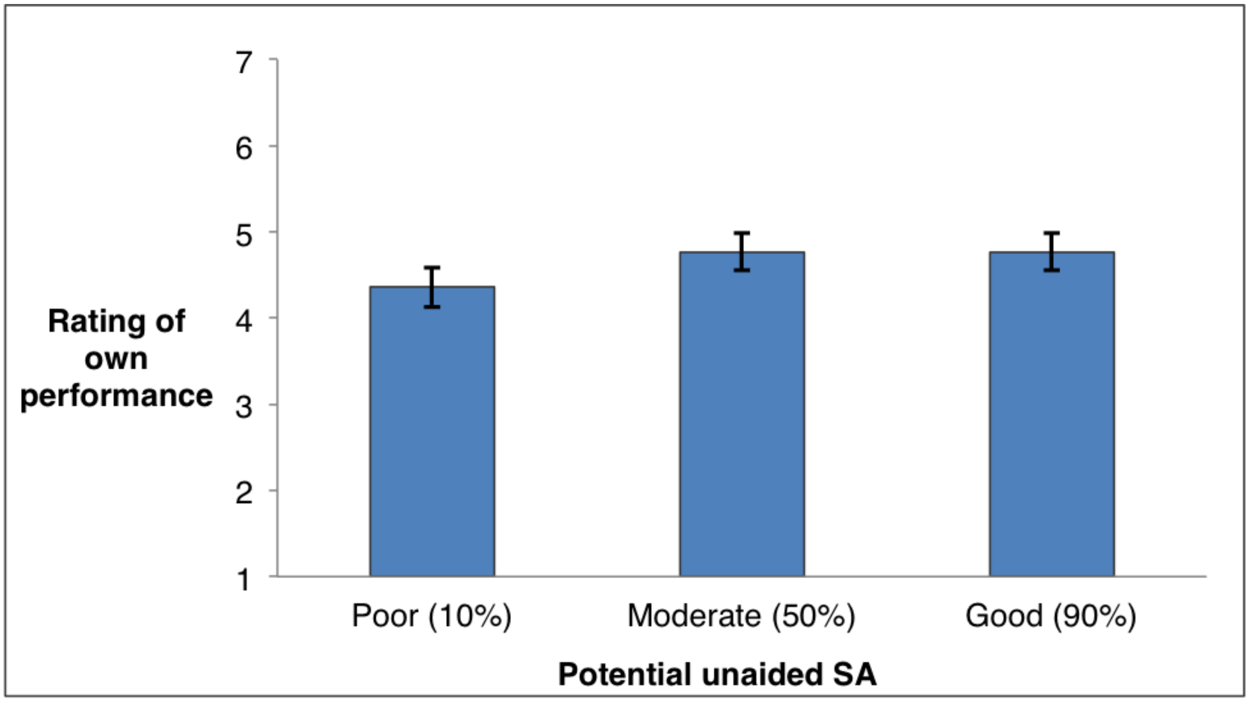


Figure 32. Participants' self-rating of performance as a function of potential unaided SA. Error bars show standard errors.

Participants' ratings of the robot's performance were significantly affected by reliability, $F(2, 44) = 25.18$, $p < .001$, partial $\eta^2 = .53$ (see Figure 33). Participants rated the perfectly reliable robot higher ($M = 6.07$, $SE = 0.28$) than both 80% reliability ($M = 3.80$, $SE = 0.28$, $p < .001$, $d = 2.02$) and 60% reliability ($M = 3.47$, $SE = 0.29$, $p < .001$, $d = 2.32$). The two imperfect reliability conditions were not significantly different from each other, $p = .42$, $d = 0.29$.

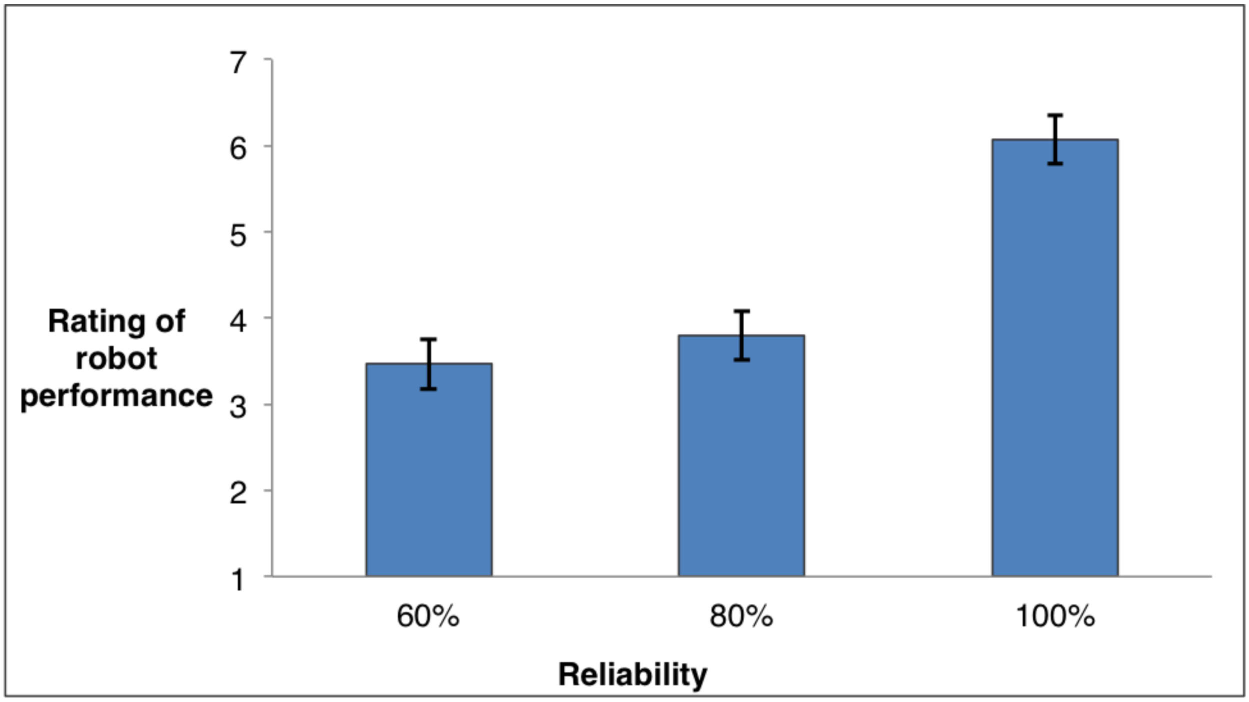


Figure 33. Participants' rating of the robot's performance as a function of reliability. Error bars show standard errors.

Participants' ratings of the human-robot team's performance were also significantly affected by reliability, $F(2, 44) = 12.15$, $p < .001$, partial $\eta^2 = .36$ (see Figure 34). Participants rated the team higher at 100% robot reliability ($M = 6.09$, $SE = 0.32$) than both 80% reliability ($M = 4.43$, $SE = 0.32$, $p = .001$, $d = 1.31$) and 60% reliability ($M = 3.97$, $SE = 0.33$, $p < .001$, $d = 1.67$). The two imperfect reliability conditions were not significantly different from each other, $p = .31$, $d = 0.37$.

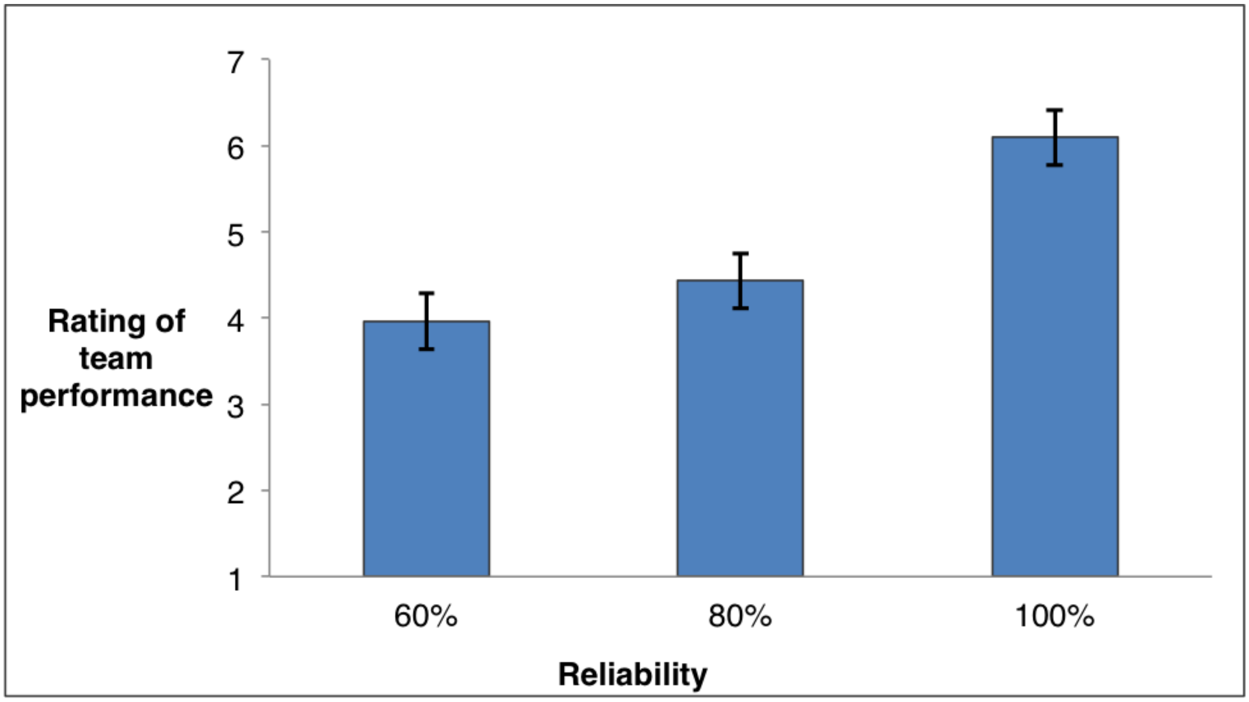


Figure 34. Participants' rating of the human-robot team's performance as a function of reliability. Error bars show standard errors.

CHAPTER FIVE: DISCUSSION

This study examined the effects of stage of diagnostic aiding, robot reliability, and unaided task performance on operator situation awareness (SA). While prior research had shown that unreliability of automated systems can critically affect their usefulness, the impact of unaided task performance had not been systematically considered. This study expanded upon prior research by: (a) studying how SA explains the relationship between automation factors and performance, and (b) studying whether the effects of automation unreliability depend on the level of automation while considering how unaided task performance may affect this relationship. Overall, the results showed that reliability is not the sole determinant of effective use of automation. Unaided operator performance affects this relationship such that unreliable automation can still provide a benefit; the SA drop-off as reliability falls is mitigated when operators cannot easily build their own SA.

Effects of Stage of Diagnostic Aiding and Reliability

Summary of Results

Hypotheses 1 and 2 served to establish effects that have been demonstrated in other automated systems. The first was a link between the stages of diagnostic aiding and the levels of SA. Specifically, support for Hypothesis 1 established that automation of information acquisition (stage 1), led to better level 1 SA. Although it was also expected that perfectly reliable information analysis (stage 2) would lead to better level 2 SA than information acquisition (stage 1), no significant differences were found to support this second effect.

Hypothesis 2 stated that operator SA would be higher at higher levels of robot reliability. This would have provided a cognitive explanation for the effects of unreliability on performance. Hypothesis 2 was supported, albeit with two caveats. First, significant differences were observed, but only between perfect and imperfect reliability. That is, no significant difference between the 60% reliability and the 80% reliability conditions were observed on either of the SA measures. Second, the relationship depended on the level of potential unaided SA; the effects were observed only under poor (10%) or moderate (50%) potential unaided SA, but not under good (90%) potential unaided SA.

Discussion

These findings provide theoretical support for the *what* of automation according to Wickens and Dixon's (2007) stages of diagnostic aiding. That is, information acquisition (stage 1) improved level 1 SA. However, the expected relationship between information analysis (stage 2) and level 2 SA was not observed. This suggests that the information analysis (stage 2) aid did not lead to better SA categorically. Understanding this finding requires further consideration of the performance data. At moderate (50%) levels of potential unaided SA (when participants were required to integrate their own information with the robot's information to the greatest extent), information analysis (stage 2) led to higher performance than information acquisition (stage 1). While the information analysis (stage 2) aid did help performance over the lower level of aiding, it did not improve SA.

As far as perfect reliability is concerned, it was hypothesized that the higher the level of the diagnostic aiding, the better for SA. Instead, it may be that even reliable diagnostic aiding can put an operator out-of-the-loop while still allowing good performance. In this scenario, operators may make heavy use of the automation, thus improving performance but not SA. Thus, operators may exhibit good performance even without good SA.

The evidence for Hypothesis 2 confirmed the extensive prior research showing the importance of automation reliability. Both level 1 SA and level 2 SA improved under perfect reliability. While performance and SA do fall as reliability drops, the degree to which this occurs differs depending on the operator's unaided performance. At poor (10%) or moderate (50%) potential unaided SA, the robot could be expected to have a greater role in the task, so its reliability was more important. It could have been the case that under good (90%) potential unaided SA, enough of the task could be performed without the robot such that it its mistakes did not affect operator SA, perhaps because the person could compare robot responses with their own evaluation of the information. This conclusion is supported by the performance data; a main effect for reliability showed that 100% reliability led to superior performance than each level of unreliability.

Potential Unaided SA

Summary of Results

Hypotheses 3 and 4 predicted SA under unreliability at different levels of potential unaided SA. Specifically, under unreliability, automation of information analysis (stage 2) should have led to lower SA when the operator would otherwise have moderate (50%) or good (90%) SA without the aid. Evidence for this effect was found at moderate (50%) potential unaided SA. Information acquisition (stage 1) was associated with better level 1 SA at all levels of reliability and better level 2 SA at 60% reliability. The effect was not observed under good (90%) potential unaided SA.

Hypothesis 4 considered the case when the operator was unable to effectively perform the task alone. When the participant could gather little information without the diagnostic aid, and reliability was 80%, information analysis (stage 2) was better. This significantly affected level 2 SA, but not level 1 SA.

Discussion

The purpose of testing hypotheses 3 and 4 was to determine whether imperfect automation would provide a benefit as unaided performance varied. Hypothesis 3 was partially supported. More automation (i.e., information analysis [stage 2]) led to lower SA under limited reliability (i.e., 60%) when the operator had moderate (i.e., 50%) access to information in the environment. As with the previous hypotheses, and in line with the performance data, these effects were evident only at the moderate (50%) level of potential unaided SA. This further demonstrates that unaided operator performance affects the relationships established by prior research. That is, this finding confirms that it is more difficult for operators to build SA when unreliable information is integrated than when unreliable information is passed directly to the operator.

An additional contribution is that the negative effect of unreliable integrated information depends upon unaided operator performance. This effect was not present when potential unaided SA was good (90%). Perhaps participants at the good (90%) level of potential unaided SA simply had enough information directly available to them that unreliability in the robot was not detrimental; they were not forced to use the robot and could choose to ignore it. At the moderate (50%) and poor (10%) levels, the robot mediated more information.

Under poor (10%) potential unaided SA, direct access to relevant information was very limited (e.g., 20% of individuals), necessitating use of the robot. Here, it was expected that even unreliable information analysis (stage 2) would provide a benefit over information acquisition (stage 1). This was the case, although with limitations. When the participant could not perform the task alone, information analysis (stage 2) was superior to information acquisition (stage 1) at 80% reliability. This illustrates that even when reliability is limited, more automation is not always bad for SA. This finding expands upon prior literature that examined the *out-of-the-loop* performance problem as an issue of *too much* automation. Prior work has aimed to find an appropriate level of automation in which operators are neither out of the loop nor overburdened. Rather than trying to find an ideal level of automation, the current research examined *what* tasks could be automated while maintaining operator SA. In addition to considering what tasks are being performed by the automation, unaided performance should be considered as well. In this study, the expected linear drop in SA was not present for poor (10%) potential unaided task performance.

In all, these findings provide evidence that the availability of information in the environment is a factor that should be considered in system design to maximize SA. When the operator can adequately perform the task alone, increasing the stage of unreliable automation will cause a detriment. However, when the operator's task performance is otherwise very poor, even unreliable automation can lead to higher SA. At the same time, this does not mean that reliability does not need to be considered; below a certain level of reliability, the benefit of increased automation disappears.

Subjective Measurement of SA

The differences observed between objective and subjective measures of SA showed again that these measurement techniques capture different facets of the SA construct. This was evident in the pattern of intercorrelations among the measures and performance. SART captured elements important to task performance that were unique from the level 1 and level 2 SA captured by the objective measures. Consequently, the SART may be relevant, but it measures something different than both task performance and behavioral SA.

The SART offers the least clear distinction between level 1 and level 2 SA. Theoretically, the SART is sensitive to different levels of SA to the extent that participants knew the distinction between the two levels and weighted their responses based on these levels. In other words, participants would have needed to know that knowledge of individual characteristics alone was insufficient and lowered their ratings accordingly for the SART to measure level 2 SA. In the present study, it was unlikely that participants made this distinction considering their unfamiliarity with the task and that they received no feedback on their task performance. Further, to be sensitive to SA, participants needed to know what they did not know. Even accepting that this may be possible under the best circumstances, the novice level of participants on this task probably made calibrating expectations for SA difficult.

Ultimately, the SART may have utility in situations where operators have a high quality mental model of the environment in which they work and thus have a basis for making selfreports of their own knowledge. The SART is a subjective measure, but it may also be a holistic measure that is affected by a complex mix of task familiarity, attitudes, and other factors.

Conclusions

Implications

The finding that the utility of unreliable automation depends on the level of automation provides additional support for the findings of Rovira, McGarry, and Parasuraman (2007) and Sarter and Schroeder (2001). Under unreliability, automation of information analysis (stage 2) led to lower SA than information analysis (stage 1) when the operator had moderate (50%) potential unaided SA. This study added to these past findings by providing cognitive explanations for previously found effects on performance. Additionally, this research considered how unaided human performance might have affected these relationships. When the participant could gather little information without the diagnostic aid, and reliability was 80%, information analysis (stage 2) lead to better level 2 SA. The findings that: (a) that the impact of automation's unreliability on SA and performance is affected by the level of unaided human performance, and (b) that SA is a mediator of this relationship, have a number of implications for both theory and practice.

Theoretical implications. This research contributes to our understanding of automation reliability by showing that having direct access to information changes the nature of the relationship between automation reliability and SA. When some information can be gathered directly, fallible information analysis (stage 2) automation hinders SA more than information acquisition (stage 1) automation, likely because operators have limited means to uncover the source of the error. When it is difficult to gather information directly, however, adding information analysis (stage 2) does not hinder SA, and it may provide a benefit, even at the same level of automation reliability. This may seem to contradict prior research suggesting a stable point below which automation hinders performance, but a likely explanation is that the effect depends on additional factors that were not previously considered. In much of the prior research, operators were able to take over when automation failed. In this study, unaided performance was manipulated. In situations where operators cannot perform the task on their own, later stages of even unreliable automation can be beneficial. Reliability remains an important factor in effective use of automation, but its impact on SA and performance depend on what the operator can accomplish without the automation.

The findings confirm that diagnostic aiding is a useful taxonomy for describing the cognitive tasks handled by automation and that reliable diagnostic aiding is useful for reducing cognitive load, improving SA, and supporting task performance. It is useful to consider *what* is being automated and not just *how much* automation is being used. This research connects Endsley's (1994) levels of SA with the stages of diagnostic aiding (Wickens & Dixon, 2007); it supports Horrey et al.'s (2009) theory that stages of diagnostic aiding have corresponding effects on levels of SA. The inclusion of information acquisition (stage 1) provides operators with a measurable increase in level 1 SA. At the same time, automation of information analysis (stage 2) leads to better performance. This is because level 1 SA is not sufficient on its own to perform the identification task; it is only when this information is integrated that it becomes level 2 SA.

This research also has implications for Endsley's (1994) model of SA. Under this model, SA at level 1 is the foundation for SA at level 2. The results of this study support this taxonomy;

the two measures were strongly correlated and consistently predicted performance, with the level 2 measure displaying slightly stronger correlations with performance. The level 1 and the level 2 SA measures were generally sensitive to the same manipulations, with the level 1 SA measure finding a more widespread effect of level of diagnostic aiding (i.e., across all levels of reliability) than level 2 SA (which found a significant effect only at 60% reliability). Thus, level 1 SA supports level 2 SA, which was a significant predictor of task performance. Because of the applied nature of the SA construct, care must be taken to avoid self-referential definitions of knowledge (e.g., using good performance to identify what needed to be known which, in turn, improved performance). In this study, I attempted to minimize this problem by creating a task where management of quantifiable elements was essential to task performance. This methodology resulted in a situation where information corresponding to level 1 SA was withheld. Under the model, these operators would have good level 2 SA and poor level 1 SA, which should not be possible if level 1 is the foundation of level 2 SA. It could also have been the case that if the information at level 1 was not needed to perform the task, it stopped being part of SA, leaving only level 2 information. The results, which do not completely settle this issue, provide more support for the former conclusion. While information acquisition (stage 1) clearly supported level 1 SA, the conditions under which information analysis (stage 2) lead to higher SA were shown on both level 1 and level 2 SA measures.

The results provide a starting point for reconciling studies finding a positive relationship between reliability and performance with those finding a negative relationship. Past studies have found a performance decrement as reliability increased but remained imperfect. In this study, in contrast, the 60% reliable aid never led to better SA or performance than the 80% reliable one.

Because participants were trained on the true reliability of the robot, this outcome suggests that a performance decrement as reliability increases (but remains under 100%) would be due to inappropriate trust, rather than operator strategy selection. Even in the lowest reliability condition with information analysis (stage 2) diagnostic aiding, 60% SA and performance could have been achieved by ignoring everything except the robot's output. Participants tended not to do this; they did not follow the robot indiscriminately. If the robot had been ignored, participants would have been limited to 10% SA and performance. Participants tended to score above this level. Even at the lowest level of reliability, the robot still had an advisory role. Thus, this study suggests that *a priori* knowledge of automation reliability may be a factor in the relationship between reliability and performance.

Practical implications. These results offer a number of recommendations for the design of future automated systems working interdependently with people. For system design, reliability improves performance, an effect that is mediated by SA. Perfect reliability is the best-case design scenario but rarely, if ever, possible. This research offers potential solutions for the problem of increasing technological capability with limited reliability. One such solution could be implemented in the robot; a future robot could anticipate the performance of its human operator and adjust its capability accordingly. For example, if the robot is uncertain of its results but knows the human operator can easily see all relevant elements in the situation, it should provide low-level information so that the operator remains in the loop. Conversely, if the robot is uncertain of its results but also anticipates poor unaided performance (perhaps smoke makes it difficult for the operator to see), it should provide integrated information. In this way, future robot systems could support an operator at the time and to the degree needed while minimizing the negative effects of unreliability. The results of SA and metacognitive analyses showed that 60% and 80% reliability were both significantly worse than 100% reliability but not significantly different from each other. However, when considering the operator's ability to perform the task unaided, 80% reliability had similar effects on SA as perfect reliability. The effectiveness of information analysis (stage 2) diagnostic aiding also depended on unaided performance, and it provided a benefit over information acquisition (stage 1) at 80% reliability. Ultimately, the information processing needs of the task, the operator's ease of acquiring information, and the difficulty of information integration may determine what levels of reliability are useful. By considering these factors, system designers can implement more effective robots.

Based on the results of this study, the capability of the system should be balanced against the ease at which the operator can override or ignore the automated system. If operators have reasonable options for turning off the automation, then providing filtered, low-level data is preferable if much of that data are likely to be wrong. If turning off the automation means missing important system or environmental information, however, then even unreliable automation may still be useful. In that case, system designers should maximize the capability of the system to allow human operators to obtain the most benefit.

Although this study was primarily about system design, these results offer a number of recommendations for selection and training. This study demonstrated that an operator's unaided performance affects the impact of reliability. While unaided performance was manipulated in this study across missions with different characteristics, selection and training may also improve unaided performance. Consequently, reliability may become more important as operators build expertise due to training or selection. In this case, it may be helpful to provide experts with the ability to request low-level data to support their SA when the data may be unreliable. Additionally, novice operators may see the most benefit from unreliable automation. When possible, training should include information about the true reliability of the system, feedback about operator performance, and strategies for mitigating the effects of unreliability.

Limitations of the Study

Due to the diversity of automated systems and the challenge of measuring applied cognition, several limitations of this study should be discussed. The first is due to the nature of the SA construct and challenges associated with its measurement. While the results show how SA may be improved, the magnitude of SA is a useful metric only for comparison across very similar tasks. Because SA is a measure of relevant knowledge, each task has different knowledge requirements that are not easily compared. Quantifying this knowledge remains a challenge in real-world situations where all relevant knowledge may not be known. The purpose of this study was to see how manipulations of diagnostic aiding, robot reliability, and unaided task performance interacted to affect SA. Using a controlled experiment allowed conclusions to be made about how SA may be affected in relative terms (i.e., identification of interventions that will improve versus hinder SA). However, specific predictions about the magnitude of SA improvement in a real world task are not possible. Because of the importance of *relevant* information underlying the SA construct, and the fact that relevance is wholly task-dependent, two dissimilar tasks cannot be compared to evaluate which one results in better SA.

In line with this is variation observed across measurement techniques. Although both the SART and the objective measure quantified SA, they measured different facets of the construct. Theoretically, a self-report measure of SA can only assess what an operator believes is known relative to what the operator believes needs to be known (Jones, 2000). A major limitation with this technique is that operators may suffer from “unknown unknowns” in that they are not aware of what knowledge is needed. In this study, participants may have had difficulty calibrating their expectations for good SA. While the objective measures did not suffer from this issue, their effectiveness depended on the knowledge measured being necessary for task performance. I sampled this knowledge by measuring objective SA multiple ways. My finding that the objective measures did not provide exactly the same pattern of results was due primarily to whether level 1 or level 2 SA was assessed. Another source of differences may have been in the strategies used by participants. For example, some participants may have prioritized awareness of uniform color after seeing it in the questionnaire, even though this was not their primary task.

Both the task and the participant pool leave some questions about expert performance unanswered. The participants in this study were novices at both the task and the domain of military operations. Because of this, the task and its training were constructed to be within a difficulty range that avoided ceiling and floor effects. While the results confirm that the task did fall within this difficulty range, it is not known how these effects may vary for experts in more complex environments. Real world systems require operators to deal with more complexity than was presented in this study. In this study, participants tracked the movement of individuals with five characteristics among three locations (outdoors, the visible room, and the hidden room). In a real world task, operators may need to track many more interacting elements.

The study conclusions are also limited to diagnostic aiding. The later stages of automation, decision selection (stage 3) and action implementation (stage 4) were not measured, but, more importantly, they were not part of the task. This was a tradeoff made to distinguish the information processing components of the task from the psychomotor components of the task. In order to ensure that the SA measure captured the knowledge needed for the task, the task itself was an information management mission. In a task spanning all possible levels of automation, operators would be responsible for applying the knowledge gained in the early stages to make a decision, such as a decision to remain in place or move to a new waypoint. Because of this, level 3 SA, projection of future states, was neither important in the task nor a focus of the study.

Areas for Future Study

A future study could add to these findings by expanding the reliability manipulation in two ways. First, it could add additional levels of reliability. While 100% reliability was different from both 80% and 60% reliability, the two levels of imperfect reliability did not significantly differ in the resulting SA. By adding additional levels of reliability between 60% and 100%, the effects of imperfect reliability can be more clearly distinguished.

Second, a future study could make more refined distinctions between reliability and perceptions of reliability. Reliability is different from operators' perceptions of system reliability, and research should look at cases where the true reliability of the system is known to the operators. Anecdotally, a few participants in the 60% reliability condition stated that the robot made a large number of mistakes or that it was wrong more often than it was right, even though that was not the case. A future study could add perceptions of system reliability or provide more exposure to the true reliability level with guided feedback, to see if alerting participants to automation errors leads to more accurate perceptions of system reliability.

Future research can also extend these findings by expanding the task to encompass the other stages of automation. In a task spanning all possible levels of automation, operators would be responsible for applying their task knowledge. An unknown question is the degree to which feedback provided to the operator about the success or failure of their decision making may influence the relationship between unaided performance and SA. The SART might be more useful when operators are provided with feedback on their decision making as they gain better understanding of the information they need to know.

Finally, a future challenge will be the application of these findings to settings in which the necessary information is difficult to model *a priori*. Future battlefields will lead to complexity in information acquisition, analysis, and decision making. The present research would be augmented by research examining how the relationship between SA and the operator's potential unaided SA, reliability, and level of automation is affected when the environment is highly dynamic, requiring adaption not only in information processing, but also in strategy selection. This could be accomplished by testing these effects in a real-world, complex environment. In doing this, future research should measure *what* is automated and not just *how much* automation is used; diagnostic aiding provides a useful framework for comparing the effects of diverse forms of automation that perform similar information processing tasks.

APPENDIX A: UCF APPROVAL OF HUMAN-SUBJECTS RESEARCH

University of Central Florida

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Approval of Human Research

From: **UCF Institutional Review Board #1**
FWA00000351, IRB00001138

To: **David A. Schuster**

Date: **January 23, 2013**

Dear Researcher:

On 1/23/2013, the IRB approved the following modifications / human participant research until 1/22/2014 inclusive:

Type of Review:	UCF Initial Review Submission Form Expedited Review Category #7
Project Title:	The effects of diagnostic aiding on situation awareness under unreliability
Investigator:	David A Schuster
IRB Number:	SBE-13-09053
Funding Agency:	General Dynamics, US Army Research Laboratory
Grant Title:	RCTA - FY 2012 H2 Situation Awareness
Research ID:	1054212

The scientific merit of the research was considered during the IRB review. The Continuing Review Application must be submitted 30days prior to the expiration date for studies that were previously expedited, and 60 days prior to the expiration date for research that was previously reviewed at a convened meeting. Do not make changes to the study (i.e., protocol, methodology, consent form, personnel, site, etc.) before obtaining IRB approval. A Modification Form **cannot** be used to extend the approval period of a study. All forms may be completed and submitted online at https://iris.research.ucf.edu .

If continuing review approval is not granted before the expiration date of 1/22/2014, approval of this research expires on that date. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

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. Participants or their representatives must receive a signed and dated copy of the consent form(s).

In the conduct of this research, you are responsible to follow the requirements of the Investigator Manual.

On behalf of Sophia Dziegielewski, Ph.D., L.C.S.W., UCF IRB Chair, this letter is signed by:

Signature applied by Joanne Muratori on 01/23/2013 12:42:33 PM EST

Joanne Muratori

IRB Coordinator

APPENDIX B:
U.S. ARMY HUMAN RESEARCH PROTECTIONS OFFICE HEADQUARTERS-LEVEL ADMINSITRATIVE REVIEW

DEPARTMENT OF THE ARMY
US ARMY RESEARCH, DEVELOPMENT AND ENGINEERING COMMAND
ARMY RESEARCH LABORATORY
BUILDING 459
ABERDEEN PROVING GROUND MD 21005-5425

REPLY TO
ATTENTION OF

March 29, 2013

Human Research Protections

David A. Schuster
Department of Psychology
University of Central Florida
Orlando, FL
E-mail: dschuster@ist.ucf.edu

Dear Mr. Schuster:

The purpose of this letter is to notify you that your protocol, titled *The Effects of Diagnostic Aiding on Situation Awareness Under Unreliability* (UCF IRB Number SBE-13-09053), has been approved for implementation by the U.S. Army Human Research Protections Office (AHRPO).

I would like to give a brief account as to why your protocol needed a review by AHRPO. The review conducted by AHRPO was a component-level administrative review, governed by regulations in Section 252.235–7004 of the Defense Federal Acquisition Regulation Supplement (DFARS), which became effective in July 2009, and not a secondary review by an Army Institutional Review Board. DFARS requires that the contractor furnish his or her assurance of compliance, protocol documents, and IRB approval to AHRPO, that AHRPO review the documents to ensure contractor compliance with Department of Defense component policies, and that AHRPO approve the assurance as appropriate for the research under the Statement of Work. In lieu of a copy of the contractor's assurance, AHRPO may accept evidence of an existing assurance, if an appropriate assurance has been approved in connection with previous research. I submitted evidence of an existing UCF assurance to AHRPO.

Please note that DFARS requires you to notify the contracting officer, Dr. Susan Hill, of any suspensions or terminations of UCF's assurance.

Good luck with your study.

Sincerely,

Paul Rose

Paul N. Rose, Ph.D.
Human Protection Administrator
Army Research Laboratory

APPENDIX C:
CONFOUNDING VARIABLE TABLE

Table 16
Confounding variables

Confounding Construct	Study Constructs Affected	Directionality & Type	Effect size	Method of control	Measure	References
Robot reliability	SA	Interactive effects with level of automation, potential unaided SA	Large	Manipulated (IV) and measured		
Situation awareness (SA)				Measured (DV)		
Level of automation	SA	Interactive effects with potential unaided SA, robot reliability	Large	Manipulated (IV)		
Potential unaided situation awareness	SA	Interactive effects with level of automation, robot reliability	Large	Manipulated (IV) and measured		
Situation assessment	SA	Positive main effects	Full mediation	None		
Sensation, Perception, Attention	SA Level 1	Positive main effects	Medium			Wickens & Horrey, 2001
Cognition, Integration, Working Memory	SA Level 2	Positive main effects	Medium			Wickens & Horrey, 2001
Reaction time	Potential unaided SA, SA	Negative main effects	Low	Exclude: measure and covary	Reaction time	
Visual acuity	SA Level 1	Positive main effects	Large	Exclude outliers from sample	Self-report	
Spatial ability	SA Level 2	Positive main effects	Large	Exclude: measure and covary	Guilford-Zimmerman spatial orientation	Barnes, Jentsch, Chen, Haas, & Cosenzo, 2008
Decision making	SA	Interactive effects with level of automation	Small	Ignore: minimize decision making in task		

Confounding Construct	Study Constructs Affected	Directionality & Type	Effect size	Method of control	Measure	References
Reading comprehension	Potential unaided SA, SA	Positive main effects	Very low	Ignore		
Motivation	SA	Positive main effects	Small	Ignore		Riley et al., 2010
Attitudes towards robots	Potential unaided SA, SA	Positive main effects	Low	Ignore		Riley et al., 2010
Attitudes towards computers	Potential unaided SA, SA	Positive main effects	Low	Ignore		Riley et al., 2010
Attitudes towards military	Potential unaided SA, SA	Positive main effects	Low	Ignore but include in study recruitment materials		
Military experience	Potential unaided SA, SA	Positive main effects	Low	Exclude: measure and covary		
Robot speed	SA	Nonlinear relationship	Low	Control: fix at level		Riley et al., 2010
Robot mobility (vs. stationary)	SA	Negative main effects (mobility increases complexity)	Low	Control: fix at level		Riley et al., 2010
Risk to operator	Potential unaided SA, SA	Nonlinear relationship	High	Control: fix at level		Riley et al., 2010
Operator mobility	Potential unaided SA, SA	Negative main effects (mobility increases complexity)	Medium	Control: fix at level		Riley et al., 2010
Robot dynamics	SA	Nonlinear relationship	Low	Ignore		Riley et al., 2010
Distractors	Potential unaided SA, SA	Negative main effects	High	Control: fix at level		Riley et al., 2010
Task familiarity	Potential unaided SA, SA	Positive main effects	High	Control: fix at level		Riley et al., 2010
Terrain	Robot mobility, operator mobility, Potential unaided SA, SA	Negative main effects (rougher terrain increases complexity)	Low	Control: fix at level		Riley et al., 2010
Display size	Potential unaided SA, SA	Interactive effects with visual acuity	Low	Control: fix at level		Riley et al., 2010

Confounding Construct	Study Constructs Affected	Directionality & Type	Effect size	Method of control	Measure	References
Display resolution	Potential unaided SA, SA	Interactive effects with visual acuity	Low	Control: fix at level		Riley et al., 2010
HRI modality	SA	Interactive effects with level of automation, motor skills	Medium	Control: fix at level		Riley et al., 2010
Robot control devices	SA	Interactive effects with individual differences, task factors	Low	Control: fix at level		Riley et al., 2010
Latency of robot communication	SA	Negative main effects	High	Control: fix at level		Riley et al., 2010
Base rate of each possible decision	Potential unaided SA, SA	Nonlinear relationship	High	Control: fix at level		
Number of potential decisions	Potential unaided SA, SA	Negative main effects	High	Control: fix at level		
Task complexity	Potential unaided SA, SA	Negative main effects	High	Control: fix at level		
Robot authority to act	Level of automation, SA	Unknown	High	Control: fix at level		
Robot capability to perform work	Level of automation, reliability, SA	Positive main effects		Control: fix at level		
Trust	Perceived complexity (workload)	Interactive effects with reliability	Medium	Exclude: measure and covary		
Mental model quality	SA	Interactive effects with training effectiveness	Medium	Control: fix at level		
Communication with robot - frequency	SA	Nonlinear relationship	Low	Control: fix at level		
Communication with robot - accuracy	Robot reliability, SA	Positive main effects	Low	Control: fix at level		
Automation adaptiveness	SA	Positive main effects	Medium	Control: fix at level		Riley et al., 2010
General intelligence	Potential unaided SA, SA	Positive main effects	Low	Ignore		

Confounding Construct	Study Constructs Affected	Directionality & Type	Effect size	Method of control	Measure	References
Training effectiveness	Potential unaided SA, SA	Positive relationship	Medium	Control: fix at level		
Closure speed	Potential unaided SA, SA	Positive main effects	Medium	Exclude: measure and covary		
Motor skills	Potential unaided SA, SA	Positive main effects	Low	Ignore		
Reliance	SA	Nonlinear relationship	Medium	Ignore but measure		Parasuraman, Mouloua, & Molloy, 1996
Compliance	SA	Nonlinear relationship	Medium	Ignore but measure		
Confidence	Potential unaided SA, SA	Nonlinear relationship	Medium	Ignore but measure		
Fatigue	Potential unaided SA, SA	Negative main effects	High	Control: fix at level		
Skill degradation	Potential unaided SA, SA	Negative main effects	Medium	Control: fix at level		
Strategy effectiveness	Potential unaided SA, SA	Positive main effects	High	Control: fix at level		
Stress	Potential unaided SA, SA	Nonlinear relationship	Medium	Control: fix at level		
Perceived robot reliability	SA	Nonlinear relationship	High	Measure		
Environmental complexity	Potential unaided SA, SA	Negative relationship	High	Control: fix at level		

APPENDIX D:
GRAPHS OF HYPOTHESIZED INTERACTIONS

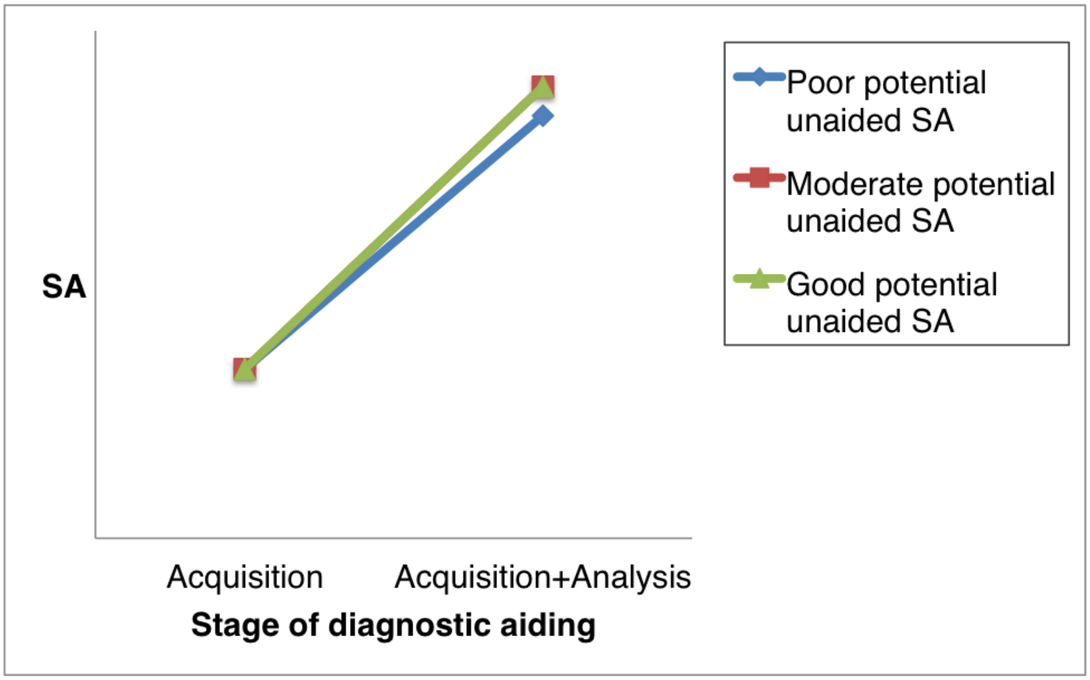


Figure 35. Graph of the hypothesized interaction between stage of diagnostic aiding and potential unaided SA at 100% robot reliability.

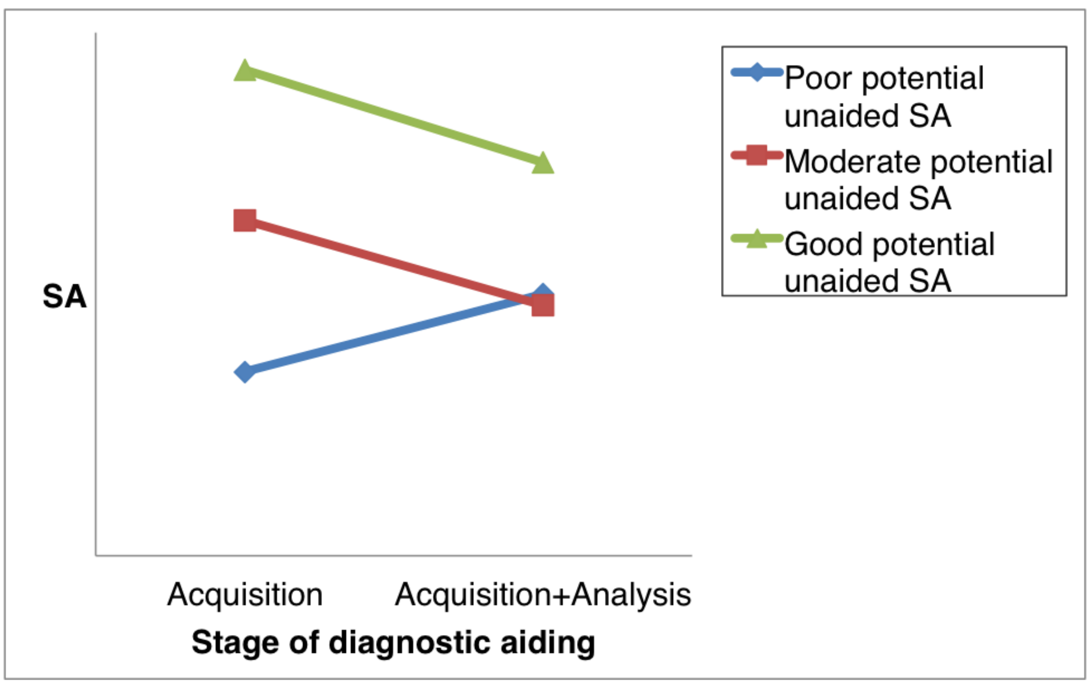


Figure 36. Graph of the hypothesized interaction between stage of diagnostic aiding and potential unaided SA at 80% robot reliability.

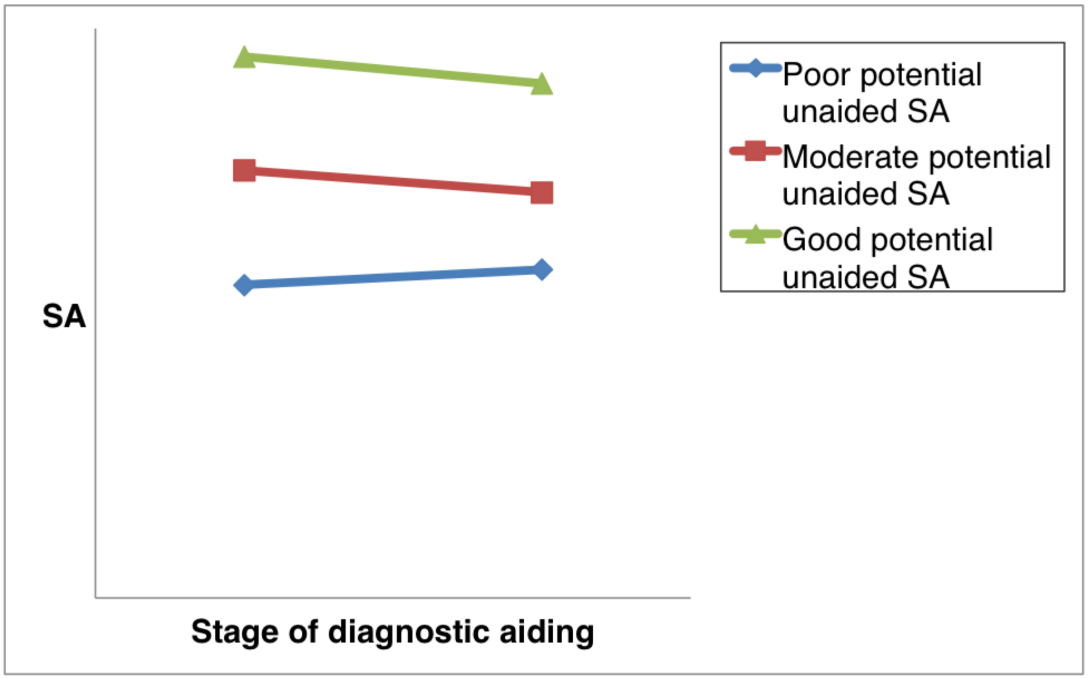


Figure 37. Graph of the hypothesized interaction between stage of diagnostic aiding and potential unaided SA at 60% robot reliability.

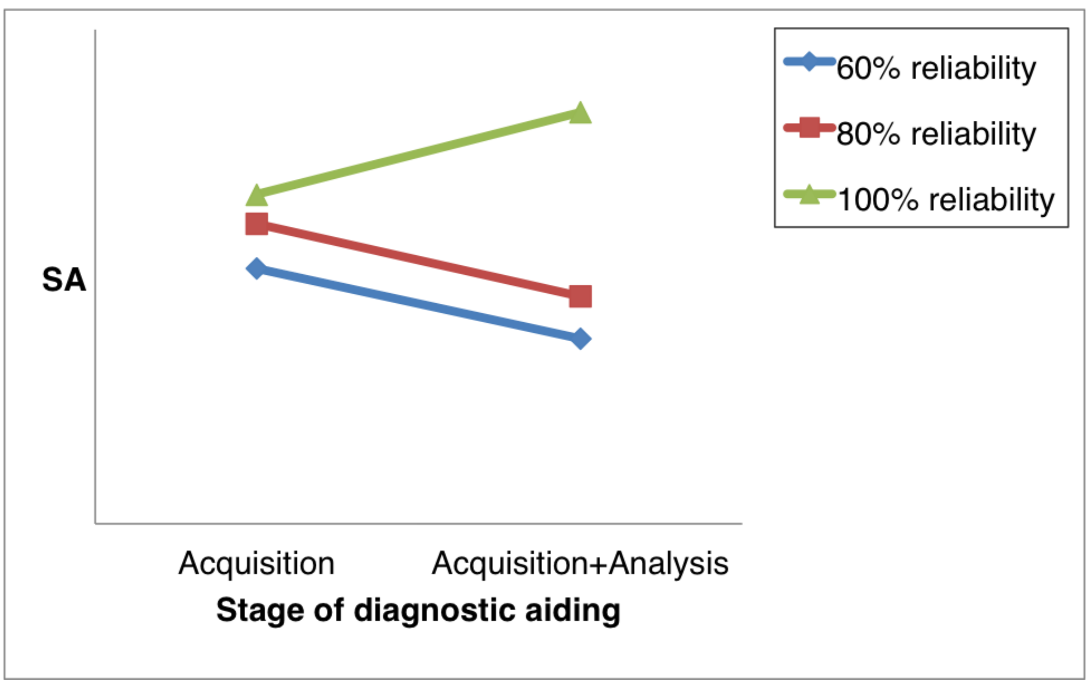


Figure 38. Graph of the hypothesized interaction between stage of diagnostic aiding and reliability at good (90%) potential unaided SA.

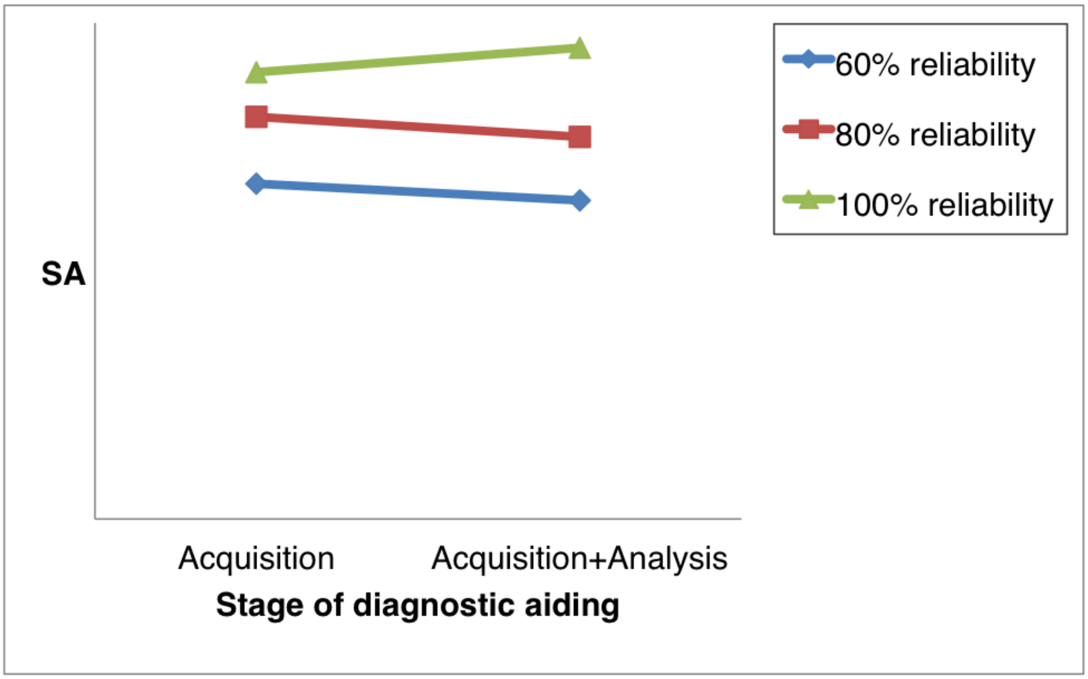


Figure 39. Graph of the hypothesized interaction between stage of diagnostic aiding and reliability at moderate (50%) potential unaided SA.

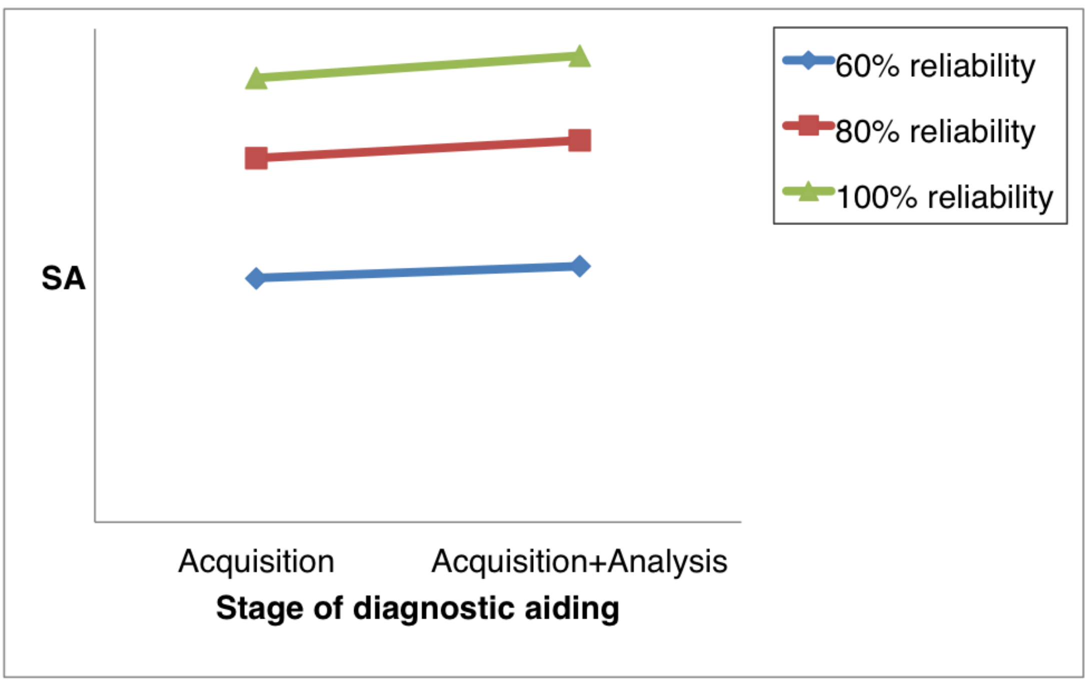


Figure 40. Graph of the hypothesized interaction between stage of diagnostic aiding and reliability at poor (10%) potential unaided SA.

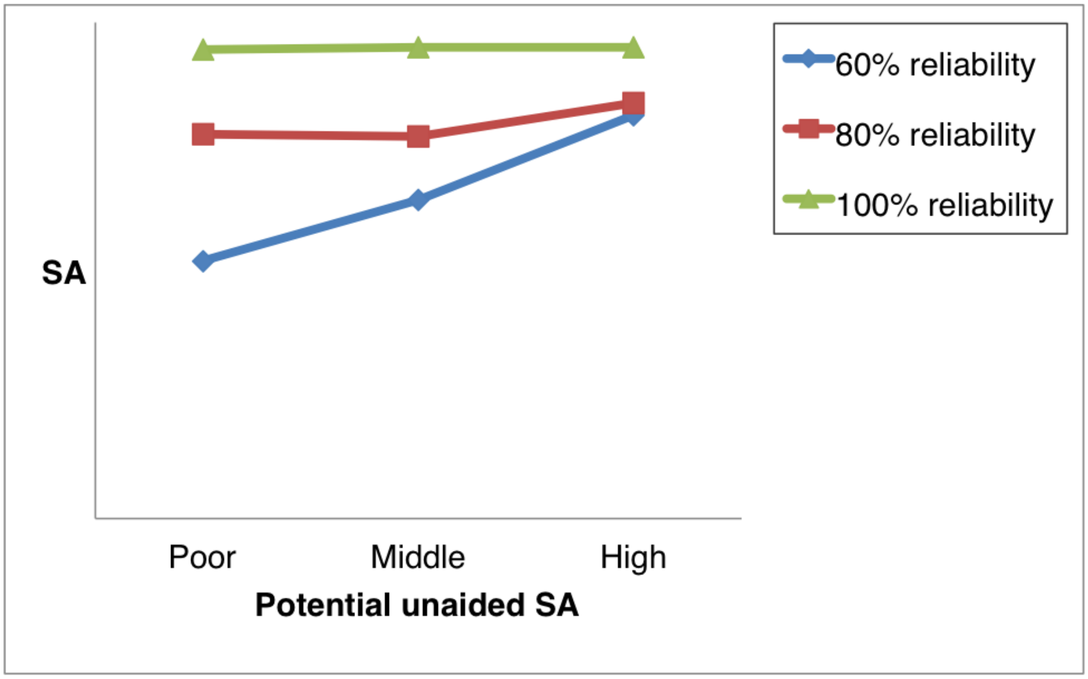


Figure 41. Graph of the hypothesized interaction between potential unaided SA and reliability at acquisition with analysis (stage 2) diagnostic aiding.

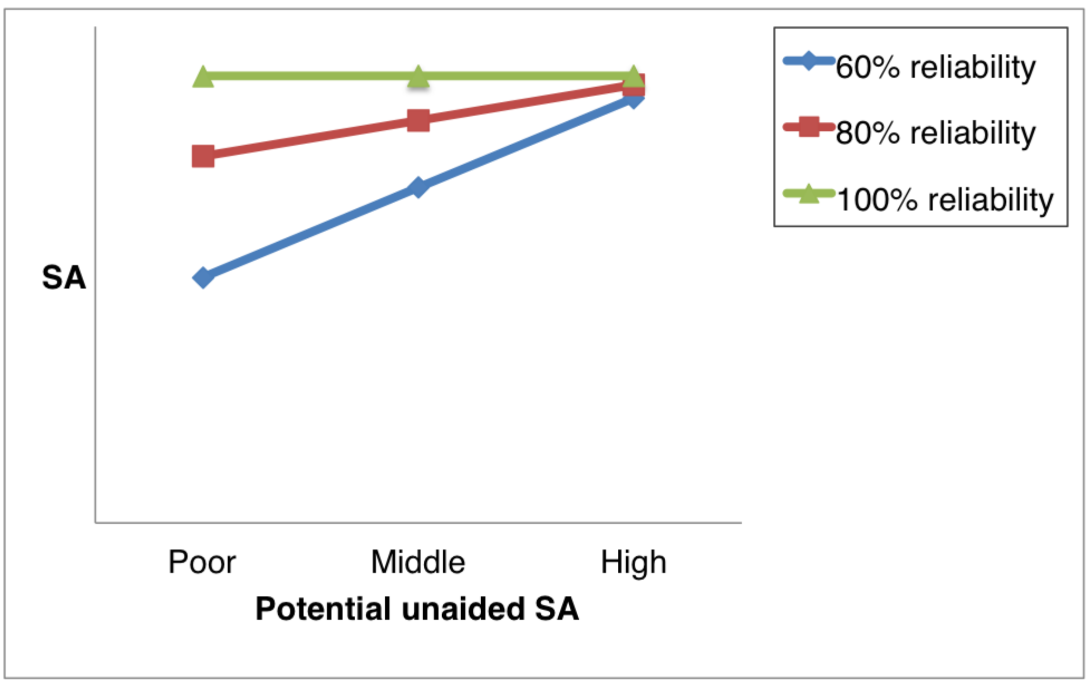
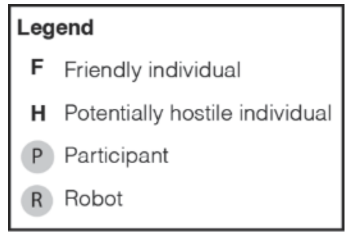


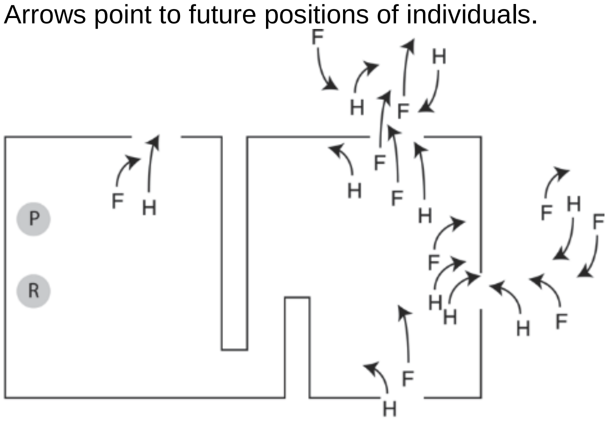
Figure 42. Graph of the hypothesized interaction between potential unaided SA and reliability at acquisition (stage 1) only diagnostic aiding.

APPENDIX E:
EXAMPLE OF ROBOT ERRORS

Scenario: Information acquisition (stage 1), 80% reliable robot, Poor (10%) potential unaided SA
All images show the same point in time during the mission.



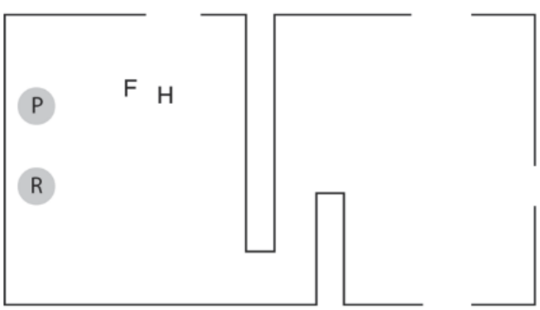
1. Ground truth:
   Arrows point to future positions of individuals.



Characteristics of the 10 individuals inside building:

ID	Status	Uniform Color	Armed	Direction	Firing
1	Friendly	Blue	Armed	Retreating	Not firing
2	Hostile	Green	Armed	Approaching	Not firing
3	Friendly	Blue	not armed	Approaching	Not firing
4	Hostile	Green	Armed	Stationary	Not firing
5	Friendly	Green	Not armed	Approaching	Not firing
6	Hostile	Green	Armed	Approaching	Not firing
7	Friendly	Blue	Not armed	Approaching	Not firing
8	Hostile	Red	Not armed	Approaching	Not firing
9	Hostile	Red	Armed	Approaching	Not firing
10	Friendly	Green	Not armed	Approaching	Not firing

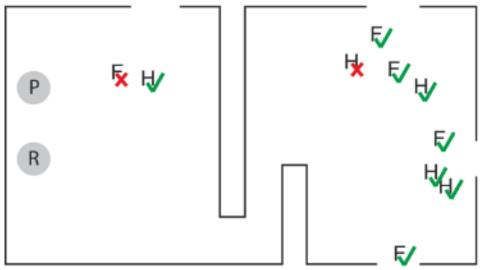
1. What the human can see:



Characteristics of the 2 individuals the human can see:

Uniform Color	Armed	Direction	Firing
Blue	Not armed	Approaching	Not firing
Green	Armed	Stationary	Not firing

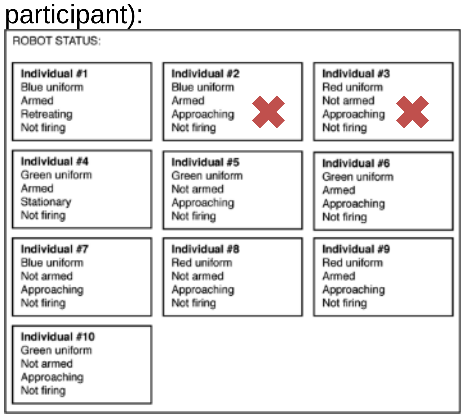
1. What the robot can see (in this scenario the robot has 80% reliability; red X indicates error in robot perception):

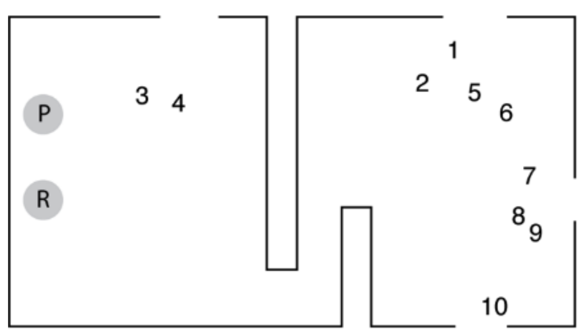


Characteristics as perceived by the robot:

ID	Status	Uniform Color	Armed	Direction	Firing
1	Friendly	Blue	Armed	Retreating	Not firing
2	**Friendly** (wrong)	**Blue** (wrong)	Armed	Approaching	Not firing
3	**Hostile** (wrong)	**Red** (wrong)	not armed	Approaching	Not firing
4	Hostile	Green	Armed	Stationary	Not firing
5	Friendly	Green	Not armed	Approaching	Not firing
6	Hostile	Green	Armed	Approaching	Not firing
7	Friendly	Blue	Not armed	Approaching	Not firing
8	Hostile	Red	Not armed	Approaching	Not firing
9	Hostile	Red	Armed	Approaching	Not firing
10	Friendly	Green	Not armed	Approaching	Not firing

2. What the robot reports to the human (red X indicates errors, which will not be shown to the participant):



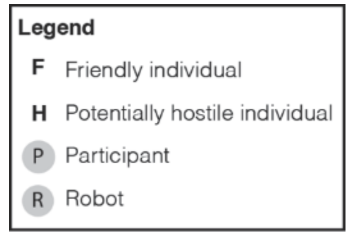


APPENDIX F:
EXAMPLES OF EXPERIMENTAL SCENARIOS

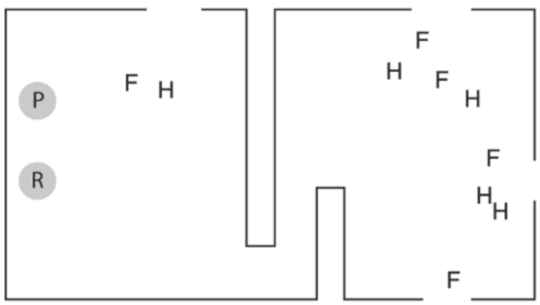
Scenario Example 1

Information acquisition with analysis (stage 2), 100% reliable robot, Poor (10%) potential unaided SA

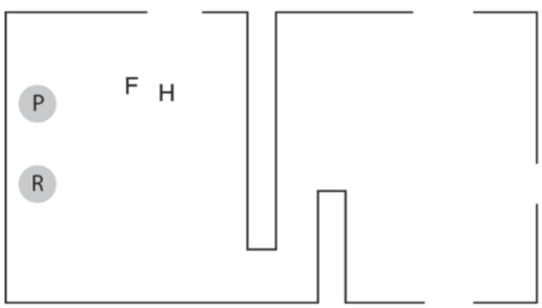
All images show the same point in time during the mission.



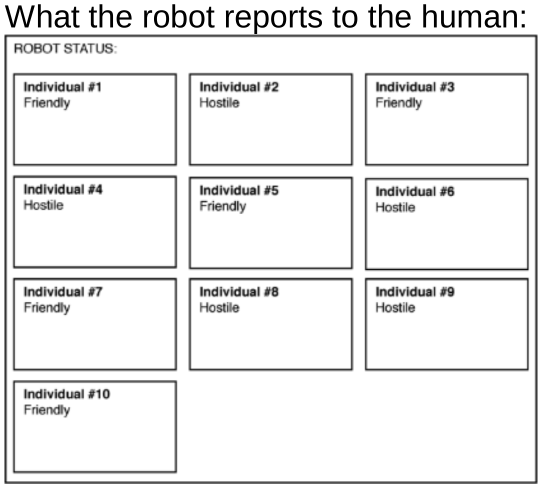
1. What the robot can see:

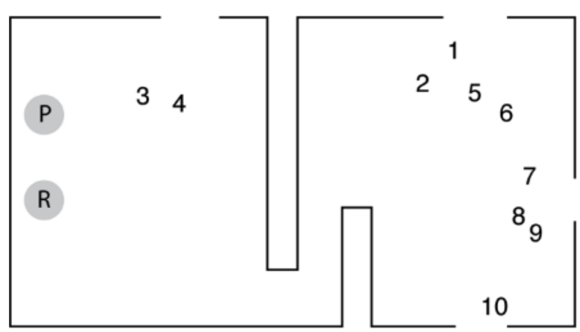


2. What the human sees:

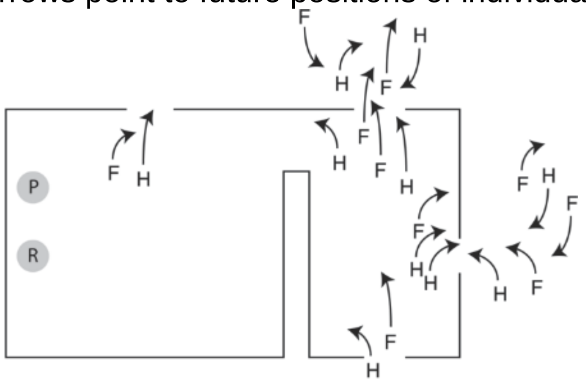


3. What the robot reports to the human:

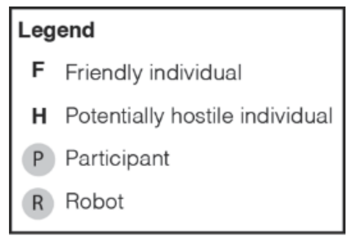




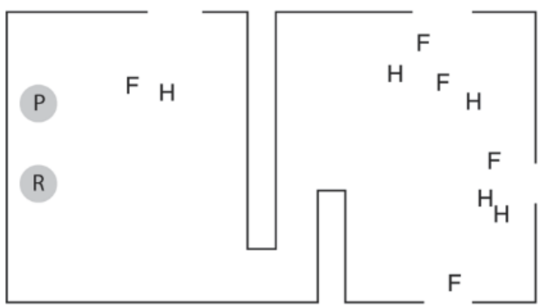
4. Ground truth:
   Arrows point to future positions of individuals.



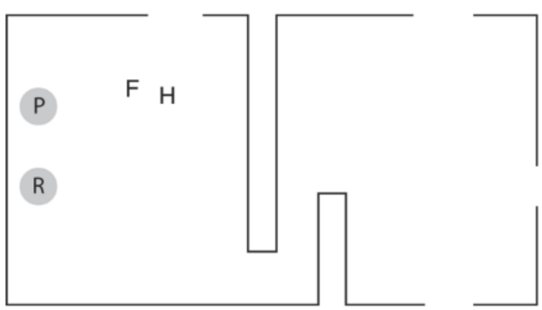
Scenario Example 2
Information acquisition with analysis (stage 2), 80% reliable robot, Poor (10%) potential unaided SA
All images show the same point in time during the mission.



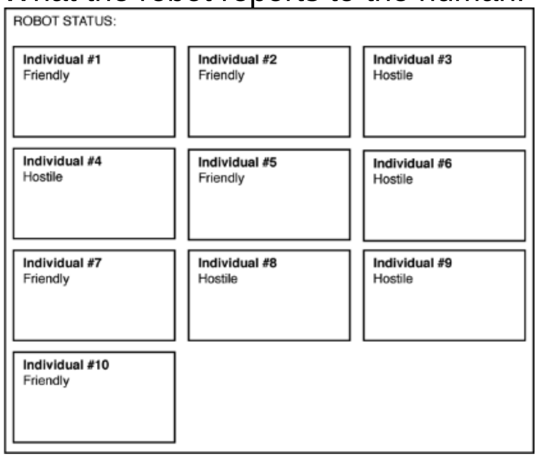
1. What the robot can see:

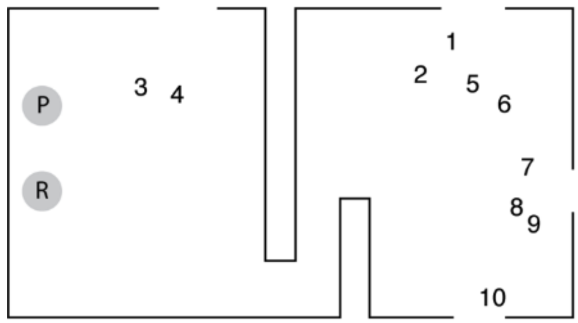


2. What the human sees:

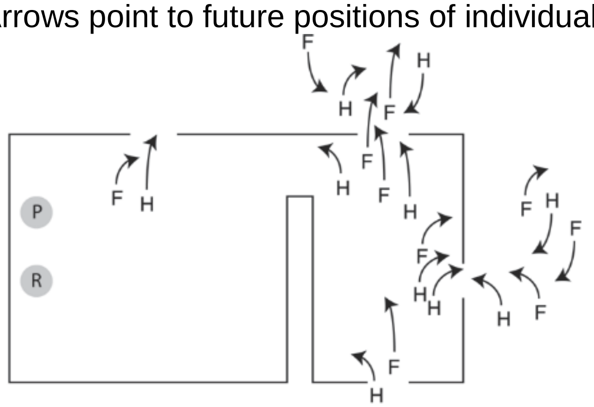


3. What the robot reports to the human:

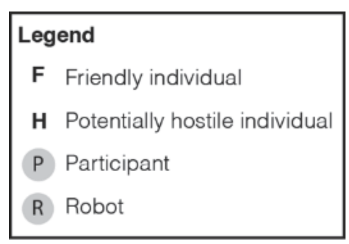




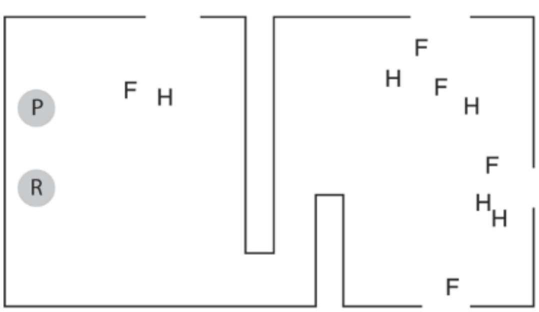
4. Ground truth:
   Arrows point to future positions of individuals.



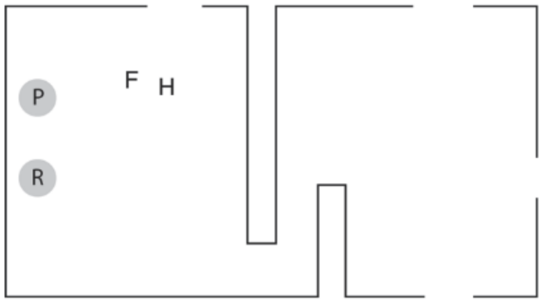
Scenario Example 3
Information acquisition with analysis (stage 2), 60% reliable robot, Poor (10%) potential unaided SA
All images show the same point in time during the mission.



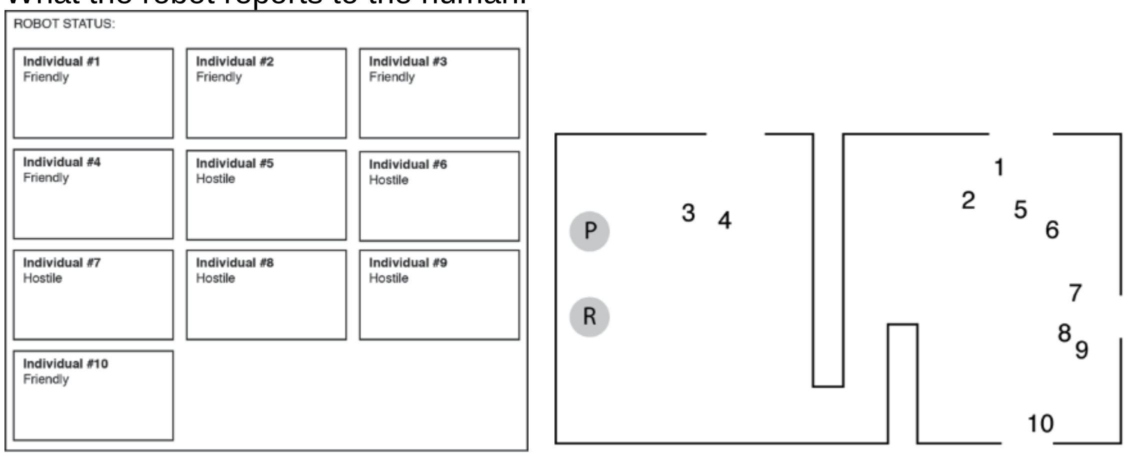
1. What the robot can see:



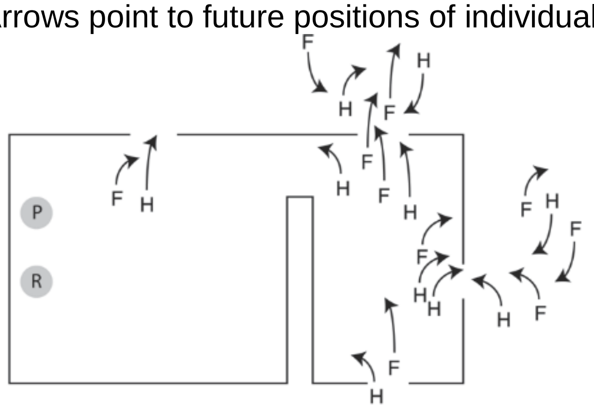
2. What the human sees:



3. What the robot reports to the human:



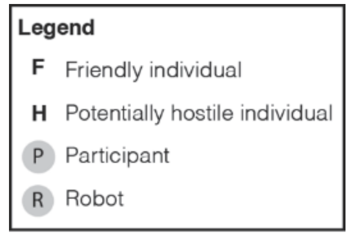
4. Ground truth:
Arrows point to future positions of individuals.



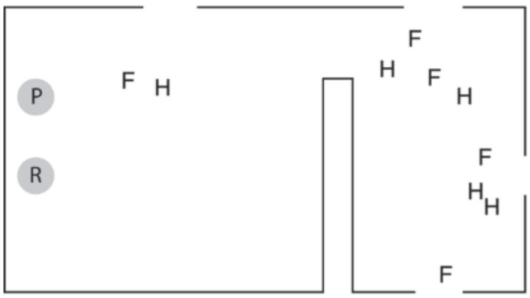
Scenario Example 4

Information acquisition with analysis (stage 2), 100% reliable robot, Moderate (50%) potential unaided SA

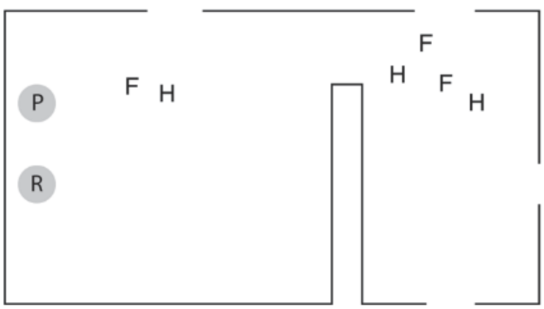
All images show the same point in time during the mission.



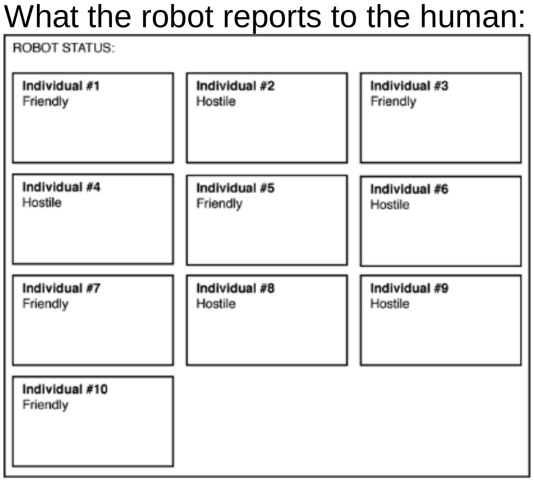
1. What the robot can see:

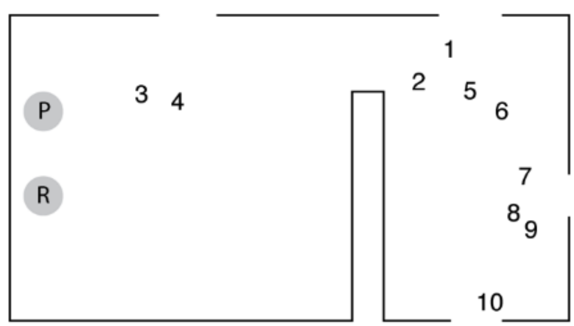


2. What the human sees:

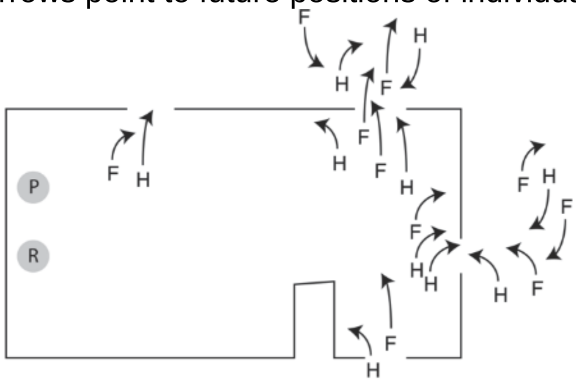


3. What the robot reports to the human:



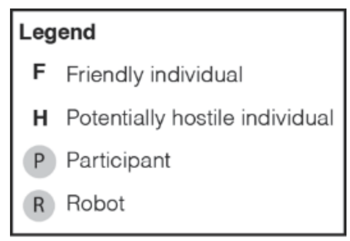


4. Ground truth:
   Arrows point to future positions of individuals.

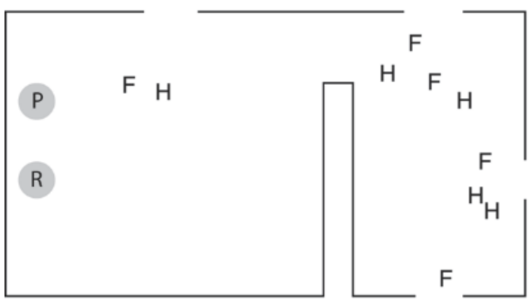


Scenario Example 5

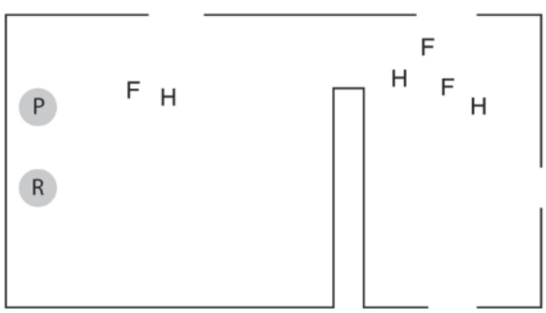
Information acquisition with analysis (stage 2), 80% reliable robot, Moderate (50%) potential unaided SA
All images show the same point in time during the mission.



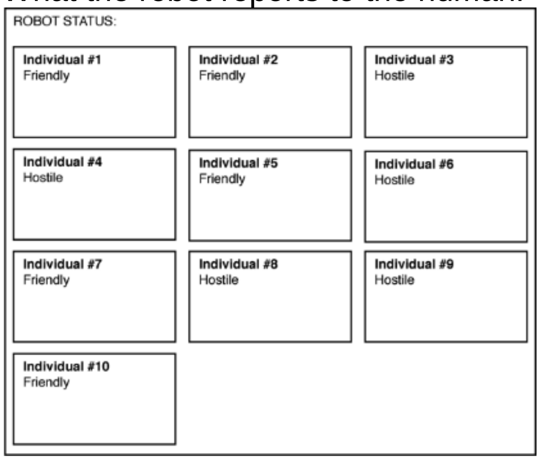
1. What the robot can see:

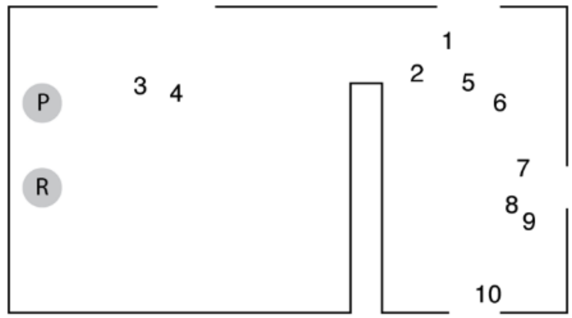


2. What the human sees:

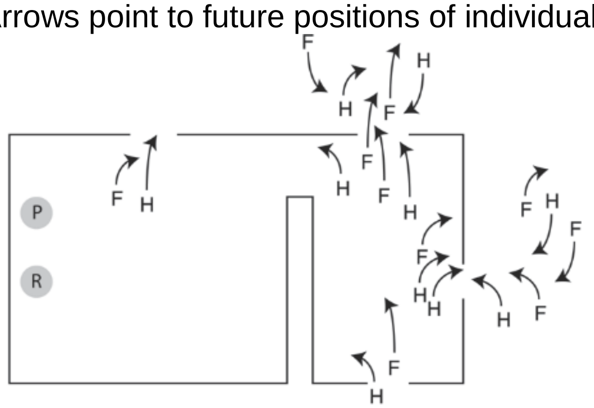


3. What the robot reports to the human:

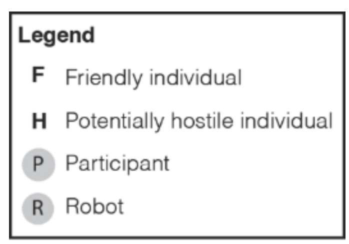




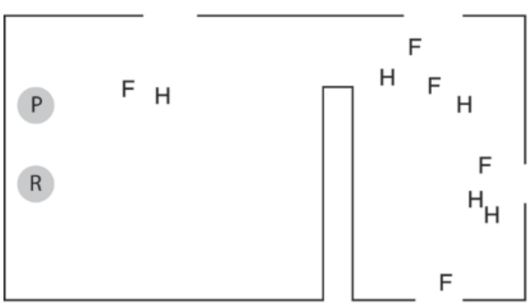
4. Ground truth:
   Arrows point to future positions of individuals.



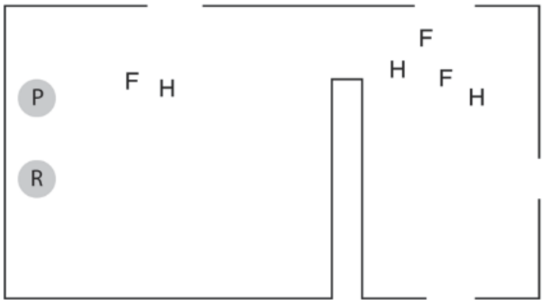
Scenario Example 6
Information acquisition with analysis (stage 2), 60% reliable robot, Moderate (50%) potential unaided SA
All images show the same point in time during the mission.



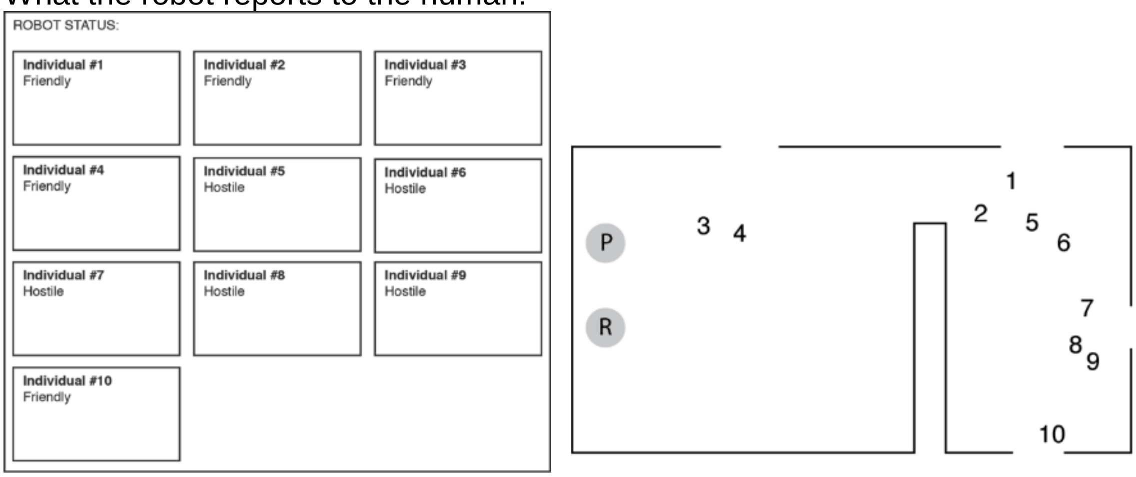
1. What the robot can see:



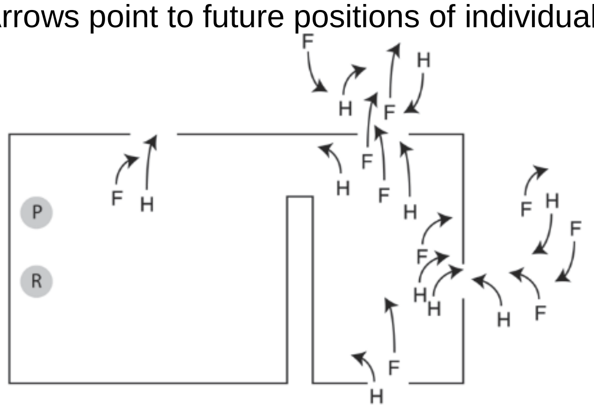
2. What the human sees:



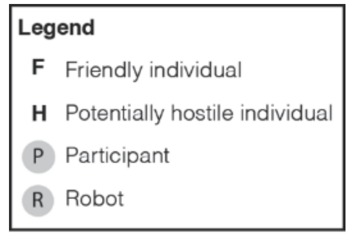
3. What the robot reports to the human:



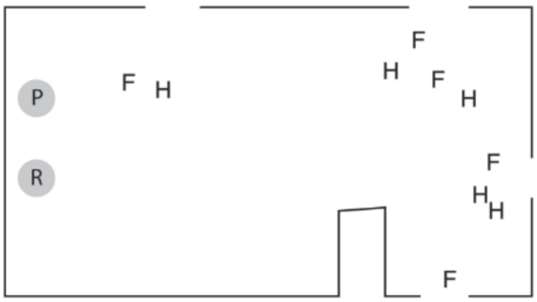
4. Ground truth:
Arrows point to future positions of individuals.



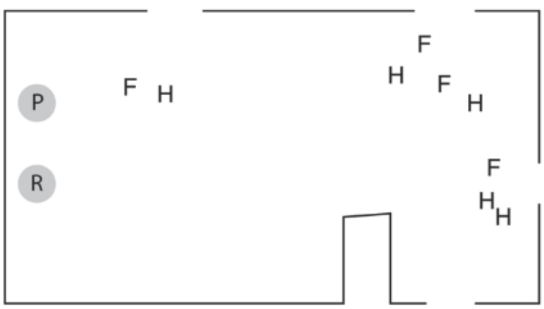
Scenario Example 7
Information acquisition with analysis (stage 2), 100% reliable robot, Good (90%) potential unaided SA
All images show the same point in time during the mission.



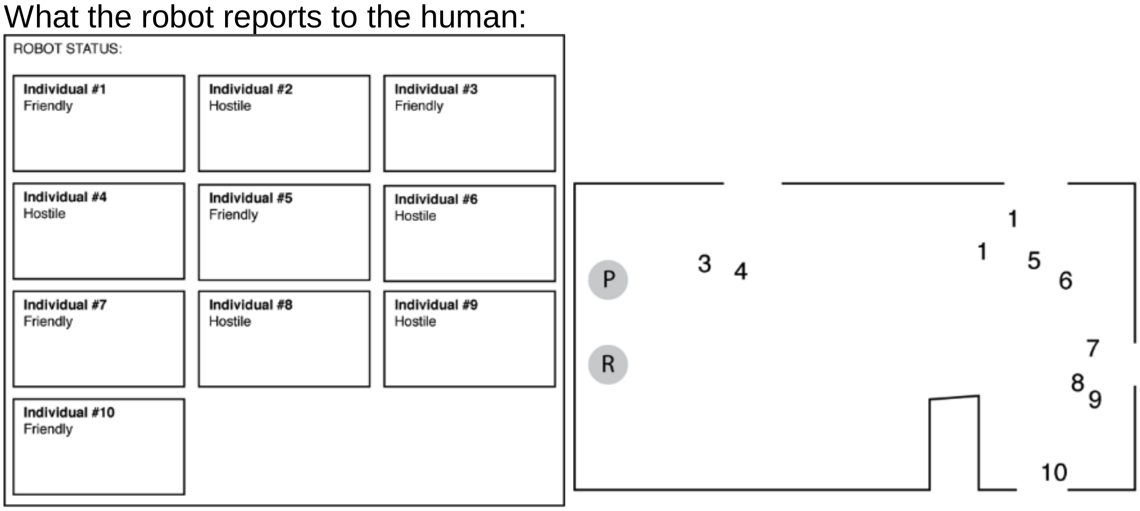
1. What the robot can see:



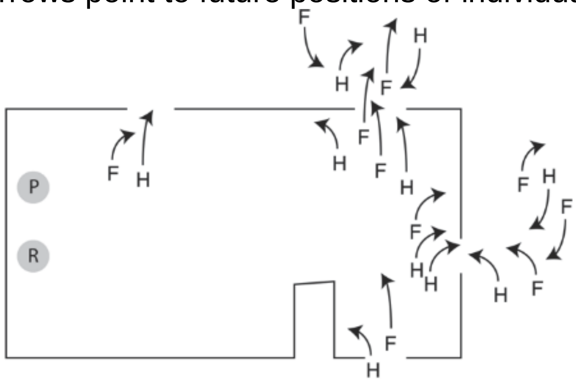
2. What the human sees:



3. What the robot reports to the human:



4. Ground truth:
   Arrows point to future positions of individuals.

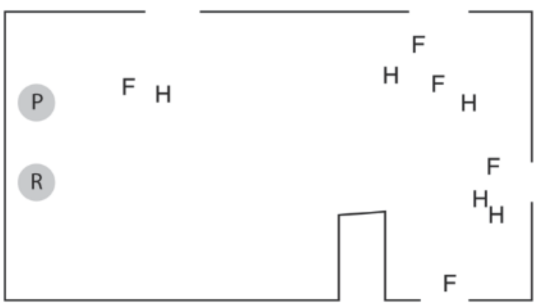


Scenario Example 8

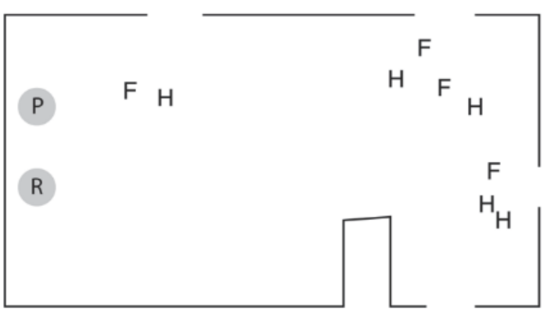
Information acquisition with analysis (stage 2), 80% reliable robot, Moderate (50%) potential unaided SA
All images show the same point in time during the mission.

Legend
- F Friendly individual
- H Potentially hostile individual
- P Participant
- R Robot

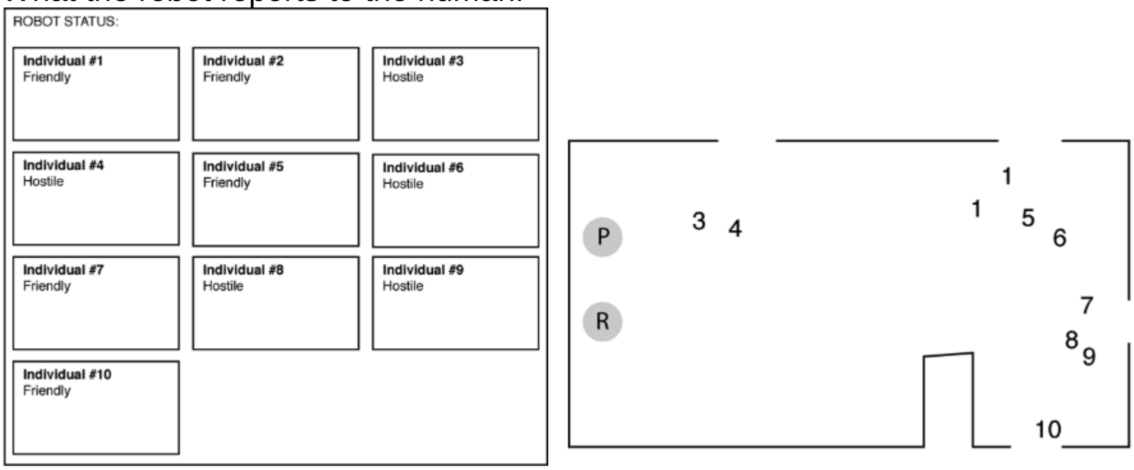
1. What the robot can see:



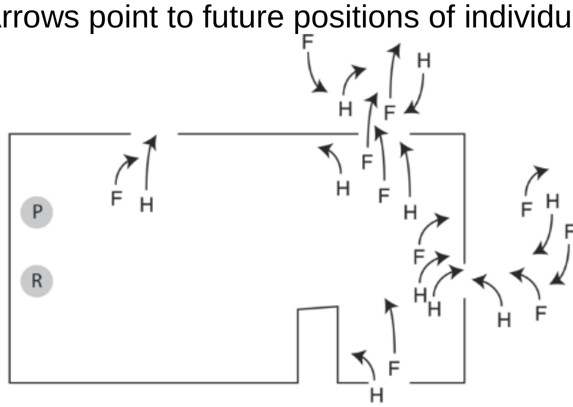
2. What the human sees:



3. What the robot reports to the human:

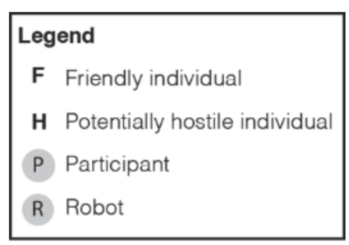


4. Ground truth:
   Arrows point to future positions of individuals.

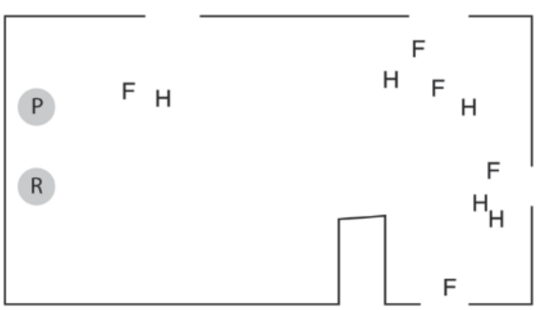


Scenario Example 9

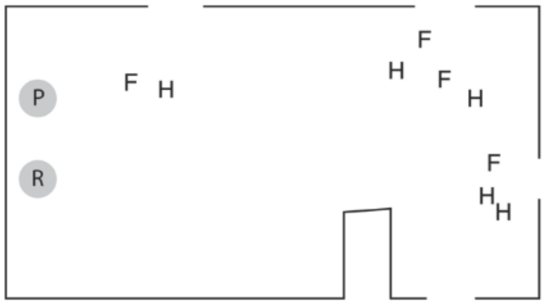
Information acquisition with analysis (stage 2), 60% reliable robot, Moderate (50%) potential unaided SA
All images show the same point in time during the mission.



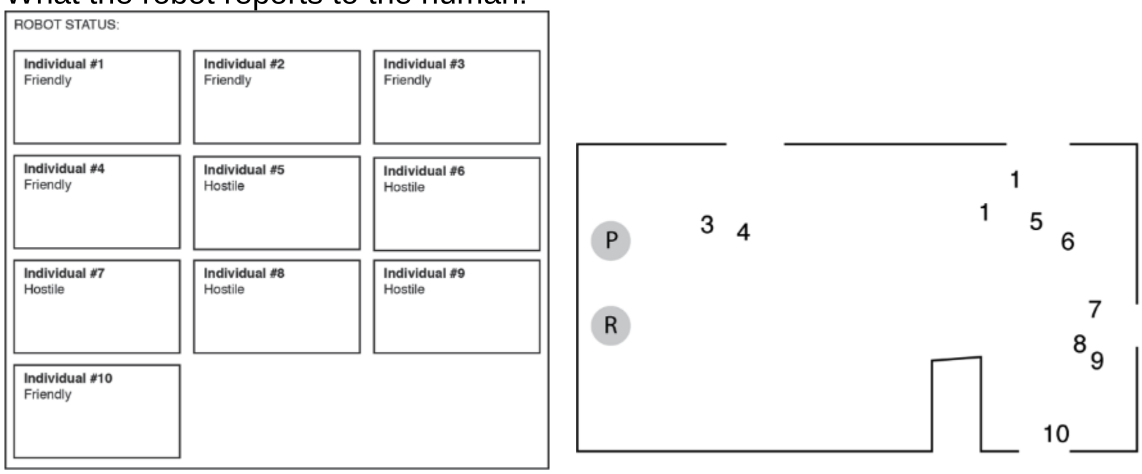
1. What the robot can see:



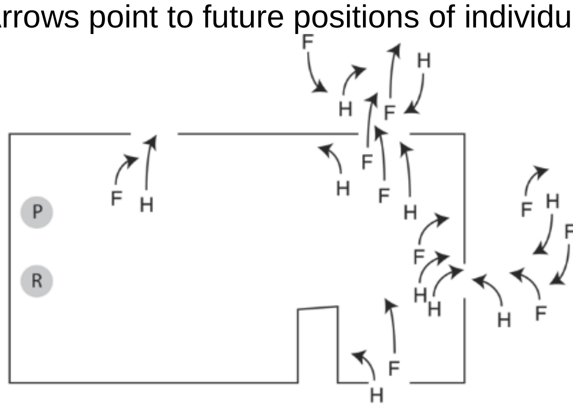
2. What the human sees:



3. What the robot reports to the human:



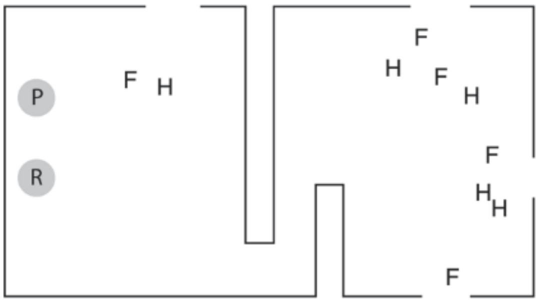
4. Ground truth:
Arrows point to future positions of individuals.



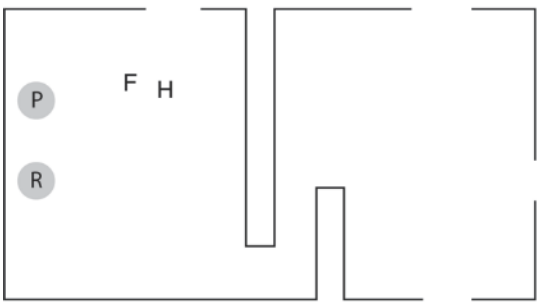
Scenario Example 10
Information acquisition (stage 1), 100% reliable robot, Poor (10%) potential unaided SA
All images show the same point in time during the mission.

Legend
F Friendly individual
H Potentially hostile individual
P Participant
R Robot

1. What the robot can see:



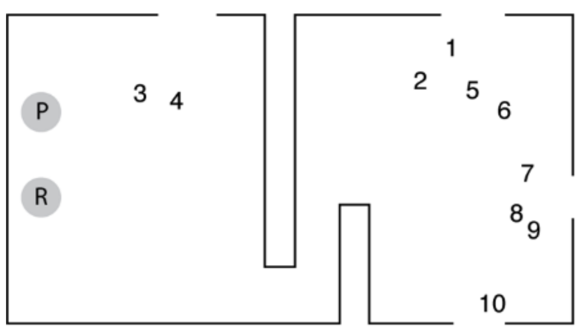
2. What the human sees:



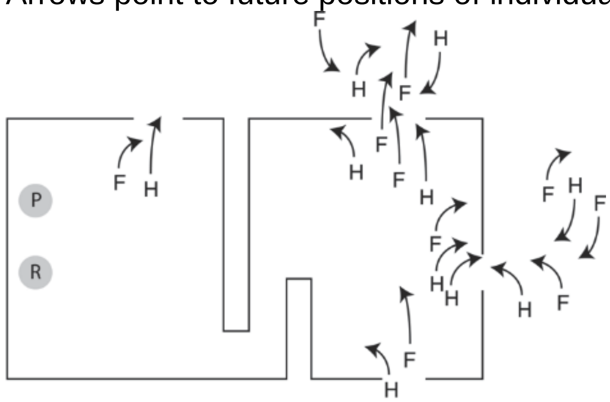
3. What the robot reports to the human:

ROBOT STATUS:

Individual #1 Blue uniform Armed Retreating Not firing	**Individual #2** Green uniform Armed Approaching Not firing	**Individual #3** Blue uniform Not armed Approaching Not firing
Individual #4 Green uniform Armed Stationary Not firing	**Individual #5** Green uniform Not armed Approaching Not firing	**Individual #6** Green uniform Armed Approaching Not firing
Individual #7 Blue uniform Not armed Approaching Not firing	**Individual #8** Red uniform Not armed Approaching Not firing	**Individual #9** Red uniform Armed Approaching Not firing
Individual #10 Green uniform Not armed Approaching Not firing		



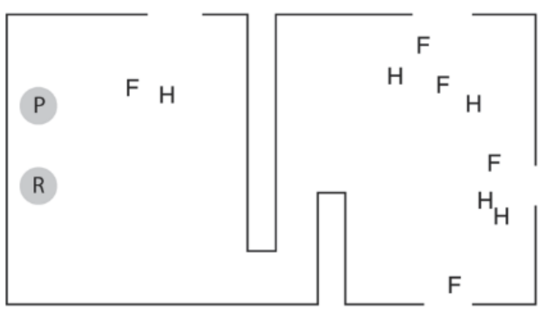
4. Ground truth:
   Arrows point to future positions of individuals.



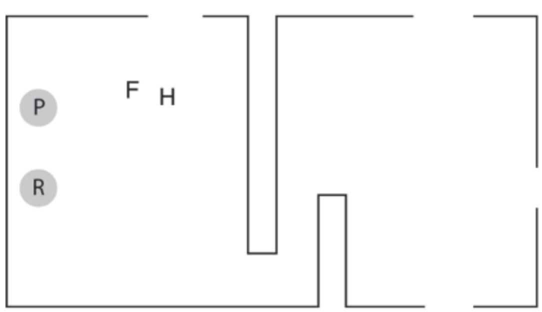
Scenario Example 11
Information acquisition (stage 1), 80% reliable robot, Poor (10%) potential unaided SA
All images show the same point in time during the mission.

Legend
- **F** Friendly individual
- **H** Potentially hostile individual
- **P** Participant
- **R** Robot

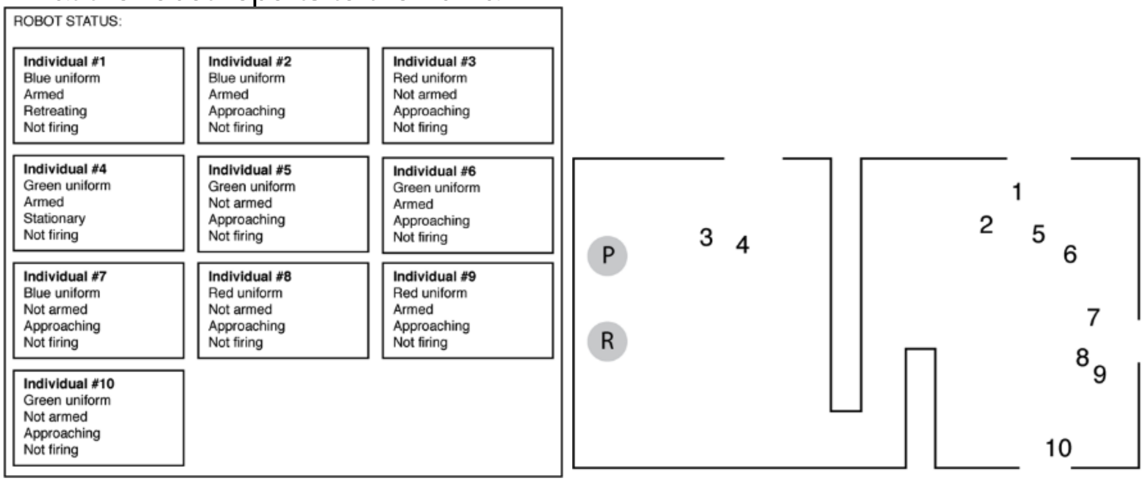
1. What the robot can see:



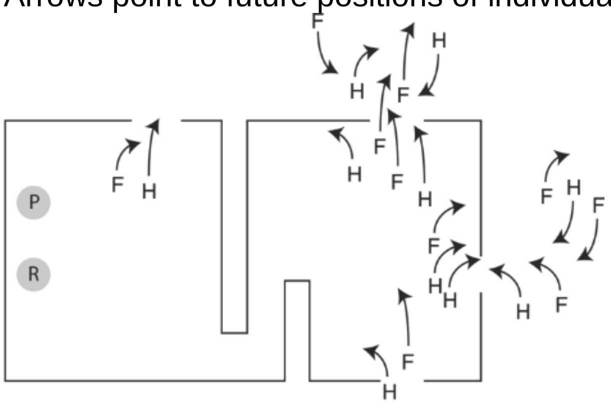
2. What the human sees:



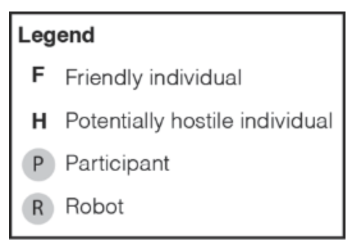
3. What the robot reports to the human:



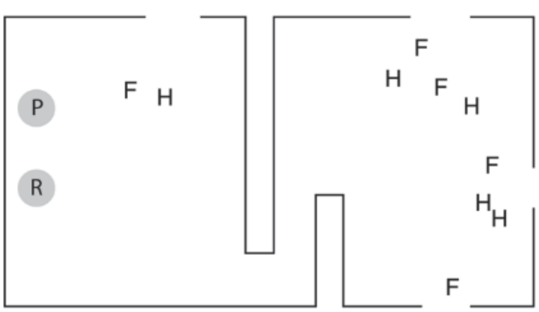
4. Ground truth:
Arrows point to future positions of individuals.



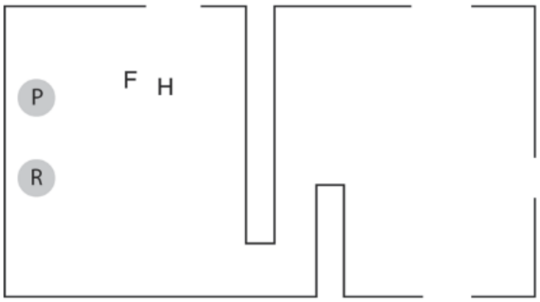
Scenario Example 12
Information acquisition (stage 1), 60% reliable robot, Poor (10%) potential unaided SA
All images show the same point in time during the mission.



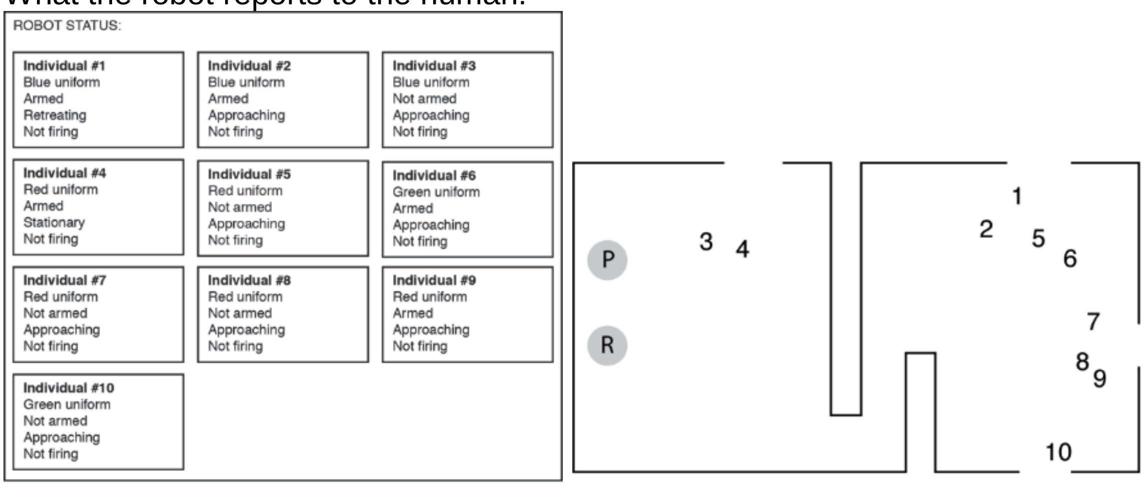
1. What the robot can see:



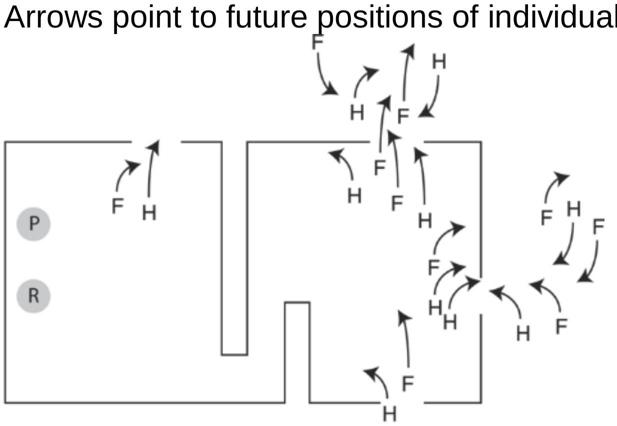
2. What the human sees:



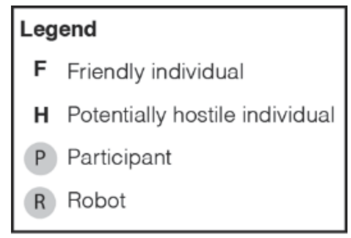
3. What the robot reports to the human:



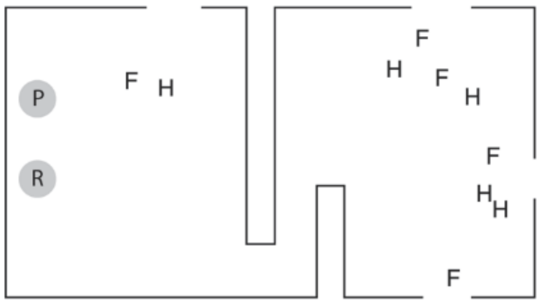
4. Ground truth:
Arrows point to future positions of individuals.



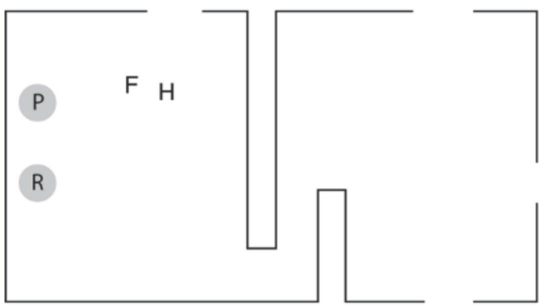
Scenario Example 13
Information acquisition (stage 1), 100% reliable robot,
Moderate (50%) potential unaided SA
All images show the same point in time during the mission.



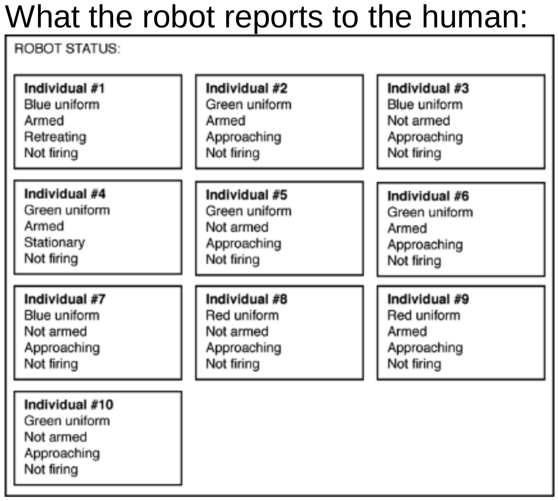
1. What the robot can see:



2. What the human sees:



3. What the robot reports to the human:


ROBOT STATUS:

Individual #1
Blue uniform
Armed
Retreating
Not firing

Individual #2
Green uniform
Armed
Approaching
Not firing

Individual #3
Blue uniform
Not armed
Approaching
Not firing

Individual #4
Green uniform
Armed
Stationary
Not firing

Individual #5
Green uniform
Not armed
Approaching
Not firing

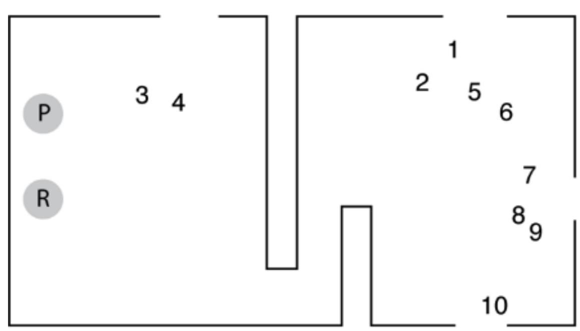
Individual #6
Green uniform
Armed
Approaching
Not firing

Individual #7
Blue uniform
Not armed
Approaching
Not firing

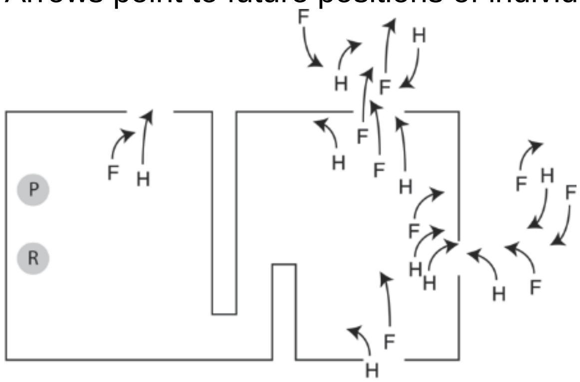
Individual #8
Red uniform
Not armed
Approaching
Not firing

Individual #9
Red uniform
Armed
Approaching
Not firing

Individual #10
Green uniform
Not armed
Approaching
Not firing



4. Ground truth:
   Arrows point to future positions of individuals.

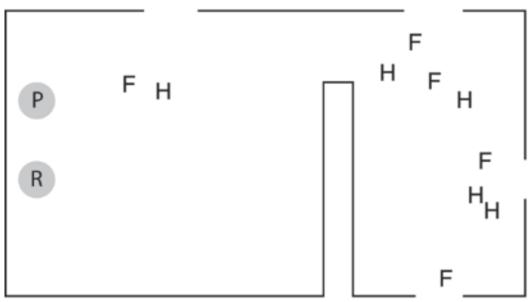


Scenario Example 14

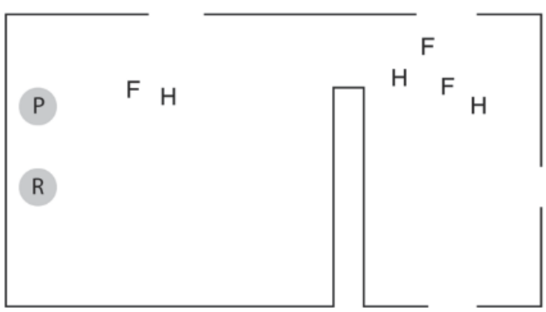
Information acquisition (stage 1), 80% reliable robot, Moderate (50%) potential unaided SA
All images show the same point in time during the mission.

Legend
- **F** Friendly individual
- **H** Potentially hostile individual
- P Participant
- R Robot

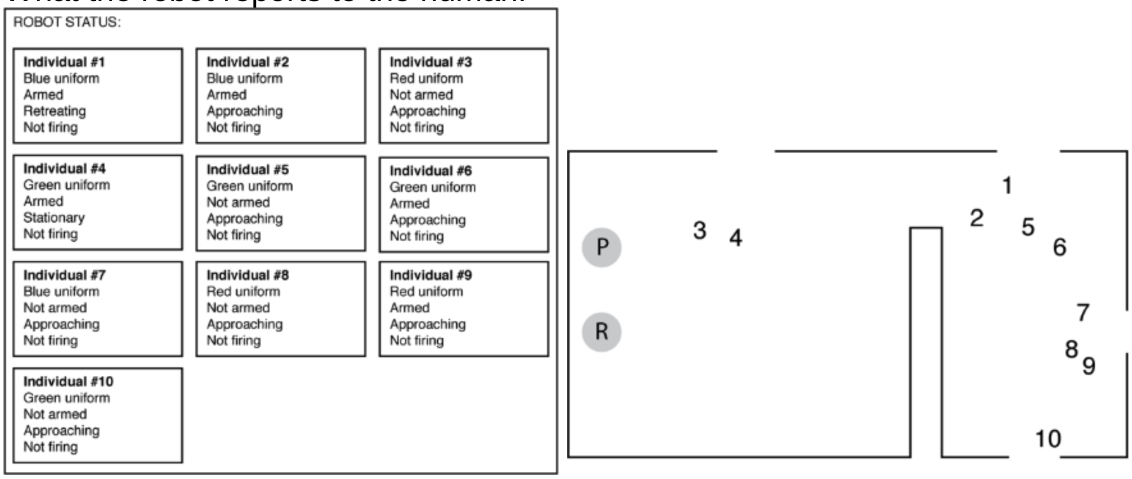
1. What the robot can see:



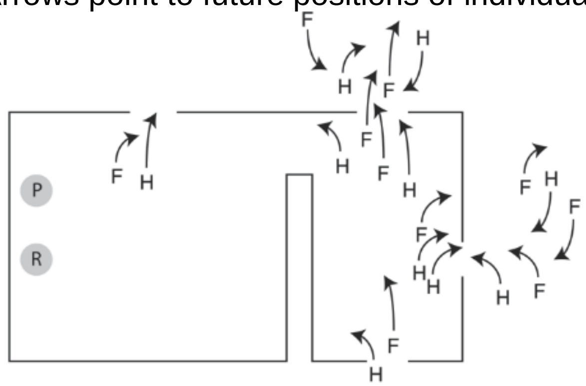
2. What the human sees:



3. What the robot reports to the human:

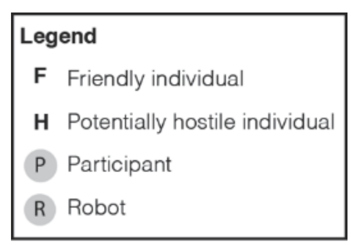


4. Ground truth:
   Arrows point to future positions of individuals.

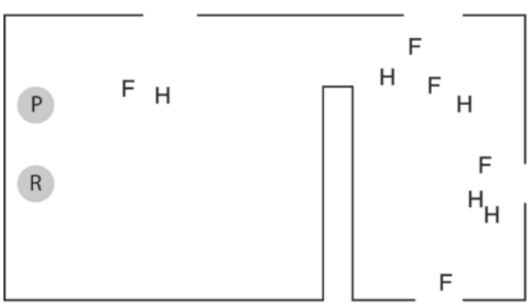


Scenario Example 15

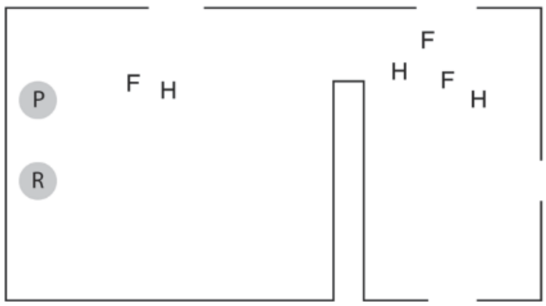
Information acquisition (stage 1), 60% reliable robot, Moderate (50%) potential unaided SA
All images show the same point in time during the mission.



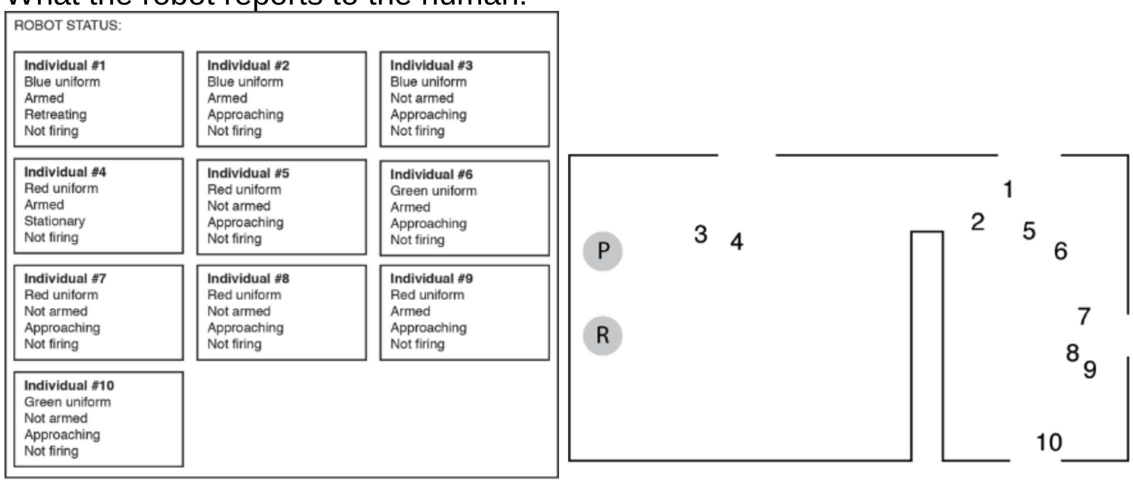
1. What the robot can see:



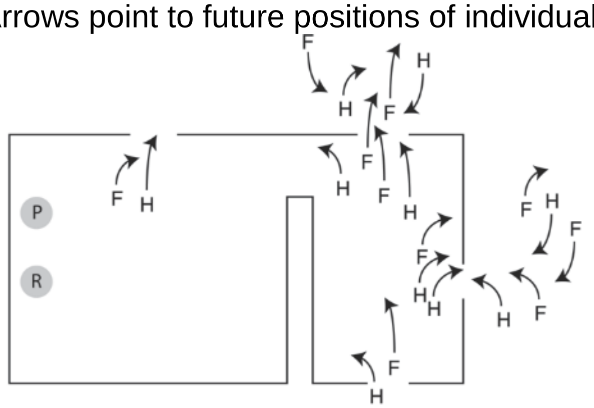
2. What the human sees:



3. What the robot reports to the human:



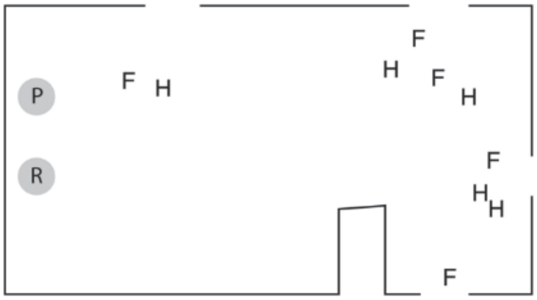
4. Ground truth:
Arrows point to future positions of individuals.



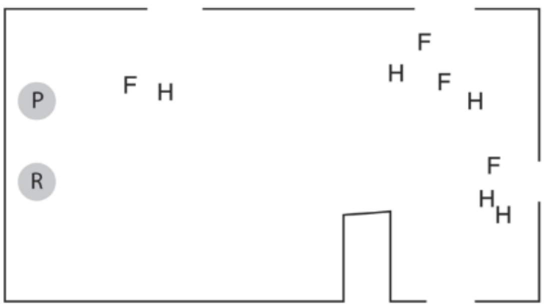
Scenario Example 16

Information acquisition (stage 1), 100% reliable robot, Good (90%) potential unaided SA
All images show the same point in time during the mission.

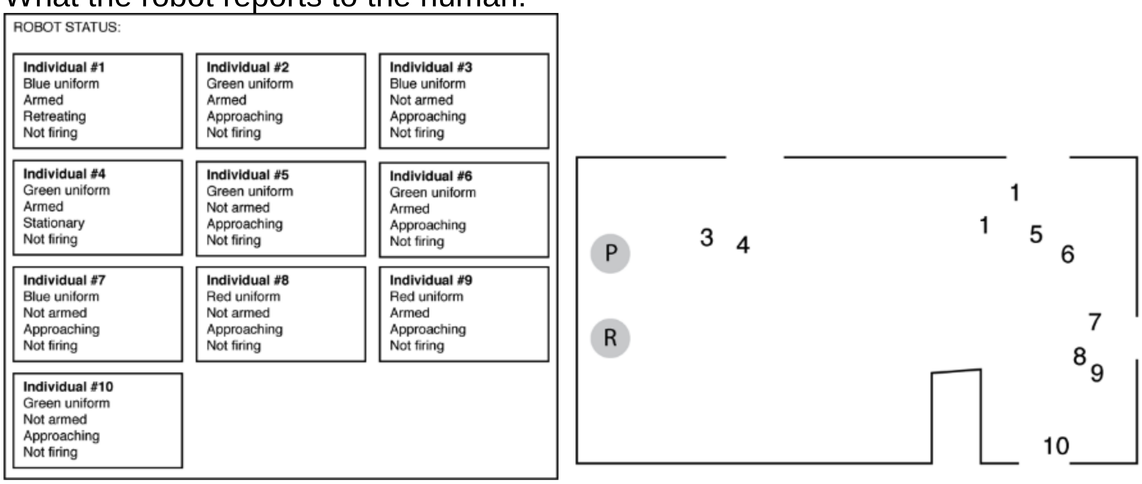
1. What the robot can see:



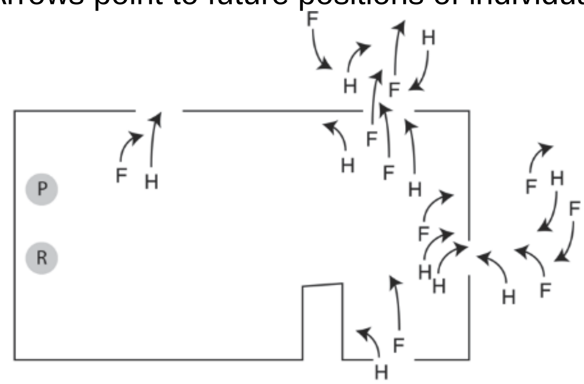
2. What the human sees:



3. What the robot reports to the human:

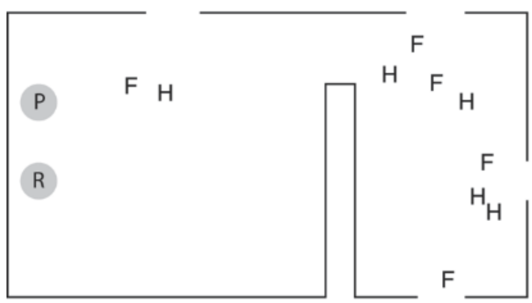


4. Ground truth:
   Arrows point to future positions of individuals.

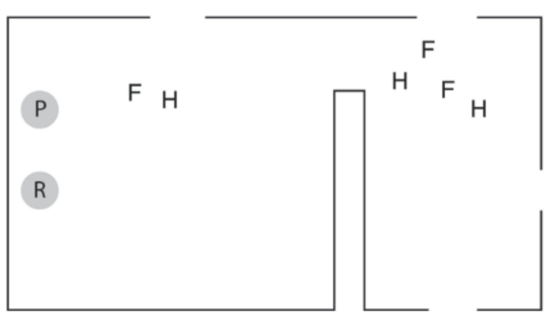


Scenario Example 17
Information acquisition (stage 1), 80% reliable robot, Moderate (50%) potential unaided SA
All images show the same point in time during the mission.

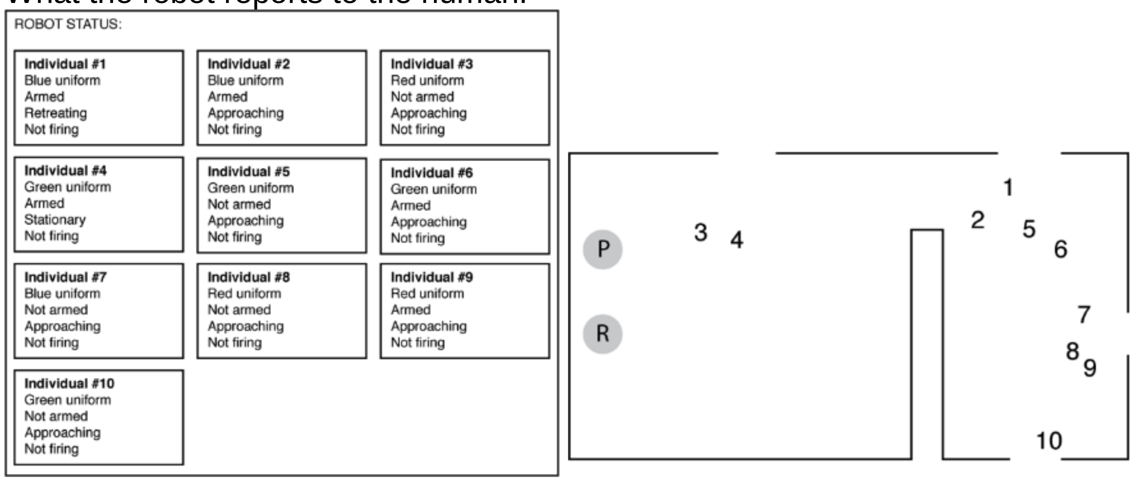
1. What the robot can see:



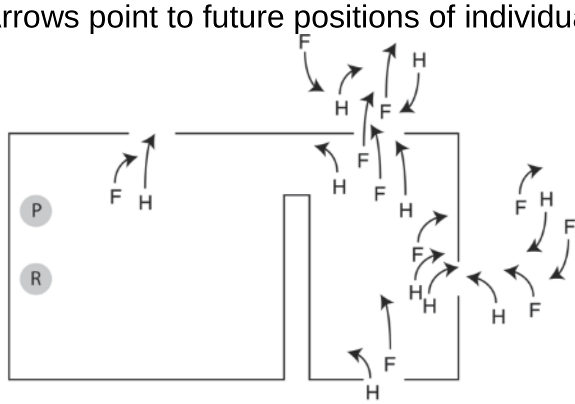
2. What the human sees:



3. What the robot reports to the human:



4. Ground truth:
   Arrows point to future positions of individuals.

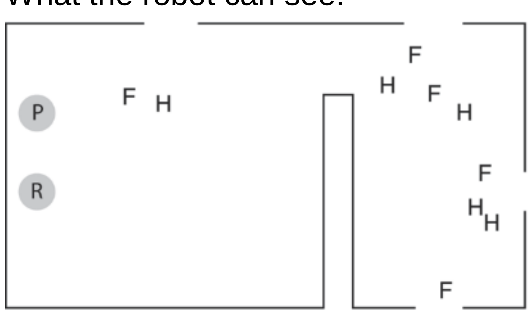


Scenario Example 18

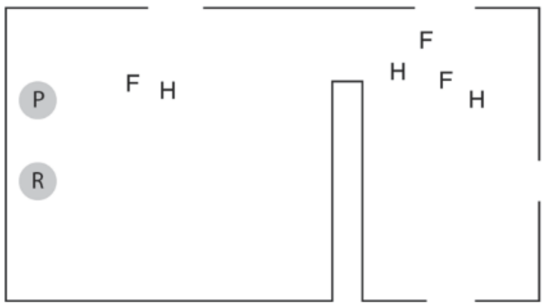
Information acquisition (stage 1), 60% reliable robot, Moderate (50%) potential unaided SA

All images show the same point in time during the mission.

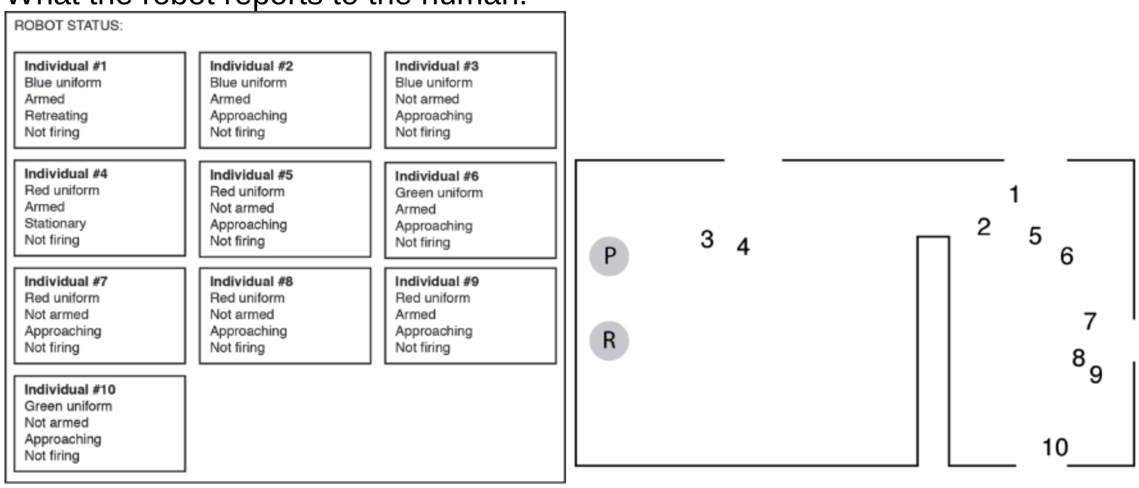
1. What the robot can see:



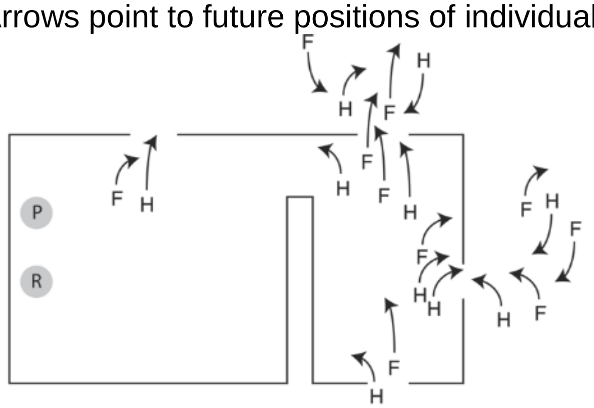
2. What the human can see:



3. What the robot reports to the human:



4. Ground truth:
   Arrows point to future positions of individuals.



APPENDIX G: INFORMED CONSENT

University of Central Florida

The effects of diagnostic aiding on situation awareness under unreliability

Informed Consent

Principal Investigator:	David Schuster, M.A.
Sub-Investigators:	Andrew B. Talone Elizabeth Phillips Scott Ososky Adriane Smith (Research assistant)
Faculty Supervisor:	Florian Jentsch, Ph.D.
Sponsor:	This research is being funded by the United States Department of Defense, the United States Army Research Laboratory, and General Dynamics
Investigational Sites:	University of Central Florida Main Campus 4000 Central Florida Blvd., Orlando, FL 32817 Psychology Building, Room 303G
OR	
	Institute for Simulation & Training 3100 Technology Parkway Orlando, FL 32826 Partnership II Building room 117A

Introduction:
Researchers at the University of Central Florida (UCF) study many topics. To do this we need the help of people who agree to take part in a research study. You are being invited to take part in a research study, which will include about 50 people at UCF. You have been asked to take part in this research study because you are an undergraduate student enrolled in a psychology class at UCF. You must be 18 years of age or older to be included in the research study.

The person doing this research is David Schuster of the College of Sciences at UCF. Because the researcher is a graduate student, Dr. Florian Jentsch, a UCF faculty supervisor in the Department of Psychology, is supervising him. UCF students learning about research are helping to do this study as part of the research team. Their names are: Andrew B. Talone, Elizabeth Phillips, and Scott Ososky, and Adriane Smith.

What you should know about a research study:

- Someone will explain this research study to you.
- A research study is something you volunteer for.
- Whether or not you take part is up to you.
- You should take part in this study only because you want to.
- You can choose not to take part in the research study.
- You can agree to take part now and later change your mind.
- Whatever you decide it will not be held against you.
- Feel free to ask all the questions you want before you decide.

Purpose of the research study:
The purpose of this study is to determine how a person's situation awareness is affected by an unreliable robot at different levels of unaided human performance and at different levels of robot assistance. Past studies have looked at these effects, but they either were not specific to robots or did not examine different levels of unaided human performance. The long-term goal of this research effort is to provide understanding of how robots can support operator situation awareness and offer guidelines, based on the findings, for the development of future Army robots so that humans are most likely to use them effectively.

What you will be asked to do in the study:
You will be randomly assigned to one of three reliability conditions. These conditions involve the frequency with which the robot provides incorrect information to the participant. Once assigned to a condition, you will not be permitted to transfer into another condition. To begin the study you will be asked fill out a biographical questionnaire and complete a test of spatial ability. Then you will be familiarized with the nature of the task that you will be asked to complete. You will then be asked to complete a practice mission before completing 12 live missions. During each mission, you will be working with a simulated robot to identify individuals located within a building. You will view a camera feed of one room of the building, and the robot will provide status updates about people inside and outside of the room, but within the building. Randomly during each mission, you will be asked to report on your knowledge of the individuals inside the building. At the conclusion of each mission, you will be asked to fill out a questionnaire asking you to rate your awareness during the mission. Once all missions have been completed, you will be given post-participation information regarding the nature of the study and its purposes as well as contact information for the principal investigator and faculty advisor should you wish to find out about the results of the study in the future or if you would like to address general comments or concerns. In addition, you will be given the opportunity to evaluate the research team on their performance and treatment of you as a research participant. The completion of this evaluation form is entirely optional, however. A general description of the course of the study is below.

1. Informed consent	
2. Biographical data form	5 min
3. Guildford Zimmerman Spatial Orientation test	10 min
4. Training session explaining how to complete missions.	12 min
5. Practice mission	6 min
6. Mission presentations (1-12)	74 min
i. SPAM measure is presented during the mission.	
ii. SART measure at the end of each mission.	
7. Post-participation information	5 min
8. Presentation of researcher contact information	
9. Researcher evaluation form [optional]	5 min
Anticipated time to completion	**117-120 min**

Please note that you do not have to answer every question or complete every task. You will not lose any benefits if you skip questions or tasks.

Location:

The research will be conducted at **either** the Psychology building at UCF room 303G **or** the UCF Institute of Simulation and Training Partnership II building located at 3100 Technology Parkway, Room 117A, Orlando, FL 32826.

Time required:
We expect that you will be in this research study for up to approximately 2 hours in one session.

Funding for this study:
This research study is being paid for by the United States Army Research Laboratory and General Dynamics as a part of the Robotics Collaborative Technology Alliance.

Risks:
Researchers believe that the likelihood of participant risk is very low. However, there may be concern that a military scenario or the suggestion of a robotic teammate may invoke a negative response to those sensitive to issues associated with military conflict or artificial intelligence. Other risks associated with participation in this research study are unforeseeable.

Benefits:
We cannot promise any benefits to you or others from your taking part in this research. Participants will be immersed in an environment of scholarly research, which may help to augment their research education.

Compensation or payment:
Participants may expect to spend approximately 120 minutes performing experimental tasks, for which they may receive Sona Systems credit for the amount of time they participate. Maximum Sona Systems credit will be 2 credits, the equivalent of 120 minutes. There is no direct compensation for taking part in this study. It is possible, however, that extra credit may be offered for your participation, but this benefit is at the discretion of your instructor.

Confidentiality:
We will limit your personal data collected in this study to people who have a need to review this information. However, due to the requirement to come into the research facility, we cannot promise complete secrecy. Additionally, you are required to sign the Informed Consent document. This document will not be attached to the data we collect from you, however we will still be in the possession of the Informed Consent document which may have to be disclosed to the IRB. In addition, because this research is sponsored by the Department of Defense and the U.S. Army, the Army Human Research Protections Office is eligible to review the research records.

Study contact for questions about the study or to report a problem:
If you have questions, concerns, complaints, or think the research has hurt you, please feel free to contact David Schuster, a graduate student in the Applied Experimental and Human Factors Psychology Ph.D. program, or, the principal investigator and faculty supervisor, Dr. Florian Jentsch, in the Department of Psychology. Their contact information is as follows:

David Schuster (Principal Investigator)
dschuster@ist.ucf.edu
AEHF Psychology Ph.D. program
Department of Psychology, UCF

Dr. Florian Jentsch (Faculty Advisor)
Florian.Jentsch@ucf.edu
Department of Psychology
(407) 882-0304

IRB contact about your rights in the study or to report a complaint:
Research at the University of Central Florida involving human participants is carried out under the oversight of the Institutional Review Board (UCF IRB). This research has been reviewed and approved by the IRB. For information about the rights of people who take part in research, 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. You may also talk to them for any of the following:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You want to get information or provide input about this research.

Withdrawing from the study:
Participation in this research study is completely voluntary. Refusal to participate or choosing to withdraw from the study at any time will involve no penalty or loss of benefits associated with the study. If you decide to leave the research study, you will receive credit in accordance with time spent in the study prior to withdrawal. For example, if you decide to withdrawal from the study after 1 hour has passed, you will receive 1 Sona credit. If you decide to leave the study, contact the investigator so that the investigator can stop administering the survey instruments or stop the simulation. The researcher will then thank you for your time, explain how you will be credited for your time, and instruct you to the exit. The person in charge of the research study or the sponsor can remove you from the research study without your approval. Possible reasons for removal include failure to follow instructions of the research staff, disorderly conduct, improper treatment of the research staff or other participants, or if the research staff feels that the study is no longer in your best interests. We will tell you about any new information that may affect your health, welfare or choice to stay in the research.

Your signature below indicates your permission to take part in this research.

DO NOT SIGN THIS FORM AFTER THE IRB EXPIRATION DATE BELOW

____
Name of participant

______________________________ ______________
Signature of participant Date

______________________________ ______________
Signature of person obtaining consent Date

____
Printed name of person obtaining consent

UCF
University of Central Florida IRB
IRB NUMBER: SBE-13-09053
IRB APPROVAL DATE: 1/23/2013
IRB EXPIRATION DATE: 1/22/2014

APPENDIX H: POST-PARTICIPATION INFORMATION

The effects of diagnostic aiding on situation awareness under unreliability

Post-Participation Information

This study was designed to investigate the way in which people maintain awareness when working with a robotic teammate. We want to see if people benefit from a robot that integrates information for you more than a robot that provides raw data. We want to see if this type of robot is helpful when it is unreliable, and whether an unreliable robot is useful if the task is very difficult. Understanding how people maintain awareness will help the scientific community develop artificial intelligence in robots so that future robots can work better with their human teammates.

We could not do our research without your help, and your participation is greatly appreciated. Please ask any questions you may have about the procedure or the study in general. If you want to learn more about the study or receive the results of the study when they become available, please contact the principal investigator, David Schuster.

The data you have contributed to this study will be held in strict confidentiality by the researchers and will not be revealed to anyone other than the researchers and their immediate assistants.

Thank you again for your participation.

Contact Information:

David Schuster, M.A. (Principal investigator)
Institute for Simulation and Training
University of Central Florida
Phone: 407-374-3283
E-mail: dschuster@ist.ucf.edu

Dr. Florian Jentsch (Faculty advisor)
Director, Team Performance Laboratory
Institute for Simulation and Training
University of Central Florida
Phone: 407-882-0304
E-mail: florian.jentsch@ucf.edu

APPENDIX I: PARTICIPANT INSTRUCTION VIDEO SLIDES

Mission Training Video

Background

- Your mission: identify all the people in a building.
- 12 missions (5 minutes each)
- Questions will be asked during and after the missions.

2

Your goal: identification

3

Identification

- Everyone is either FRIENDLY or HOSTILE.
- 4 characteristics tell you if people are FRIENDLY or HOSTILE.

4

The 4 characteristics of people in the building...

1. Uniform color:

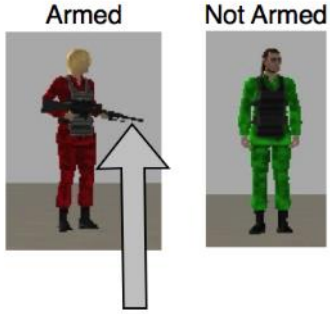
Red Green Blue

5

The 4 characteristics of people in the building...

2. Either:

Armed Not Armed

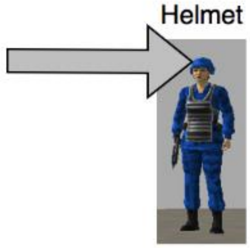


6

The 4 characteristics of people in the building...

3. Either:

Helmet



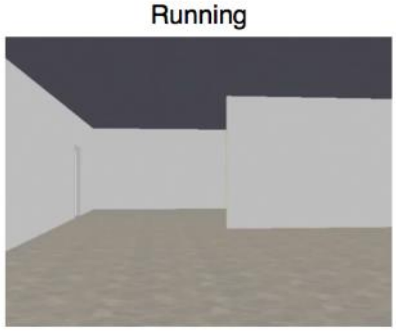
No Helmet

7

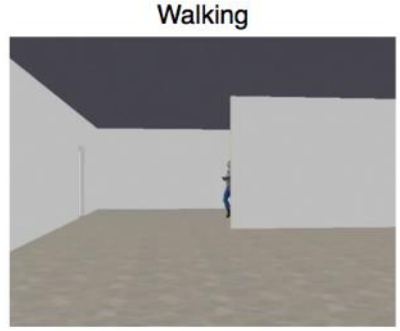
The 4 characteristics of people in the building...

4. Either:

Running



Walking



8

When is someone HOSTILE or FRIENDLY?

9

When is someone HOSTILE or FRIENDLY?

- An individual is HOSTILE in one of these cases:

green uniform +
armed

armed +
wearing a helmet

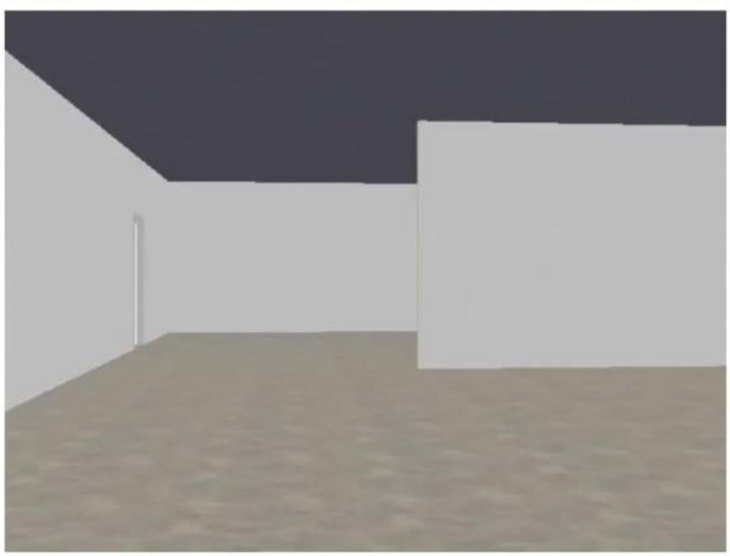
red uniform +
running

10

Try two examples...

11

FRIENDLY or HOSTILE?



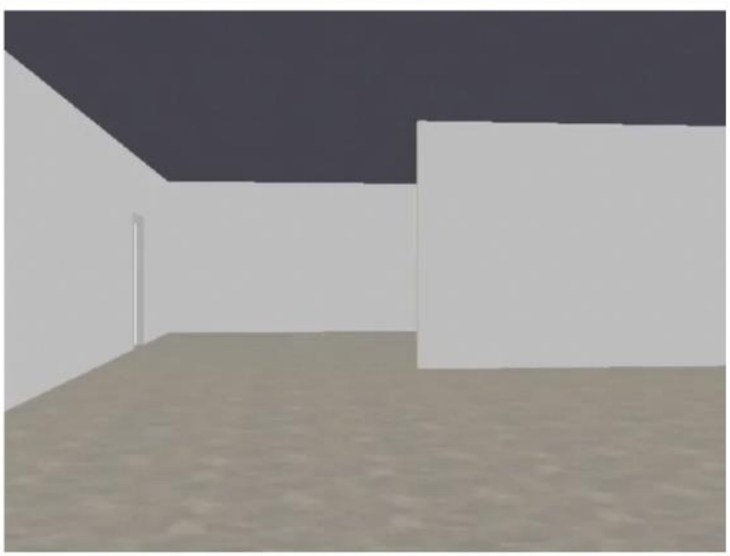
12

FRIENDLY or HOSTILE?

- This person is HOSTILE.
- Red uniform + running means HOSTILE.

13

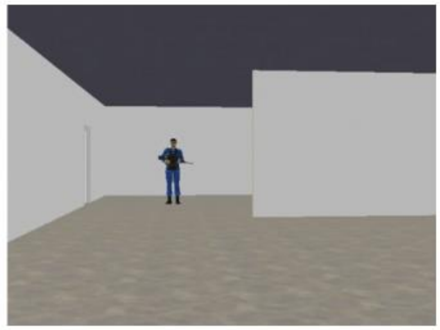
FRIENDLY or HOSTILE?



14

FRIENDLY or HOSTILE?

- This person is FRIENDLY.
- Armed, but not green.
- Armed, but not wearing a helmet.
- Running, but wearing a blue uniform.



15

To review

16

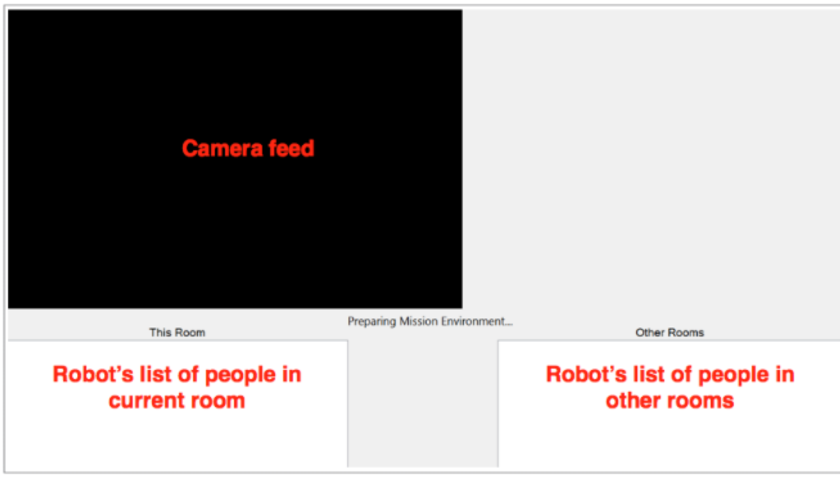
You need information about the entire building, but you can only see one room.

17

Your robot

18

Your mission display



More about your robot

- The robot sometimes provides characteristics:

Person #1-- Blue uniform. Not armed. Running. No helmet.

- Other times it provides a conclusion:

Person #2-- Hostile.

Awareness questions

- Each mission will randomly pause multiple times.
- You will be asked:
  - ...how many FRIENDLY and HOSTILE are in the building right now
  - ...about the last individual to enter the building
- Always answer with the current state of the building.

1

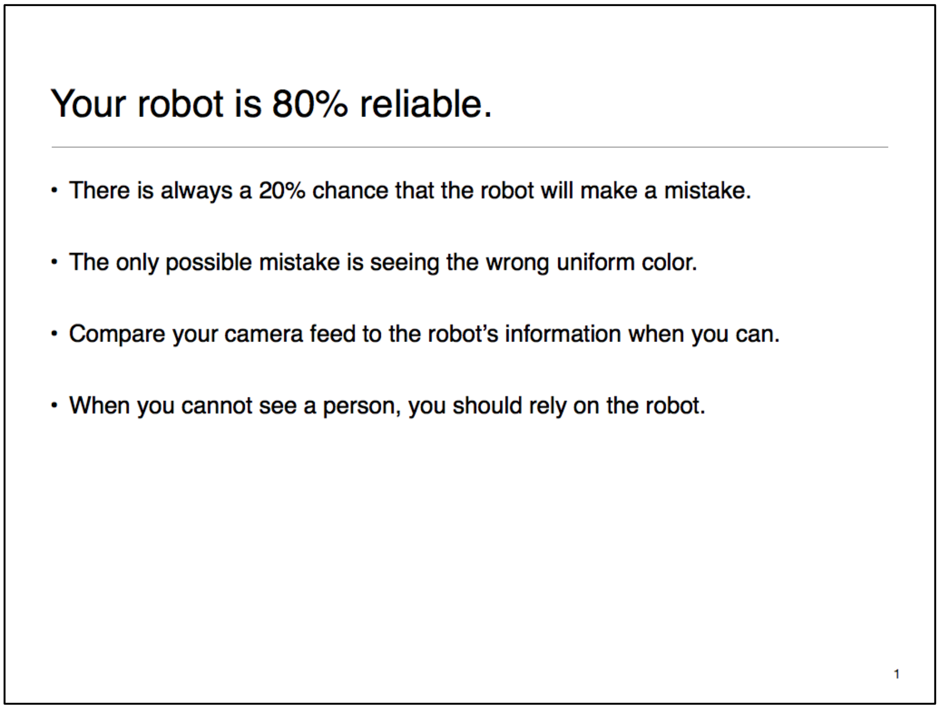
100% Reliability Only

Your robot is perfectly reliable.

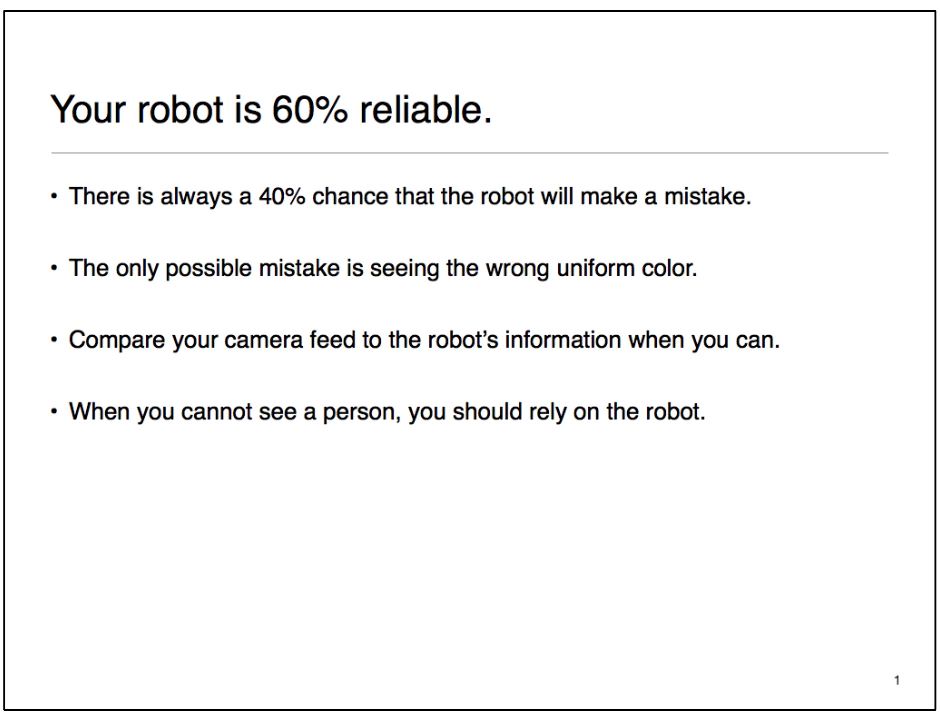
- The robot does not make mistakes.
- You should rely on the robot as much as possible.

1

80% Reliability Only



60% Reliability Only



To Review

- You will be doing 2 practice missions followed by 12 live missions.
- You will be asked questions randomly.
  - How many FRIENDLY and HOSTILE in the entire building?
  - Is the last person who entered the building FRIENDLY or HOSTILE?
  - What color was the uniform of the last person who entered the building?
- On some missions, the robot gives you characteristics. In others, it gives you conclusions.

2

End of Training

You can rewind this video using your mouse.
Or, wait for the video to end.

3

APPENDIX J:
PARTICIPANT INSTRUCTION VIDEO SCRIPT

START SLIDESHOW:

ALL CONDITIONS

NARRATOR

Soldiers and first responders use robots to gather information in dangerous environments. In these missions, a robot may help maintain better awareness. Pay close attention to this mission training video, because it will help you do well in the challenging task you will be performing in a few minutes. But first, let me introduce you to your mission goals.

In this study, you will be working with a robot to identify people within a building. You will complete 12 missions, and each mission is about 5 minutes long. During the missions, you will be prompted to answer questions measuring your awareness. After each mission, you will be asked to rate how well you and the robot did at identifying people in the building.

In each mission, your objective is the same: identification. All the people within the building are either FRIENDLY or HOSTILE. You can determine if someone is friendly or hostile based 4 characteristics. Every person has these four characteristics:

First, a uniform color. Uniforms are red, green, or blue.

Narrator pauses so participant can examine uniform colors.

Second, people are either armed or not armed. When armed, they are holding a rifle.

Narrator pauses so participant can examine the examples.

Third, people will be wearing a helmet or not wearing a helmet.

Narrator pauses so participant can examine the examples.

Finally, people will either be walking or running. People will always run or they will always walk. So even if a runner stops for a few seconds, they are still a runner. This is an example of running. This is an example of walking.

Narrator clicks on each video example to play it and pauses while each plays.

These characteristics are used to determine if someone is friendly or hostile. There are three, and only three, conditions under which a person should be considered HOSTILE.

People are HOSTILE if they have a green uniform and are armed. On the left is an example of a hostile individual because he has green uniform and is armed.

Or, they are armed and wearing a helmet. In the middle is an example of a hostile individual because she is armed and wearing a helmet.

Or, they are wearing a red uniform and are running. The individual on the right has a red uniform and is running, so he is hostile.

If one of those three conditions is not met, then people are FRIENDLY. For example, someone who walks and is unarmed would be friendly, because they don't meet any of the three conditions.

Try these two practice examples. Is this individual FRIENDLY or HOSTILE?

This individual is hostile. They were running, and they wore a red uniform.

Is this individual FRIENDLY or HOSTILE?

Narrator clicks on video and waits for it to play.

This individual is friendly. He doesn't meet any of the three conditions for being hostile. He is armed, but wearing blue. He is armed, but not wearing a helmet. And he was running, but his uniform is blue.

To review: Your mission is to know the friendly or hostile status of every individual in the building. You will determine friendly or hostile status based on the four characteristics. Each mission will pause at random times. When the mission pauses, you will be asked about how many friendly or hostile people are in the building. You will also be asked about the friendly or hostile status of the last person who entered the building, and the uniform color of the last person who entered the building. Always answer with the last person who entered the building at either entrance, no matter if they have left or not.

There is one catch. When answering these questions, you need to answer about the entire building. The buildings in every mission have multiple rooms, but your camera feed only shows one room. There might be people in the other rooms. In order to answer the mission questions, you will need information about people in other rooms. To help you, you will have the aid of a robot.

Your robot is a small vehicle with a camera mounted on its front. It is able to detect people, and it can see through walls. During the missions, you won’t see the robot, but it will provide you with information about people in the building.

This is what your mission screen will look like. On top is your camera feed. Below are two lists. On the left, the robot will give you information about people in the current room.

Narrator clicks to advance to the next annotation.

These are the same people you can see in the camera feed. On the right, the robot will give you information about people in other rooms.

Narrator clicks to advance to the next annotation.

These people you cannot see yourself.

People sometimes move from one room to another. When this happens, they will disappear off of one list and reappear in the other.

Remember, when you are asked questions about the number of friendly and hostile people and the last person who entered, these questions are always about the entire building, not just the room shown in the camera feed. So, you will need information from the robot about the other rooms.

There are a few more things you should know about your robot.

Depending on the mission, your robot may be able to provide one of two kinds of information. Either the robot will provide you with the characteristics of each person- their uniform color, armed status, helmet status, and walking speed. Or, the robot will summarize this information and provide you with a conclusion about whether the person is FRIENDLY or HOSTILE.

At random times in the middle of each mission, you will be asked about two things: One, The number of friendly and hostile people in the building right now. Do not report how many friendly or hostile people have been in the building total. Instead, report how many friendly and hostile people are in the building right now. If two friendly people are in the building and one leaves, you would report 1. Second, whether the last individual who entered the building is FRIENDLY or HOSTILE, and, the color of their uniform. When you report about the last individual in the building, it does not matter if that individual is still present. Always report on the last person who entered, no matter if they have already left or not.

Again, always answer these questions about the state of the building at the current time.

100% RELIABILITY CONDITION

NARRATOR

Your robot is perfectly reliable. That means that it never makes a mistake detecting friendly versus hostile individuals, and it never confuses the characteristics of any person. Furthermore, your robot never misses people who are in the building, and it never sees people in the building who are never there. For these reasons, you should rely upon the robot's information as much as possible.

80% RELIABILITY CONDITION

NARRATOR

Your robot has an 80% reliability rate. That means that for each person who comes into the building, there is a 20% chance that the robot will make a mistake in seeing the person's uniform color. The robot makes these mistakes independently; every time a new person comes into the building, the odds are exactly the same - 80% chance of the robot being correct. It doesn't matter if the robot hasn't made a mistake in a while or just made 2 mistakes in a row. Each new person is 80% likely to be seen correctly. When the robot does make a mistake, it confuses the uniform color. This means that the robot will give you the wrong uniform color and the wrong friendly/hostile label. The only way you can know if the information is wrong is if you can see the individual at the same time. Then you can compare the robot's information to your own information. When you cannot see the person, you should rely on the robot. Note that confusing uniform color and giving the wrong friendly/hostile label is the only kind of mistake the robot can make. Your robot never misses people who are in the building, and it never sees people in the building who are never there.

60% RELIABILITY CONDITION

NARRATOR

Your robot has a 60% reliability rate. That means that for each person who comes into the building, there is a 40% chance that the robot will make a mistake in seeing the person's uniform color. The robot makes these mistakes independently; every time a new person comes into the building, the odds are exactly the same - 60% chance of the robot being correct. It doesn't matter if the robot hasn't made a mistake in a while or just made 2 mistakes in a row. Each new person is 60% likely to be seen correctly. When the robot does make a mistake, it confuses the uniform color. This means that the robot will give you the wrong uniform color and the wrong friendly/hostile label. The only way you can know if the information is wrong is if you can see the individual at the same time. Then you can compare the robot's information to your own information. When you cannot see the person, you should rely on the robot. Note that confusing uniform color and giving the wrong friendly/hostile label is the only kind of mistake the robot can make. Your robot never misses people who are in the building, and it never sees people in the building who are never there.

ALL CONDITIONS

NARRATOR

To recap, you will be doing 12 missions, starting with 2 practice missions. At random times during the mission, the mission will pause, and you will be asked how many friendly and hostile people are in the building. You will also be asked about the last person who entered the building.

Your robot can provide you with information about people in the building. On some missions, it will give you a friendly/hostile labels. On other missions, it will give you the characteristics of individuals and you will have to decide if they are friendly or hostile.

Remember the conditions under which someone is hostile, and refer to the card next to you if you forget.

You can go back in this video by moving your mouse to the bottom of the screen and dragging the slider back. Ask the researcher if you need help.

END SLIDE SHOW:

THE END

LIST OF REFERENCES

Autonomy [Dictionary entry]. (2011). Retrieved from http://www.etymonline.com/index.php?term=autonomy

Barnes, M., Jentsch, F., Chen, J., Haas, E., & Cosenzo, K. (2008). *Five things you should know about soldier - robot teaming*. Paper presented at the meeting of the Army Science Conference, Orlando, FL.

Building unmanned ground vehicles requires more funding, greater focus [Press release]. (2003). Retrieved from http://www8.nationalacademies.org/onpinews/newsitem.aspx?RecordID=10592

Burke, J. L., Murphy, R. R., Coovert, M. D., & Riddle, D. L. (2004). Moonlight in Miami: A field study of human-robot interaction in the context of an urban search and rescue disaster response training exercise. *Human-Computer Interaction*. doi:10.1080/07370024.2004.9667341

Čapek, K. (1921). *R.U.R. (Rossum's Universal Robots)*. Prague: Aventinum.

Chen, J. Y. C., & Barnes, M. J. (2012). Supervisory control of multiple robots: Effects of imperfect automation and individual differences. *Human Factors*, *54*(2), 157-174. doi:10.1177/0018720811435843

Chen, J. Y. C., Barnes, M. J., & Harper-Sciarini, M. (2011). Supervisory control of multiple robots: Human-performance issues and user-interface design. *IEEE Transactions on Systems, Man, and Cybernetics- Part C: Applications and Reviews*, *41*(4), 435-454. doi:10.1109/TSMCC.2010.2056682.

Chen, J. Y. C., Haas, E. C., & Barnes, M. J. (2007). Human performance issues and user interface design for teleoperated robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 37*(6), 1231-1245.

Cosenzo, K., Parasuraman, R., & De Visser, E. (2010). Automation strategies for facilitating human interaction with military unmanned vehicles. In M. Barnes & F. Jentsch (Eds.), *Human-Robot Interactions in Future Military Operations* (pp. 103-124). Surrey, England: Ashgate.

Crocoll, W. M., & Coury, B. G. (1990). Status or recommendation: Selecting the type of information for decision aiding. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society. doi: 10.1177/154193129003401922

Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. *AIAA 1st Intelligent Systems Technical Conference*, *1*, 33–40. Chicago, IL: American Institute of Aeronautics and Astronautics.

Dexter, F., Willemsen-Dunlap, A., & Lee, J. D. (2007). Operating room managerial decision-making on the day of surgery with and without computer recommendations and status displays. *Anesthesia & Analgesia*. doi:10.1213/01.ane.0000268539.85847.c9

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2006). How do automation false alarms and misses affect operator compliance and reliance? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(1), 25-29. Santa Monica, CA: Human Factors and Ergonomics Society.

Durso, F. T., & Sethumadhavan, A. (2008). Situation awareness: Understanding dynamic environments. *Human Factors, 50*, 442-448. doi:10.1518/001872008X288448

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. doi:10.1016/S1071-5819(03)00038-7

Ehrlich, K., Kirk, S., Patterson, J., Rasmussen, J., Ross, S., & Gruen, D. (2011, February). *Taking advice from intelligent systems: The double- edged sword of explanations*. Paper presented at the meeting of the International Conference on Intelligent User Interfaces, Palo Alto, CA.

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). *IEEE 1988 National Aerospace and Electronics Conference*, *3*, 789-795. doi: 10.1109/NAECON.1988.195097

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors: The Journal of the Human Factors and Ergonomics Society, 37(1), 32-64. doi:10.1518/001872095779049543

Endsley, M. R. (2000a). Direct measurement of situation awareness: Validity and use of SAGAT. In M. R. Endsley & D. J. Garland (Eds.), *Situation Awareness Analysis and Measurement* (pp. 147-173). Mahwah, NJ: Lawrence Erlbaum Associates.

Endsley, M. R. (2000b). Theoretical underpinnings of situational awareness: A critical review. In M. R. Endsley & D. J. Garland (Eds.), *Situation Awareness Analysis and Measurement* (pp. 3-32). Mahwah, NJ: Lawrence Erlbaum Associates.

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*(3), 462-492. doi:10.1080/001401399185595

Endsley, M., Sollenberger, R. L., Nakata, A., & Stein, E. S. (2000). Situation awareness in air traffic control: *Enhanced displays for advanced operations* (Report No. DOT/FAA/CT-TN00/01). Atlantic City International Airport, NJ: Federal Aviation Administration.

Eski, I., Erkaya, S., Savas, S., & Yildirim, S. (2010). Fault detection on robot manipulators using artificial neural networks. *Robotics and Computer-Integrated Manufacturing*. doi:10.1016/j.rcim.2010.06.017

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191.

Franklin, S., & Graesser, A. (1997). Is it an agent or just a program? A taxonomy for autonomous agents. In J. Müller, M. Wooldridge, & N. Jennings (Series Eds.), *Intelligent Agents: Agent Theories, Architectures, and Languages* (Vol. 1193, pp. 21-35). Heidelberg: Springer Berlin. doi:10.1007/BFb0013570

Galster, S. M., Bolia, R. S., & Parasuraman, R. (2002, October). The application of a qualitative model of human-interaction with automation: Effects of unreliable automation on performance. In W. Tielemans (Chair), *He Role of Humans in Intelligent and Automated Systems*. Symposium conducted at the meeting of the North Atlantic Treaty Organization, Warsaw, Poland.

Goodrich, M. A., McLain, T. W., Anderson, J. D., Sun, J., & Crandall, J. W. (2007). *Managing autonomy in robot teams: Observations from four experiments*. Paper presented at the meeting of the ACM International Conference on Human-Robot Interaction, Arlington, Virginia.

Harwood, K., Barnett, B., & Wickens, C. (1988). *Situational awareness: A conceptual and methodological framework*. Paper presented at the meeting of the 11th Symposium of Psychology in the Department of Defense, Colorado Springs, Colorado.

Horrey, W. J., & Wickens, C. D. (2001). *Supporting battlefield situation assessment through attention guidance and diagnostic aiding: A cost-benefit and depth of processing analysis* (Report No. ARL-01-16/FED-LAB-01-1). Savoy, IL: Aviation Research Lab.

Horrey, W. J., Wickens, C. D., Strauss, R., Kirlik, A., & Stewart, T. R. (2009). Supporting situation asessment through attention guidance and diagnostic aiding: The benefits and costs of display enhancement on judgment skill. In A. Kirlik (Ed.), *Adaptive Perspectives on Human-Technology Interaction: Methods and Models for Cognitive Engineering and Human-computer Interaction* (pp. 55-70). Oxford, England: Oxford University Press.

International Organization for Standardization (1994). *ISO standard 8373:1994*. Geneva, Switzerland: Author.

Jentsch, F., Fincannon, T., Sellers, B., Keebler, J., Ososky, S., Phillips, E., & Schuster, D. (2012). *Safe operation of autonomous robot teams: impact of autonomy, teaming, and workload on operator trust and performance* (Task Order 132). Orlando, FL: University of Central Florida.

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C., Sierhuis, M., & Van Riemsdijk, B. (2010). Toward coactivity. *HRI '10: Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, *5*, 101-102. New York, NY: Association for Computing Machinery. doi:10.1145/1734454.1734492

Johnson, R. C., Saboe, K. N., Prewett, M. S., Coovert, M. D., & Elliott, L. R. (2009). Autonomy and automation reliability in human-robot interaction: A qualitative review. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *53*, 1398-1402. Santa Monica, CA: Human Factors and Ergonomics Society. doi:10.1177/154193120905301849

Jones, D. G. (2000). Subjective measures of situation awareness. In M. R. Endsley & D. J. Garland (Eds.), *Situation Awareness Analysis and Measurement* (pp. 113-128). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kaber, D. B., & Endsley, M. R. (1997). Out-of-the-loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety. *Process Safety Progress, 16*(3), 126-131.

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*(2), 113-153. doi:10.1080/1463922021000054335

Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition*, *26*(2), 169-181.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243-1270.

Luck, M., & D'Inverno, M. (1995). A formal framework for agency and autonomy. *Proceedings of the First International Conference on Multiagent Systems*, *1*, 254-260. San Francisco, CA: AAAI Press/MIT Press.

Madhavan, P., & Phillips, R. R. (2010). Effects of computer self-efficacy and system reliability on user interaction with decision support systems. *Computers in Human Behavior*, *26*(2), 199–204. doi:10.1016/j.chb.2009.10.005

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277-301. doi:10.1080/14639220500337708

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57-87. doi:10.1177/1555343411433844

May, P. A. (1993). *Effects of automation reliability and failure rate on monitoring performance in a multi-task environment* (Unpublished master's thesis). Catholic University of America, Washington, DC.

Oakley, B., Mouloua, M., & Hancock, P. (2003). Effects of automation reliability on human monitoring performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 188-190. Santa Monica, CA: Human Factors and Ergonomics Society.

Oleson, K. E., Billings, D. R., Kocsis, V., Chen, J. Y. C., & Hancock, P. A. (2011). Antecedents of trust in human-robot collaborations. *IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*. Miami, FL: IEEE CogSIMA. Retrieved from http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5744986

Parasuraman, R., Mouloua, M., & Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. *Human Factors*, *38*(4), 665-679.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230-253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, *30*(3), 286-297. doi:10.1109/3468.844354

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(3), 511-520. doi:10.1518/001872008X312198

Pew, R. W. (1994). Situation awareness: The buzzword of the 90's. *CSERIAC Gateway, 5*(1), 1-4.

Rehfeld, S. A. (2006). *The impact of mental transformation training across levels of automation on spatial awareness in human-robot interaction* (Unpublished doctoral dissertation). University of Central Florida, Orlando, Florida.

Riley, J. M., Strater, L. D., Chappell, S. L., Connors, E. S., & Endsley, M. R. (2010). Situation awareness in human-robot interaction: Challenges and user interface requirements. In M.

Barnes & F. Jentsch (Eds.), *Human-Robot Interactions in Future Military Operations* (pp. 171-191). Surrey, England: Ashgate.

Rousseau, R., Tremblay, S., & Breton, R. (2004). Defining and modeling situation awareness: A critical review. In S. Banbury & S. Tremblay (Eds.), *A Cognitive Approach to Situation Awareness: Theory and Application* (pp. 3-21). Burlington, VT: Ashgate Publishing Company.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*(1), 76-87. Santa Monica, CA: Human Factors and Ergonomics Society. doi:10.1518/001872007779598082

Rudisill, M. (2000). *Crew/automation interaction in space transportation systems: Lessons learned from the glass cockpit*. Hampton, VA: NASA Langley Research Center.

Ruff, H. A., Narayanan, S., & Draper, M. H. (2002). Human interaction with levels of automation and decision-aid fidelity in the supervisory control of multiple simulated unmanned air vehicles. *Presence, 11*(4), 335-351. doi:10.1162/105474602760204264

Salas, E., Cannon-Bowers, J., Fiore, S. M., & Stout, R. J. (2001). Cue-recognition training to enhance team situation awareness. In M. McNeese, E. Salas, & M. Endsley (Eds.), *New Trends in Collaborative Activities: Understanding System Dynamics in Complex Environments* (pp. 169-190). Santa Monica, CA: Human Factors and Ergonomics Society.

Sarter, N. B., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors*, *43*(4), 573-583.

Schuster, D., Keebler, J., Zuniga, J., & Jentsch, F. (2012, New Orleans, LA). *Individual differences in SA measurement and performance in human-robot teaming*. Poster session presented at the meeting of the IEEE Conference on Cognitive Methods in Situation Awareness and Decision Support, New Orleans, LA.

Sheridan, T. B. (2000). Function allocation: Algorithm, alchemy or apostasy? *International Journal of Human-Computer Studies*, *52*(2), 203-216. doi:10.1006/ijhc.1999.0285

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Report No. N00014-77-C-0256). Cambridge, MA: United States Office of Naval Research.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*, 991-1006.

Stancliff, S., Dolan, J. M., & Trebi-Ollennu, A. (2005). *Towards a predictive model of robot reliability*. Pittsburgh, Pennsylvania: Carnegie Mellon University.

Tenney, Y. J., & Pew, R. W. (2006). Situation awareness catches on: What? So what? Now what? *Reviews of Human Factors and Ergonomics*, *2*(1), 1-34.

United States Army (2009). *Tactics in counterinsurgency* (Field Manual No. 3-24.2). Washington, DC: Author.

United States Army (2008). *The warrior ethos and soldier combat skills* (Field Manual No. 3-21.75). Washington, DC: Author.

United States Army (2006). *Cordon and search: Multi-service tactics, techniques, and procedures for cordon and search operations* (Field Manual No. 3-06.20). Washington, DC: Author.

United States Army Research Laboratory (2011). *Robotics collaborative technology FY 2011 annual program plan*. Adelphi, MD: Author. Retrieved from http://www.arl.army.mil/www/pages/392/rcta.fy11.ann.prog.plan.pdf

United States Department of Defense (1998). *Department of defense modeling and simulation glossary* (Directive No. 5000.59-M). Washington, DC: Author.

Vidulich, M., Dominguez, C., Vogel, E., & McMillan, G. (1994). *Can SA be defined?* (Report No. AL/CF-TR-1994-0085). Wright-Patterson Air Force Base, OH: Air Force Systems Command.

Wickens, C. D. (2002). *Spatial awareness biases* (Report No. AHFD-02-6/NASA-02-4). Moffett Field, CA: NASA Ames Research Center.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201-212.

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010). Stages and levels of automation: An integrated meta-analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(4), 389-393. Santa Monica, CA: Human Factors and Ergonomics Society.

Wickens, C. D., Li, H., Sebok, A., & Sarter, N. B. (2010). Stages and levels of automation: An integrated meta-analysis. *Proceedings of the Human Factors and Ergonomics Society*

Annual Meeting, 54, 389-393. Santa Monica, CA: Human Factors and Ergonomics Society.

Yeh, M., Wickens, C. D., & Seagull, J. F. (1999). Target cuing in visual search: The effects of conformality and display location on the allocation of visual attention. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *41*(4), 524-542.

Your view: How would you define a robot? [Blog post]. (2007, July 16). Retrieved from http://www.cbc.ca/technology/technology-blog/2007/07/your_view_how_would_you_define.html