ARL-TR-8956 • MAY 2020

Agent Transparency for an Autonomous Squad Member: Depth of Reasoning and Reliability

by Julia L Wright, Jessie YC Chen, Shan G Lakhmani, and Anthony R Selkowitz

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-8956 ● MAY 2020

Agent Transparency for an Autonomous Squad Member: Depth of Reasoning and Reliability

Julia L Wright, Jessie YC Chen, Shan G Lakhmani
Human Research and Engineering Directorate, CCDC Army Research Laboratory

Anthony R Selkowitz
Oak Ridge Associated Universities

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
May 2020	Technical Report	September 2016–June 2017

4. TITLE AND SUBTITLE
Agent Transparency for an Autonomous Squad Member: Depth of Reasoning and Reliability

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Julia L Wright, Jessie YC Chen, Shan Lakhmani, and Anthony Selkowitz

5d. PROJECT NUMBER
ARL-16-084

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
CCDC Army Research Laboratory
ATTN: FCDD-RLH-FD
Aberdeen Proving Ground, MD 21005

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-8956

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

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

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

13. SUPPLEMENTARY NOTES
ORCID ID(s): Julia L Wright, 0000-0003-3026-1538; Shan Lakhmani, 0000-0001-6052-439X

14. ABSTRACT
Agent transparency is an important contributor to effective human–agent teaming. However, agent transparency's effects on human performance when the agent is unreliable have yet to be examined. This report examines how the transparency and reliability of a robotic autonomous squad member (ASM) affected a human observer's task performance, workload, situation awareness (SA), trust in the robot, and perceptions of the robot. In a 2 (ASM transparency) × 2 (ASM reliability) within-subject design experiment, participants monitored a simulated soldier squad that included an ASM as it traversed a simulated training environment, while concurrently monitoring the environment for targets. There was no difference in participants' performance on the target detection task, workload, or SA due to either ASM transparency or reliability. ASM reliability influenced participant trust and perceptions of the robot. Results suggest that reliability may be a stronger influence on the human's perceptions of the robot than transparency. Robot errors had a profound and lasting effect on the participants' perception of the robot's future reliability and resulted in reduced confidence in their assessments of the robot's reliability. These findings could have important implications for the continued use of automated systems when the user is aware of system errors.

15. SUBJECT TERMS
agent transparency, human–agent teaming, display design, visualization, situation awareness

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Julia L Wright
Unclassified	Unclassified	Unclassified	UU	110	19b. TELEPHONE NUMBER (Include area code) 407-208-3348

Contents

List of Figures

List of Tables

1. Introduction

1.1 The Autonomous Squad Member Project

Autonomous robotic agents for military operations are becoming increasingly sophisticated and independent. As robotic agent autonomy increases, it becomes paramount for human team members to understand the agent's behavior, reasoning, and outcome projections. The autonomous squad member (ASM) project is a multi-experiment endeavor, funded as part of the Department of Defense's Autonomy Research Pilot Initiative (ARPI) program, which explores human interaction issues between a human team member and a robotic team member within a simulated environment (Military.com 2013). The focus of the ARPI ASM project is developing a transparent interface for ASM communications with the Soldier teammates, and then examining how this transparency impacts the human's awareness of the ASM's actions, reasoning, perceptions, and projected outcomes, as well as the human's perceptions of the ASM itself. Prior ASM studies explored how agent transparency can be facilitated via the ASM's display design (Boyce et al. 2015) and level of information (Selkowitz et al. 2016).

This experiment investigated the effects of depth of information and ASM reliability on a human teammate's awareness and perceptions of the ASM. Participants monitored the progress of a simulated dismounted Soldier team as it progressed through a series of training courses, each time accompanied by a different ASM. Each ASM varied in the depth of information it provided, as well as its reliability. During each training mission, the participant's situation awareness (SA) was assessed, and after each course was completed, their perceived workload, trust, and perceptions of the ASM were evaluated.

Surface-level information would be the ASM's goals, motivation, projected outcome, and perception of the squad's actions. In a prior study, it was found that when the agent shared not only its current goal but also its motivation and projected outcome, the human teammate had improved SA and trust in the agent (Selkowitz et al. 2016). In-depth information would be the factors and/or reasoning behind these goals, motivators, and projections. For instance, if the surface-level information is that the robot has chosen a plan because it is trying to preserve its mechanical integrity, the in-depth information behind this motivation may be that it has observed explosions or gunfire in the area. The level and depth of information were developed using the principles of the SA-based Agent Transparency (SAT) model (Chen et al. 2014).

1.2 The SA-based Agent Transparency (SAT) Model

The SAT model is a foundation for displaying transparency information to facilitate effective interactions between a human and an autonomous agent by supporting an operator's SA (Chen et al. 2014). Agent transparency has been defined as "the descriptive quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process," as such, the SAT model suggests that increasing agent transparency will assist the operator in maintaining SA of the mission environment (Chen et al. 2014). Making the underlying processes that the autonomous agent uses to make its projections, decisions, and actions available to the operator should effectively create a shared understanding of the autonomous agent's environment and decision-making process between the operator and an autonomous agent. This shared understanding has been cited as an essential factor for the operator's trust calibration regarding an autonomous agent (Lee and See 2004; Lyons 2013; Lyons and Havig 2014). Thus, the SAT model can be used to inform the information architecture design so that it facilitates the operator's trust calibration regarding the autonomous agent and allows the operator to maintain SA of the agent (Chen et al. 2014).

As seen in Fig. 1, the SAT model is composed of three levels of information (Chen et al. 2014). Level 1 is basic information about the autonomous agent's actions and plans. It also describes the agent's knowledge of the environment and events within it. For example, Level 1 information could be any of the following: the agent's actions, enemy combatants, buildings, or its human teammates. This level is meant to foster a shared understanding between the participant's perception of the world and the autonomous agent's perception of the world. Level 2 is the agent's reasoning behind its actions/decisions. An example of Level 2 information could be that the agent indicates why it chose one route over another or logic used to make decisions. Information from this level assists the participant's understanding of why the autonomous agent has taken a particular action. Level 3 includes the agent's predictions about its future actions and state of uncertainty. Information supporting Level 3 understanding is meant to aid the participant by communicating the agent's projected consequences of its actions. An example of Level 3 information would be the agent indicating its projections and predictions.

Fig. 1 SAT model (Chen et al. 2014)

In the current study, the effect of the ASM sharing in-depth information for the three surface levels of information on the operator's SA, trust, and perceived workload was examined. In-depth information would be the factors and/or reasoning behind these actions, reasoning, and projections. For instance, if the surface-level information is that the robot has chosen a plan because it is trying to preserve its mechanical integrity, the in-depth information behind this reasoning may be that it observed explosions or gunfire in the area. In this manner, SAT information can be thought of as a way in which the robot can share its SA with the operator to support the operator's SA.

1.3 Situation Awareness

Developing appropriate SA is a mission-critical goal for human–robot teams (Evans 2012). SA refers to an individual's dynamic understanding of "what is going on" in a given system (Endsley 1995). Several conceptions of SA exist; the most popular is Endsley's (1995) information-processing-based model, which suggests that SA is comprised of three hierarchical levels: perception of elements within the environment, the comprehension of their meaning, and a projection of their status in the near future. As the SAT model was designed to support the operator's SA, the operator's SA at each level was assessed to evaluate the effectiveness of the SAT information.

1.4 Reliability

Automation reliability rates can influence a user's task performance and trust in the automation. Wickens and Dixon (2007) found that around 70% automation reliability was the point below which performance using such automation was worse than not using any automation at all. Parasuraman et al. (1993) found that human–automation groups with variable automation reliability (87.5% and 56.25%

alternating every 10-min block) were better at detecting automation errors than constant reliability groups (87.5% group or 56.25% group). On the other hand, having a higher, albeit still imperfect, reliability rate may increase operator dependency and lead to complacency (McBride et al. 2014).

In the current study, ASM reliability was manipulated to examine the interaction of the depth of information and reliability. Errors were operationalized as a situation in which the autonomous robotic agent performed a normal action that was inappropriate for the given context—otherwise known as an error of commission (Lewis 1998). Previous research showed that participants who were aware of robot errors had lower trust and reliance in the robot compared to those who had a robot who made no errors (Salem et al. 2015), although this reduced trust did not prevent the participants from complying with robot commands when the consequences were not irrevocable. This indicates that the relationship between robot reliability and operator trust may be dependent upon the task and subsequent consequences of task outcomes.

1.5 Trust

Operator trust has been defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See 2004). A robot needs to provide meaningful insight into its actions and why it is performing them to provide proper support for the operator's trust calibration (Chen et al. 2014). Appropriate trust is essential to effective performance; too much trust can result in operator complacency or misuse, while too little trust can result in disuse (Parasuraman and Riley 1997).

Operators' attitude toward automation influences their level of trust in the automation (Chen et al. 2014). Operators' explicit attitudes, which are conscious and cognitively effortful, can be measured using self-report measures (Merritt et al. 2013). On the other hand, implicit attitudes toward automation, unconscious "gut reactions", can still influence operators' perception of information, and subsequent behavior. Unlike their explicit counterparts, implicit attitudes are determined through measures of speed and evaluative representation (Merritt et al. 2013).

A primary research question for the current study was how the participant's trust in the agent was affected by the errors that the agent committed, and if the display of information to support transparency information aided in mitigating the impact of the errors. To that end, the participant's implicit attitude toward automation was evaluated, and then their explicit attitudes toward each simulated robot were assessed.

1.6 Workload

The addition of SAT information and uncertainty information may increase visual complexity since each increase in information adds more visual elements to be displayed. Previous studies have shown that as visual complexity increases, mental workload also increases (Kroft and Wickens 2002; Lohrenz et al. 2009). Parasuraman et al. (2008) defined mental workload as "the relation between the function relating the mental resources demanded by a task and those resources available to be supplied by the human operator." The investigation of workload in the current study was a priority to assess if the increase in SAT information and uncertainty in the display resulted in a concomitant increase in workload. Workload was assessed, both subjectively and objectively, through the use of eye-movement analysis.

Eye movements are an effective way of assessing underlying cognitive activities (Beatty 1982; Jacob and Karn 2003). Blink duration and pupil diameter have been shown to correlate positively with cognitive workload (Ahlstrom and Friedman-Berg 2006; Peavler 1974). Fixation count correlated negatively with increased mental workload, while (correspondingly) longer fixation durations correlate positively with increased mental workload (Goldberg and Kotval 1999; Van Orden et al. 2001). In the current study, eye-tracking metrics were evaluated as objective measures of mental workload as they are sensitive to differences in workload that subjective measures may not reveal (Ahlstrom and Friedman-Berg 2006; Wright et al. 2014).

1.7 Humanness

Since trust is one of the primary attributes being studied, the humanness and intelligence of the robot is of interest as well (Lee and See 2004). Hinds et al. (2004), while investigating the human-like attributes of a robot in a human–robot teaming task, found that users felt less responsible for the task when they collaborated with a humanlike robot than with a machine-like robot. This finding suggests that when the robot is more humanlike, the operator is more willing to cede responsibility for the outcome of the task to their robot teammate. In a previous study investigating the effectiveness of SAT-based information in the ASM display, it was shown that when the robot displayed information regarding its uncertainty and projections, the operator rated the robot as being more humanlike, compared to conditions that displayed only the ASM's reasoning and current understanding of its environment (Lakhmani et al. Forthcoming 2020). In addition to more humanlike, the ASM was also rated as more trustworthy, even though the ASM's reliability remained unchanged. This finding illustrates the need to

investigate further how attributions of humanness and intelligence in robots would be affected by manipulating the robots' error rate and level of information to support transparency.

1.8 Individual Differences

Additionally, the effects of several individual difference (ID) variables that could affect an operator's performance in a multitasking environment were investigated. Previous studies showed that perceived attentional control, spatial ability, and video game experience contribute to performance in simulated environments and robot supervisory tasks (Chen and Barnes 2012). Individuals who reported high perceived attentional control (PAC) outperformed their low attentional control counterparts on tasks that require attention focus and shifting of attention (Chen and Barnes 2012; Wright et al. 2013). Differential effects on performance due to spatial ability (SPA) have been found on teleoperation tasks, robotic operation, and target detection tasks (Chen et al. 2010; Chen et al. 2008; Lathan and Tracey 2002), as well as improved SA and target-detection performance (Wright et al. 2013). Working memory capacity (WMC) differences affect performance in multi-robot supervisory tasks and SA (Ahmed et al. 2014; Endsley 1995). It has been demonstrated that frequent video gamers can maintain better situation awareness in dynamic tasking environments than infrequent video gamers (Chen and Barnes 2012). In the current experiment, we examined the differential effects of PAC, SPA, WMC, and gaming experience on operator SA and perceived workload.

1.9 Current Study

1.9.1 Overview

The present research investigated how the transparency of agent reasoning and agent reliability, within the context of human–agent teaming, influences operator behavior and attributions regarding the agent in a dynamic simulated environment. Transparency of agent reasoning was manipulated by varying the depth of SAT model information displayed; surface-level SAT model information (S) only versus surface-level plus the in-depth factors for the SAT model information (D). Agent reliability was manipulated by varying the ASM's error rate (67% versus 100% reliable) for its responses to events occurring in the environment. In a 2 × 2, counterbalanced, mixed-factor design experiment, each participant completed four trials: two at each information level and two at each error rate. Within-subjects evaluations compared differences in operator behavior and attributions regarding the agent across information level and agent error rate. Between-subjects evaluations assessed how IDs affected these outcomes.

During the experiment, participants monitored a simulated dismounted Soldier team, accompanied by an ASM, as it traversed a training course four times. Each time the team completed the course, it was accompanied by a different ASM. As the team traversed the training course, events would occur, and the team (and ASM) would have to respond. Participants were required to identify the event, as well as other potential threats, as soon as they became aware of it using on-screen buttons. Six events occurred during each training course; in the 100% reliable (R100) condition the ASM always responded correctly to the event, and in the 67% reliable condition (R67) the ASM did not respond correctly to the event two out of six times. The Soldier team always responded correctly to the events. The ASM display would show the robot's perceptions of the team's actions, its current objective, the reason behind this objective, and projected resource loss associated with this objective. In the in-depth information condition, the ASM display would also show the underlying factors behind the current objective. Participants were required to monitor the environment and the ASM interface and received SA queries throughout each trial.

The findings of this study are expected to inform how the amount of agent transparency available to the operator interacts with the agent's reliability to influence operator perceptions and attributions of the robot. While a high error rate may reduce an operator's trust in a robot, increased transparency of agent reasoning may mitigate these effects. However, increased transparency may also increase operator workload, as they have more information to process. Additionally, this work investigated how several ID factors influenced the human–agent relationship in terms of agent transparency, and their effect on the related human performance issues.

1.9.2 Hypotheses

1.9.2.1 Target Detection Task

H1: Access to in-depth agent reasoning will not affect participant's target detection performance when the agent is perfectly reliable (R100), SR ≈ DR, but will be detrimental to the participant's target detection performance when the agent is less reliable (R67), SU > DU.

H2: Reduced agent reliability will reduce the participant's target detection performance, R67 < R100, regardless of the information level.

1.9.2.2 Perceived Workload

H3: Access to in-depth agent reasoning will not increase operator mental workload when the agent is perfectly reliable (R100), SR ≈ DR, but will increase operator mental workload when the agent is less reliable (R67), SU < DU.

H4: Reduced agent reliability will increase operator mental workload, R67 > R100, regardless of information level.

1.9.2.3 Situation Awareness

H5: Access to in-depth agent reasoning will increase operator SA of the agent when the agent is perfectly reliable (R100), SR < DR, but will not increase operator SA of the agent when the agent is less reliable (R67), SU ≈ DU.

H6: Reduced agent reliability will decrease operator SA of the agent, R67 < R100, regardless of the information level.

1.9.2.4 Operator Trust

H7: Access to in-depth agent reasoning will increase operator trust in the agent when the agent is perfectly reliable (R100), SR < DR, but will not increase operator trust when the agent is less reliable (R67), SU ≈ DU.

H8: Reduced agent reliability will decrease operator trust in the agent, R67 < R100, regardless of information level.

1.9.2.5 Robot Humanness and Intelligence

H9: Access to in-depth agent reasoning will increase the robot's perceived humanness and intelligence when the agent is perfectly reliable (R100), SR < DR, but will not increase the robot's perceived humanness and intelligence when the agent is less reliable (R67), SU ≈ DU.

H10: Reduced agent reliability will reduce the robot's perceived humanness and intelligence, R67 < R100, regardless of information level.

1.9.2.6 ID Factors

The effects of IDs in PAC, SPA, and WMC on the operator's task performance, trust, workload, and SA were also investigated. Differential effects are reported.

2. Method

2.1 Participants

A total of 81 participants (between the ages of 18 and 40) were recruited from the University of Central Florida's (UCF) Institute for Simulation and Training's Sona System. UCF's Sona System is a participant recruitment system that allows students and members of the local community to participate in research. Participants received cash payment ($15/h) as compensation. There were 25 potential participants excused or dismissed from the study, or the data from those sessions were deemed useless and had to be replaced: nine were given incorrect condition sequences, which rendered that data useless; three had equipment malfunctions; two participants' data were lost because a participant moved part of the equipment setup and the world model had to be redefined; six failed the Ishihara Color Test (Ishihara 1917); one participant could not pass the training requirement; and four fell asleep or were not attending to their experimental duties and were dismissed. Those who were determined to be ineligible or withdrew from the experiment received payment for the amount of time they participated, with a minimum of one hour's pay. There were 56 participants (32 males, 23 females, 1 unreported, Min_{age} = 18 years, Max_{age} = 31 years, M_{age} = 20.5 years) who successfully completed the experiment and their data were used in the analysis.

2.2 Apparatus

2.2.1 Simulator

A custom software application, capable of showing images and video, was used to present the ASM display to the participant (Fig. 2). The simulation was coded in the HAVOK simulation engine. The simulation was delivered via a commercial desktop computer system, two 22-inch monitors, standard keyboard, and three-button mouse.

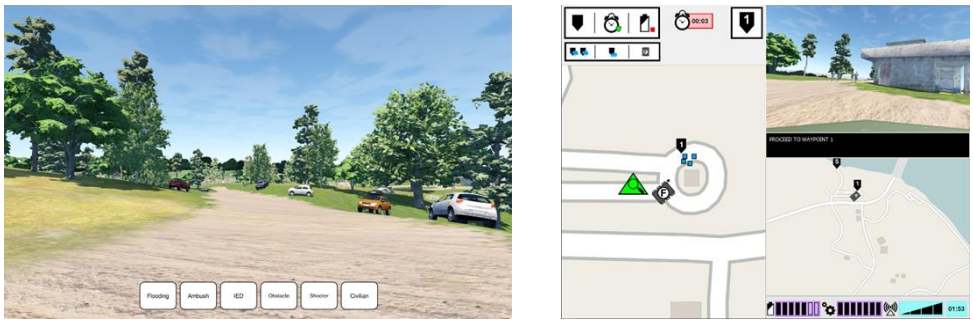


Fig. 2 ASM experimental interface. The left-side monitor displays the lead Soldier's point of view of the task environment. The right-side monitor displays the ASM's communication interface.

2.2.2 Eye Tracker

A desk-mounted Smart Eye Pro (Smart Eye AB, Gottenburg, Sweden) eye-tracking system was used to collect eye-movement data. The eye-tracker system is comprised of a pair of cameras, each with an infrared flasher, mounted under each computer monitor (Fig. 3). The Smart Eye system uses an IR camera-based tracking system, which allows noncontact operation. Eye and head movements, which can be observed at approximately 0.5° of spatial resolution and sampled at the rate of 60 Hz, along with measurement reliability data, were logged in real time and synchronized with performance data from other systems. Only the participants' eye-gaze coordinates were measured and recorded; no video of the participants' eyes and faces was recorded. The system was individually calibrated for each participant after the training exercise.

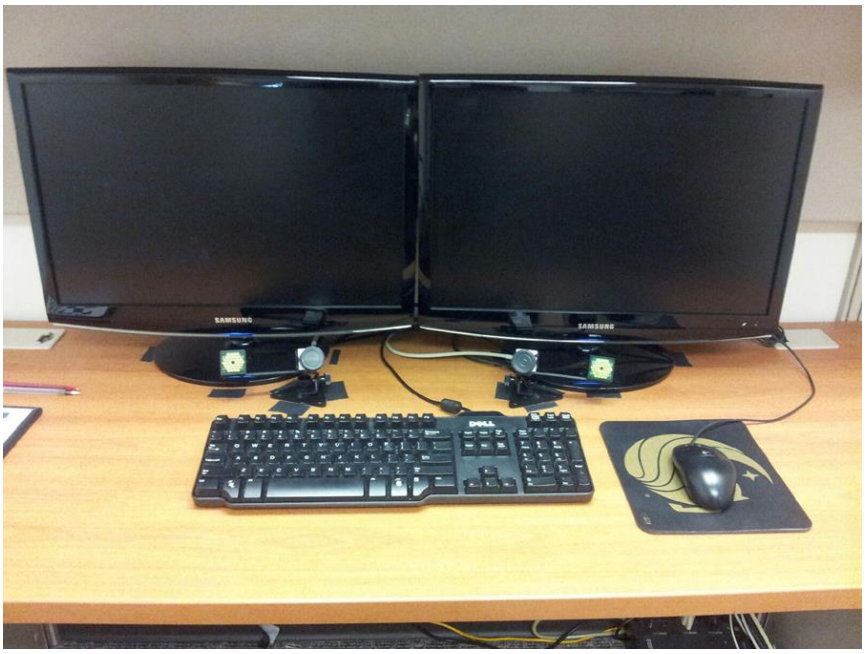


Fig. 3 Experiment station showing the eye-tracking cameras positioned in front of the monitors

2.3 Surveys and Tests

2.3.1 Demographics Questionnaire

A demographics questionnaire was administered at the beginning of the experimental session (Appendix A). Information on the participant's age, gender, education level, computer familiarity, and gaming experience (GE) was collected. Participants who played action video games at least weekly were classified as gamers (Gamers $N = 33$, NonGamer $N = 23$).

2.3.2 Ishihara Color Vision Test

An Ishihara Color Vision Test (Ishihara 1917) using nine test plates (Appendix B) was administered via PowerPoint slide presentation. Since the ASM interface employs several colors to display the plans for the ASM, normal color vision is required to interact with the system effectively. Six potential participants failed to identify at least seven of the plates correctly, so were paid for one hour ($15) and dismissed.

2.3.3 Implicit Association Test (IAT)

Implicit trust will be measured using the modified Implicit Association Test (IAT) developed by Merritt et al. (2013). During an IAT, participants are asked to categorize "good" and "bad" words into superordinate categories, such as

"automation" and "human" in this instance. Participants complete several trials, and scores are calculated by dividing the difference in response times by the pooled standard deviations (Appendix C). Faster response times infer stronger associations. The scoring procedure outlined in Greenwald et al. (2003) was used, which produces a statistic similar to Cohen's *d* and indicates the implicit preference for automation over humans. Raw scores were converted to Z-scores, then reversed. As such, higher scores indicate higher implicit trust in automation; (Min_{IAT} = –1.69, Max_{IAT} = 3.04, Mdn_{IAT} = –0.03, M_{IAT} = 0.00, SD_{IAT} = 1.00, IAT_{LOW} N = 28, IAT_{HIGH} N = 28).

2.3.4 Spatial Ability Tests

Two aspects of spatial ability were assessed: visual scanning (VS) and spatial orientation (SO).

The Matrix Scanning Task (VS; Appendix D), developed by Barrett et al. (1982), assesses an individual's ability to scan a static display to extract information. It is a timed test, delivered via computer, wherein an individual is presented with a screen displaying a 4 × 5 matrix of triangles for 3 s. Each triangle has a line through it; however, in each matrix one, two, three, or four of the triangles do not have lines. The individual must correctly identify how many triangles did not have lines in each trial. Participants score 1 point for each correct trial; total score is calculated as the number correct across 20 trials (Min_{VS} = 16, Max_{VS} = 20, Mdn_{VS} = 19.0, M_{VS} = 19.0, SD_{VS} = 1.2). There was a ceiling effect of the scores for this measure, as 42 of 56 participants scored 19 or 20 points; as such, it was not used in any analyses.

The Spatial Orientation Test (SO; Appendix E) measures an individual's ability to orient themselves in a 3-D world (Gugerty and Brooks 2004). It is a computerized test consisting of a brief training segment and 32 test questions. The score is based on both accuracy and response time (RT). Scores are calculated by dividing average RT by total number correct, and higher performance is indicated by lower scores. Lower scores indicate better performance. High/low group membership was determined by median split of all participants' scores (Min_{SO} = 2.1, Max_{SO} = 30.6, Mdn_{SO} = 9.2, M_{SO} = 11.1, SD_{SO} = 7.3, SO_{LOW} N = 28, SO_{HIGH} N = 28).

2.3.5 National Aeronautics and Space Administration–Task Load Index (NASA–TLX)

Participants' perceived workload was evaluated with the computerized version of the NASA–TLX questionnaire (Appendix F), which uses a pairwise comparison weighting procedure (Hart and Staveland 1988). The NASA-TLX is a self-reported questionnaire of perceived demands in six areas: mental demand, physical demand, temporal demand, effort (mental and physical), frustration, and performance.

Participants evaluated their perceived workload in these areas on 10-point scales as well as completing pairwise comparisons for each subscale after completing each experimental trial.

2.3.6 Godspeed Measure

The Godspeed measures (Appendix G) assess an individual's perceptions of a robot on such attributes as anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety via a series of bipolar Likert scale evaluations (Bartneck et al. 2009). Participants completed a Godspeed measure after each experimental trial.

2.3.7 Functional Trust Survey

The Functional Trust Survey (Appendix H) was developed to further distinguish the basis of an individual's trust in an autonomous agent. It comprises the Trust in Automation Survey (i.e., questions 1–12; Jian et al. 2000) and select questions from the Trust Perception Scale - HRI (i.e., questions 13–16; Schaefer 2016) that have been modified to assess trust along the four functions of automation use as identified by Parasuraman et al. (2000) (i.e., A: gathering and filtering information, B: integrating and displaying information, C: suggesting or making decisions, and D: executing actions), along with related usability questions (i.e., questions 17–23) from the Technology Acceptance Measure (TAM; Davis 1989). Participants completed a Functional Trust Survey after each experimental trial.

2.3.8 RSPAN

Verbal WMC was assessed using the automated reading span task (RSPAN; Appendix I), which has high internal (partial score $\alpha = 0.86$) and test-retest ($\alpha = 0.82$) reliability (Redick et al. 2012; Unsworth et al. 2005). WMC was evaluated by using the participants' letter set score (total number of letters in perfectly recalled letter sets), and higher values indicate greater WMC, ($Min_{RSPAN} = 12$, $Max_{RSPAN} = 54$, $Mdn_{RSPAN} = 35.50$, $M_{RSPAN} = 35.09$, $SD_{RSPAN} = 11.16$). High/low group membership was determined by the median split of all participants' scores, $RSPAN_{LOW}$ $N = 28$, $RSPAN_{HIGH}$ $N = 28$.

2.3.9 Attentional Control Survey

A questionnaire on attentional control (Appendix J) was used to measure participants' perceived attentional control by evaluating their perception of their attention focus and shifting (Derryberry and Reed 2002). The Attentional Control Survey consists of 20 items scored on a 1–4 point Likert scale, with half of the items reverse-scored. The score range is 20–80 points, with higher scores indicating

better attentional control. The scale has been shown to have good internal reliability ($\alpha = 0.88$). High/Low group membership was determined by median split of all participants' scores ($Min_{PAC} = 39$, $Max_{PAC} = 79$, $Mdn_{PAC} = 56.00$, $M_{PAC} = 55.50$, $SD_{PAC} = 8.44$; PAC_{LOW} $N = 29$, PAC_{HIGH} $N = 27$).

2.4 Experiment Design and Performance Measures

2.4.1 Experiment Design

The study was a 2 × 2 mixed-factor design experiment. Within-subjects evaluations compared differences in operator behavior and attributions regarding the agent across agent transparency and agent reliability levels. Between-subjects evaluations assessed how IDs affected these outcomes.

2.4.2 Independent Variables

The independent variables were ASM transparency level and ASM reliability. Transparency of agent reasoning was manipulated by varying the depth of SAT model information displayed; surface-level SAT model information (S) only versus surface-level plus the in-depth factors for the SAT model information (D). Agent reliability was manipulated by varying the ASM's error rate (67% versus 100% reliable) for its responses to events occurring in the environment. Participants completed four missions, one in each of the conditions (Table 1). The condition sequence was fully counterbalanced across participants.

Table 1 ASM Reliability by ASM transparency level design matrix

		ASM transparency level	
		Surface-level information only (S)	Surface-level plus in-depth information (D)
ASM reliability	100% reliable (R)	100% reliable/surface-level information (SR)	100% reliable/surface-level plus in-depth (DR)
	67% reliable (U)	67% reliable/surface-level information (SU)	67% reliable/surface-level plus in-depth (DU)

2.4.3 Dependent Measures

2.4.3.1 Target Detection Task

During each mission, participants conducted a target detection task wherein they were assigned to identify a specific vehicle whenever it appeared. Each mission contained 15 target and 80 "noise" vehicles. These vehicles were distributed throughout the simulation environment, and the participant encountered them at a rate of approximately eight vehicles per min. Performance on the target detection task was assessed as an indication of whether the participant was able to stay fully engaged when agent reasoning transparency increased and/or when agent reliability decreased. All participants performed well on this task; as such, there was a ceiling effect for the number of targets identified and a floor effect on the number of false alarms. However, there was variance in how many total clicks participants made while conducting this task. Therefore, efficiency in correctness on this task was evaluated as an assessment of participant engagement. First, the correct target score was calculated by finding the ratio of correct detections to total targets (#correct/#total targets). Then, the efficiency score was calculated by finding the ratio of correct detections to the total number of clicks (#correct/#total clicks). Finally, the threat detection correctness efficiency (TDCE) score was calculated by multiplying the correct targets score by the efficiency score.

2.4.3.2 Event Identification Task

There were six events in each condition. Events could be identified ±30 s of the event trigger. Participants were instructed to use the buttons located on the lower portion of the left-side screen to identify each event as soon as they could. Incorrect answers could be corrected; the simulation logged the last answer recorded. Participants were scored on RT and whether they correctly identified the event. RT reflects participants' average time to identify events, regardless of correctness. Lower RTs indicate faster event identification attempts. If the participant did not attempt to identify the event, it was assumed they timed out; as such, they were scored as if they incorrectly identified the event and took the maximum allowable time to do so.

Event task (ET) score reflects participants' average time to identify events and assesses correctness by including a penalty for each incorrect response (i.e., adding a time penalty for each incorrect identification). The time penalty is the overall average RT from all conditions (i.e., 26.35 s). In both RT and ET, lower scores indicate better performance.

2.4.3.3 Workload

After each mission, the NASA-TLX was administered to assess the participants' perceived workload. Both global and individual factor workload scores were evaluated.

Participants' fixation count, fixation duration (in seconds), and pupil diameter (in millimeters) metrics were also collected during each scenario as real-time objective measures of cognitive workload.

2.4.3.4 SA Scores

To assess the participant's current awareness of their environment and the ASM through all three SA levels, the Situation Awareness Global Assessment Technique (SAGAT) was employed (Jones and Kaber 2004). SAGAT is a method where SA-related queries are administered to participants during predetermined freezes of the simulation during the task under analysis (Jones and Kaber 2004; Salmon et al. 2009). During each mission, the simulation was paused six times. At each pause, the participant answered three SA Level 1 queries, three SA Level 2 queries, and two SA Level 3 queries (Appendix K). SA queries were designed to assess the participants' SA at a specific SA level (i.e., SA1: level 1 SA, perception; SA2: level 2 SA, reasoning, comprehension; SA3: level 3 SA, the projection of future state; see Appendix K for example questions). SA1 and SA2 queries were scored as correct (+1) or incorrect (–1), with higher scores indicating better SA.

SA3 queries were designed to assess the participant's projection as to how reliable and accurate the ASM would be in the following encounter, based upon its performance in the current encounter. For purposes of this question, reliability was described as how well the robot's actions were suited to the event, while accuracy was defined as how well the robot correctly identifies what the squad is encountering and/or the squad's actions. SA3 queries were scored using a 4-point Likert scale (e.g., 4 = reliable, 1 = unreliable), with higher numbers indicating more reliable or accurate.

In addition to SA, the related concept of confidence in one's SA was assessed (Endsley and Jones 1997; McGuinness 2004). Each SA query had an associated confidence evaluation, where the participant rated their level of confidence in their response to the SA query using 5-point Likert scales (e.g., 5 = very confident, 1 = not confident). Higher values indicate greater confidence.

2.4.3.5 Trust

After each mission, the Functional Trust Survey was administered to assess the participant's trust and perceived usability of the ASM.

2.4.3.6 Anthropomorphic Tendencies

After each mission, the Godspeed measure was administered to assess the participant's attributions of anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of the ASM.

2.5 Procedure

After being briefed on the purpose of the study and signing the informed consent form, participants were tested for normal color vision using the Ishihara Color Vision Test (Ishihara 1917). Then they completed the IAT, demographics survey, and Spatial Orientation Test.

Participants received training and practice on their tasks. The training was self-paced and delivered by PowerPoint slides. Participants were trained on the elements of the ASM interface, their specific tasks, and how to evaluate the ASM. The training included assessments and participants had to score 90% or better to continue with the training. Those who scored too low on the assessments were allowed to review the information again and be reassessed. After completing the PowerPoint slides, there was a practice session to familiarize the participants with the experimental environment. The training session lasted approximately 45 min. Participants were allowed a short (under 5 min) break before continuing to the experiment.

After training, the eye-tracking system was calibrated for the participant and the experimental session began.

During the experiment, participants monitored a simulated dismounted Soldier team, accompanied by an ASM, as it traversed a training course four times (four missions). Each time the Soldier team completed the course, they took a different route and were accompanied by a different ASM. As the team traversed the training course, events would occur, and the team (and ASM) would have to respond accordingly. Each mission had six events (i.e., improvised explosive device [IED], sniper fire, ambush, obstruction in the road, civilians in the area, and flooding), and the order in which the events were presented varied between missions. Participants were required to identify the event, as well as other potential threats, as soon as they became aware of the threat, using on-screen buttons. Six events occurred during each training course; in the 100% reliable (R) condition, the ASM always responded correctly to the event, and in the 67% reliable condition (U), the ASM did not respond correctly to the event two out of six times. The Soldier team always responded correctly to the events.

In the Surface-Level Information (S) only condition, the ASM display would show the robot's perceptions of the team's actions, its current objective, the reason behind this objective, and projected resource loss associated with this objective. In the Surface-Level plus In-Depth Information (D) condition, the ASM display would also show the underlying factors behind the current objective. Participants were required to monitor the environment and the ASM interface and received SA queries throughout each trial.

After each mission, participants assessed their perceived workload, trust in the ASM, and anthropomorphic tendencies of the ASM. After completing four missions (one in each condition), the participants completed the working memory assessment, visual scanning task, and attentional control survey. Participants were then debriefed, and any questions they had were answered by the experimenter.

3. Results

Data analysis was performed using SPSS Version 24 software. Data were examined using repeated-measures analysis of variances (ANOVAs) ($\alpha = 0.05$), with a Bonferroni correction for multiple comparisons when applicable. Planned comparisons were conducted to examine differences between conditions, specifically, (SR) to (DR), (SU) to (DU), (SR) to (SU), and (DR) to (DU).

When there was a violation of the homogeneity of variance assumption, Welch's correction was used, and contrast tests do not assume equal variance between conditions. Means, SD, and 95% confidence interval (CI) are reported for each measure.

ID factors (i.e., IAT, VS, SO, PAC, WMC, GE) were assessed as potential covariates. When an ID factor was revealed to be a significant predictor or correlate highly with the measure of interest, these results are reported.

Preliminary GPower 3.1.3 analysis indicated that 56 participants in a repeated-measures ANOVA had an estimated power of 0.8 at a medium-to-large effect size ($f = 0.35$).

3.1 Task Performance

3.1.1 Target Detection Task

Participant performance on the target detection task was evaluated using the TDCE score (Section 2.4.3.1). ID factors were examined for correlations with TDCE scores (Appendix K, Table K-1).

Spatial orientation (SO) test scores significantly predicted TDCE scores in the surface-level information conditions, regardless of ASM reliability (see Appendix K, Table K-2 for ID factor regression analysis). There was no correlation between SO and TDCE scores in the deep-level information conditions. SO accounted for 7% of the variance in TDCE scores in condition SR and 8% in condition SU. When the ASM provided surface-level information only, participants whose SO scores indicated they were better able to orient themselves in a 3-D world outperformed their counterparts on the threat-identification task.

A repeated-measures ANOVA found no difference in TDCE scores due to information level and/or ASM reliability, Wilks' $\Lambda = 0.993$, $F(3, 53) = 0.12$, $p = 0.945$, $\omega^2 = 0.00$. Paired t-tests were used for inter-condition planned comparisons and found no significant differences in performance on the threat-detection task due to information level and/or ASM reliability. Statistical information for TDCE scores can be found in Appendix K; specifically, descriptive statistics are in Table K-3, and planned comparison results are reported in Table K-4.

3.1.2 Event Identification Task

Although there were no hypotheses relating to proper event identification or RT, these were evaluated to ensure the participants were accurately identifying each event and doing so in a timely manner. These results are crucial to support later analysis because participants were also queried as to the ASM's accuracy in identifying events and reliability in responding to events.

Participant performance on the event identification task was evaluated using the RT and ET scores (Section 2.4.3.2). ID factors were examined for correlations with RT and ET scores (Appendix K, Table K-5). Aside from several correlations that were deemed spurious, there were no meaningful correlations between any ID factor and workload scores.

Repeated-measures ANOVAs found no difference in RT or ET scores due to information level and/or ASM reliability, RT: Wilks' $\Lambda = 0.937$, $F(3, 53) = 1.20$, $p = 0.320$, $\omega^2 = 0.00$; ET: Wilks' $\Lambda = 0.953$, $F(3, 53) = 0.88$, $p = 0.457$, $\omega^2 = 0.00$. Inter-condition planned comparisons were made using paired t-tests, and no differences due to information level or agent reliability were found. Statistical information for RT and ET scores can be found in Appendix K; specifically, descriptive statistics are in Table K-6, and paired comparison results are reported in Table K-7.

3.2 Workload

3.2.1 NASA-TLX

The perceived cognitive workload was assessed using the NASA-TLX survey. In addition to the global workload score, the scores for factors mental demand, effort, performance, and temporal demand were also evaluated for differences due to changing information level or ASM reliability.

ID factors were examined for correlations with perceived workload, both overall and on each of the subscales (Appendix L, Table M-1). Aside from several correlations that were deemed spurious, there were no meaningful correlations between any ID factor and workload scores.

Repeated-measures ANOVAs were used to evaluate NASA-TLX scores to assess differences in participants' perceived workload due to information level and/or ASM reliability. Statistical information for workload analysis can be found in Appendix M; specifically, ANOVA results are in Table M-2, descriptive statistics are in Table M-3, and paired comparison results are reported in Table M-4.

There was no statistically significant difference in global NASA-TLX scores, nor on the subscales of mental demand, effort, performance, or temporal demand, due to either information level or ASM reliability. Overall, the effect sizes for differences between conditions were small.

3.2.2 Eye-Tracking Metrics

Cognitive workload was also evaluated using several ocular indices. Statistical information for workload analysis can be found in Appendix M; specifically, ANOVA results are in Table M-5, descriptive statistics are shown in Table M-6, and paired comparison results are reported in Table M-7. Not all participants had complete eye-measurement data, so this N was reduced (n = 51). Eye-tracking data were evaluated using the same planned comparisons as the subjective workload measure.

There was no significant effect on participants' pupil diameter, fixation count, or fixation duration as a result of either information level or ASM reliability level. Planned comparisons did not reach statistical significance; as such, there was no indication of any difference in cognitive workload due to information level or ASM reliability.

These ocular indices findings could indicate that the additional information offered in the in-depth information conditions (i.e., DR and DU) was not used by the participants. To verify participants were availing themselves of all information

offered, an area of interest (AOI) analysis was performed to examine participant fixation count in the ASM logic factors AOI. Repeated-measures ANOVA indicated a significant difference in AOI usage, Wilks' $\Lambda = 0.718$, $F(3, 48) = 6.287$, $p = 0.001$, $\omega^2 = 0.09$. Paired comparisons showed this difference was due to the information level, but not the ASM reliability level (Table M-9). Regardless of ASM reliability, participants in the in-depth information conditions had more fixations in the ASM logic AOI than their surface-level information counterparts. This demonstrates that the additional information was accessed by participants when available, yet it did not lead to an increase in workload.

3.3 Situation Awareness

3.3.1 SA1 (Perception) and SA2 (Comprehension)

Participant's SA was evaluated using SAGAT-style queries (Section 2.4.3.3). ID factors were examined for correlations with SA1 and SA2 scores (Appendix N, Table N-1). There was a significant correlation between WMC and SA1 scores in that those with greater WMC appeared to have higher SA1 scores in the reliable ASM conditions. There appeared to be no other meaningful correlations between ID factors and SA scores.

SA1 and SA2 scores and confidence ratings were assessed via repeated-measures ANOVAs, with paired t-tests used for inter-condition planned comparisons to evaluate differences in SA due to information level or agent reliability (Table 2). The SA1 percent-correct scores failed the sphericity test, so the Greenhouse–Geisser correction for sphericity is reported. There was no significant difference in the percentage of correct answers or participant confidence in their SA responses between the experimental conditions for all assessments. Descriptive statistics by the experimental condition are shown in Table N-2, paired t-test results are shown in Table N-3. Overall, the number of correct responses was quite high, over 83% for both SA1 and SA2 queries, regardless of condition, with confidence ratings of 4.6 and higher (out of 5).

Table 2 Repeated-measures ANOVA results for SA1 and SA2 results across conditions

Measure		df_1	df_2	F	p	ω^2
SA1	% correct	2.65	148.65	0.39	0.733	–0.01
	Confidence	3.00	168.00	1.23	0.300	0.00
SA2	% correct	3.00	168.00	1.57	0.199	0.01
	Confidence	3.00	168.00	0.81	0.492	0.00

3.3.2 SA3 (Projection of ASM Future State)

ID factors were examined for correlations with SA3 responses (Appendix N, Table N-4). There was a significant correlation between perceived attentional control and confidence scores in that those with greater attentional control appeared to have higher confidence in their assessments of projected reliability and accuracy in the low information, reliable ASM condition. There appeared to be no other meaningful correlations between ID factors and SA3 scores.

SA3 scores and confidence ratings were assessed via repeated-measures ANOVAs, with paired t-tests used for inter-condition planned comparisons to evaluate differences in projected ASM reliability and/or accuracy due to information level or agent reliability (Table 3). The Greenhouse–Geisser correction for sphericity is reported for all scores. There was a significant difference in participant projections of the ASM's reliability and accuracy, as well as a difference in their confidence in their evaluations, between the experimental conditions. Descriptive statistics by the experimental condition are shown in Table N-5, paired t-test results are shown in Table N-6. There was no significant difference in participant projections of ASM reliability or accuracy due to information level; however, there were differences due to agent reliability. In the conditions where the ASM made an error, projected reliability and accuracy scores were lower than those in the non-error conditions. Interestingly, participant confidence in their assessment of the projected reliability and accuracy was also significantly lower in the error conditions than in the non-error conditions.

Table 3 Repeated-measures ANOVA results for SA3 results across conditions

Measure		df_1	df_2	F	p	ω^2
Reliability	Score	2.39	131.49	114.29	<0.001	0.65
	Confidence	2.33	128.38	13.31	<0.001	0.56
Accuracy	Score	2.35	129.01	131.27	<0.001	0.67
	Confidence	2.38	130.60	18.63	<0.001	0.18

3.3.2.1 Effect of Errors on Participant Perception of Reliability and Accuracy

The effect of ASM reliability on participant evaluations was also examined for potential carryover effects. Each scenario was comprised of six events. The first event, and in some cases the first and second events, was always error-free. Then the ASM would make an error. The next event was (again) error-free, and in some cases, the next two events were error-free. Then there would be another error event, again followed by an error-free event. The scores across scenarios were aggregated, and the events were grouped as such: Beginning (all events from beginning of scenario up to the first error event), Error (all events where the ASM made an error),

1st Following (the first non-error event following an error event), and 2nd Following (the second non-error event following an error event). Differences in scores were evaluated by whether the participant witnessed an error during that event (E; SU and DU conditions) or not (NE; SR and DR conditions).

Repeated-measures ANOVAs were used to evaluate differences in participant projections of the ASM's reliability and/or accuracy due to agent reliability and significant differences were found (Table N-12). Descriptive statistics and between-condition comparison results are shown in Table N-13. Participants were asked to project how accurate and reliable the ASM would be in the following event, based upon its assessment and behavior in the current event. As such, if the ASM displayed accuracy in perceptions and reliability in actions, the projections for the following event should be "accurate" and "reliable" (score 4/4).

The Beginning event in each scenario was error-free, and there was no difference in assessments between the NE (no error witnessed) and E (error witnessed) groups (Figs. 4 and 5). During the Error event, there was a significant difference in projections of future reliability and accuracy between the two groups, as was expected. The first event following an Error event was also error-free; however, there was a significant difference between the groups. Although their scores were considerably higher than those for the preceding event, participants who previously witnessed an error (E) scored the ASM's projected accuracy and reliability significantly lower than those that did not witness an error (NE). When the second event following the error event was also error-free, the E-group's scores continued to trend upward but were still significantly lower than the NE-group's scores for both projected accuracy and reliability. Although participants were trained to distinguish between errors in ASM accuracy (perceptions, identifying situations) and reliability (behavior appropriate to the perceptions), in practice it appears they did not distinguish between the two, scoring accuracy and reliability similarly even though the ASM never made a behavioral error (its behavior always was appropriate for the situation as interpreted).

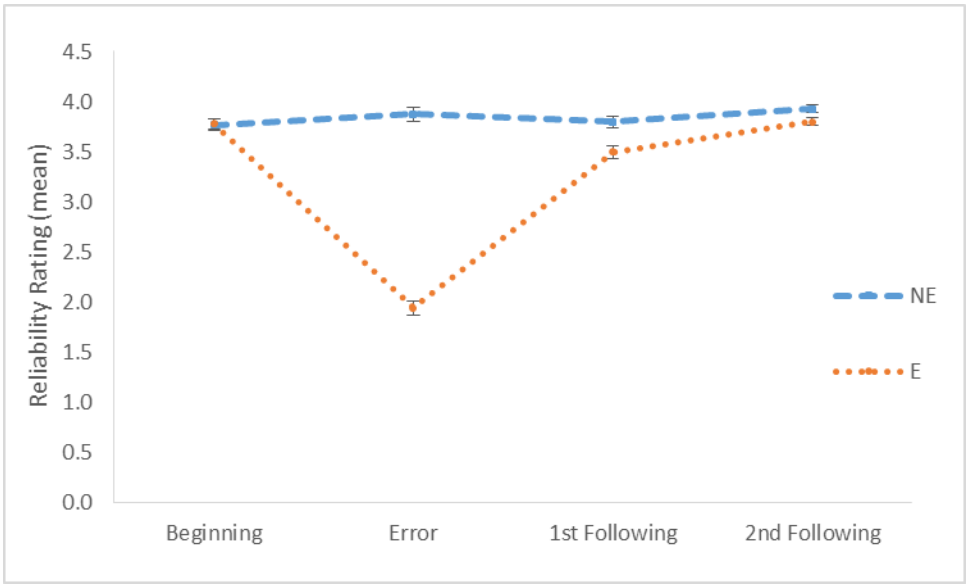


Fig. 4 Participant ASM reliability projection scores by event, sorted by whether the participant witnessed an ASM error during that event

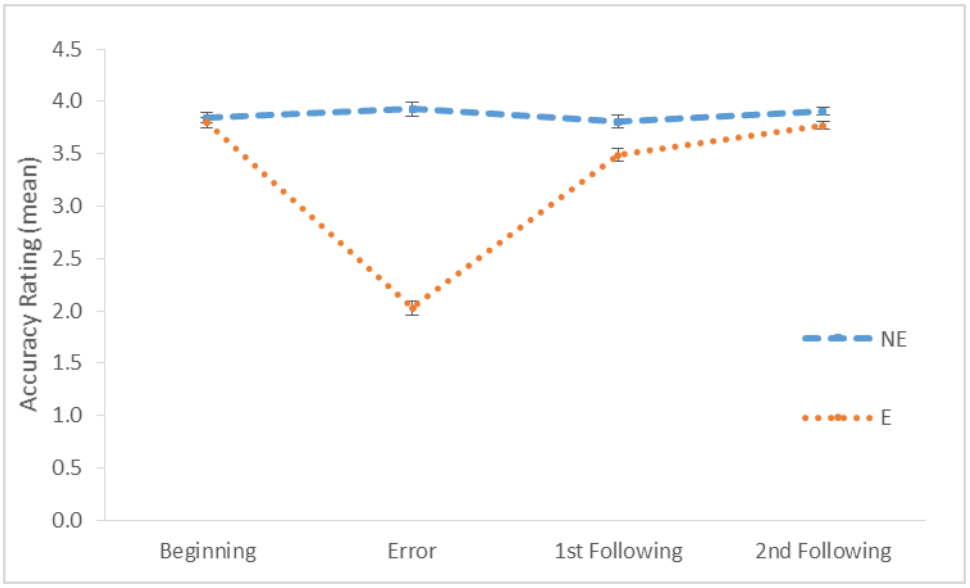


Fig. 5 Participant ASM accuracy projection scores by event, sorted by whether the participant witnessed an ASM error during that event

3.3.2.2 Effect of Errors on Participants' Confidence in their Assessment of Reliability and Accuracy

Participants' confidence in their evaluations of ASM reliability and accuracy was also examined. Repeated-measures ANOVAs found significant differences in the participants' confidence in their assessments due to agent reliability (Table N-14).

Descriptive statistics and between-condition comparison results are shown in Table N-15.

There was no difference in participant confidence between the NE (no error witnessed) and E (error witnessed) groups in the Beginning event (Figs. 6 and 7). During the Error event, there was a significant difference in participant confidence between the two groups, as the NE group had an increase in confidence ratings while the E group had a decrease. This difference in confidence ratings was consistent through the 1st and 2nd Following events. Participants who previously witnessed an error (E) reported reduced confidence in their evaluations of the ASM's projected accuracy and reliability than those that did not witness an error (NE).

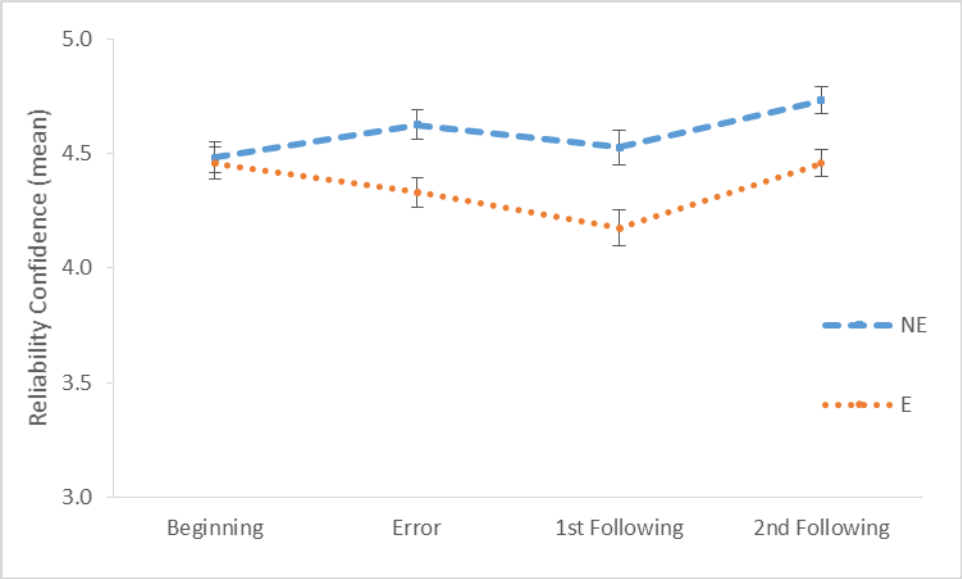


Fig. 6 Participant confidence in reliability ratings by event, sorted by whether the participant witnessed an ASM error during that event

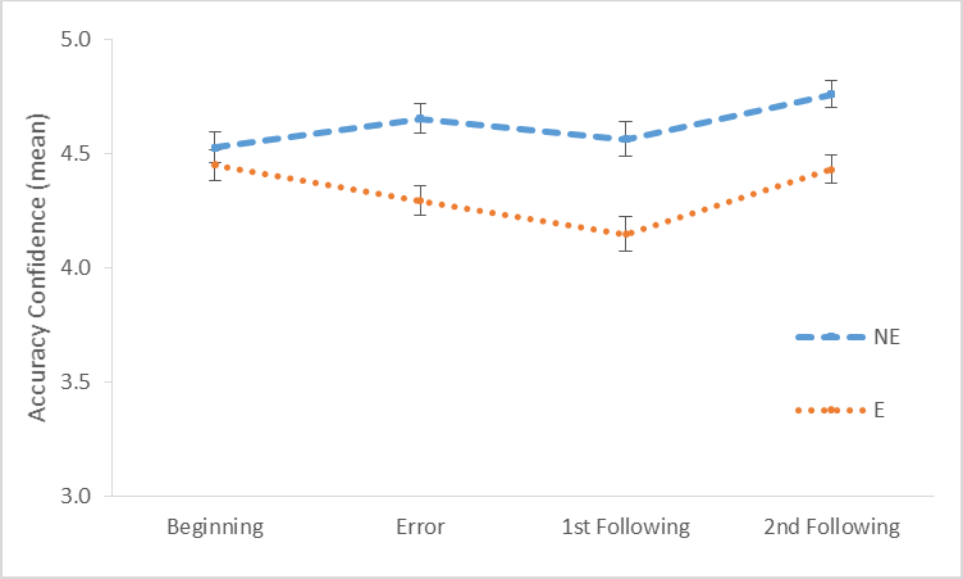


Fig. 7 Participant confidence in accuracy ratings by event, sorted by whether the participant witnessed an ASM error during that even

Confidence scores for those participants who had witnessed an error were then examined to determine whether the ASM information level had any effect. Repeated-measures ANOVAs revealed no significant difference in confidence scores between participants who were in the surface-level information condition compared to those in the deep-level information condition for either confidence in reliability ratings (Wilks' $\Lambda = 0.952$, F (4, 104) = 1.30, $p = 0.275$, $\eta^2 = .048$) or confidence in accuracy ratings (Wilks' $\Lambda = 0.980$, F (4, 104) = 0.05, $p = 0.714$, $\eta^2 = 0.020$).

3.4 Trust and Usability

3.4.1 Functional Trust Survey (Trust in Automation)

3.4.1.1 Jian Trust in Automation Survey

Questions 1–12 of the Functional Trust Survey are from the Jian et al. (2000) Trust in Automation Survey. Participant trust was evaluated both as an overall measure by condition, and then along the four automation functional stages outlined by Parasuraman et al. (2000). ID factors were examined for correlations with trust in automation (Appendix O, Table O-1).

PAC significantly predicted trust scores when the ASM was reliable (see Appendix O, Table O-2 for ID factor regression analysis). There was no correlation between PAC and trust scores when the ASM was unreliable, regardless of the information level. PAC accounted for 14% of the variance in trust scores in

condition SR, and 7% in condition DR. When the ASM made no errors, participants with greater attentional control reported higher trust in the ASM than those with lower attentional control, and this was more pronounced when the information level was low compared to when it was greater.

WMC significantly predicted trust scores when information level was high and the ASM was reliable. There was no correlation between WMC and trust when the information level was low or when the ASM was unreliable. WMC accounted for 13% of the variance in trust scores in condition DR. When the information level was high and the ASM made no errors, participants with greater WMC reported higher trust in the ASM than those with lower WMC.

Overall participant trust in the ASM was assessed via a one-way ANOVA with planned comparisons. There was concern that participant fatigue affected outcomes in later scenarios, therefore only data for the first scenario were assessed.

There was a significant difference in overall participant trust between the experimental conditions, $F(3, 52) = 4.41$, $p = 0.008$, $\omega^2 = 0.15$ (see Table O-3 for descriptive statistics). Planned comparisons indicated this difference was due to agent reliability, not varying information levels (Table 4). When surface-level information was offered, overall trust was 18% higher in the reliable condition than the unreliable condition. When more in-depth information was offered, overall trust was 15.5% higher in the reliable condition than in the unreliable condition.

Table 4 Between-condition comparisons of participant scores on the Jian Trust in Automation Survey

Planned comparison	Δ*M*	*SE*	*P*	d_s	*% difference*
SR > SU	12.21	5.29	0.032	0.90	18.25
SR ≈ DR	–2.19	3.98	0.586	–0.21	–3.28
SU ≈ DU	–3.71	5.28	0.491	–0.27	–6.77
DR > DU	10.70	3.96	0.012	0.99	15.48

Participant trust in the ASM was also assessed along the four functions of automation using one-way ANOVAs with planned comparisons. There was a significant difference in trust in the ASM's ability to perform each of the automation functions between the experimental conditions (Table 5). Although the assessment was not statistically significant in Function B, this is most likely due to the small *N*, as the effect sizes for the overall test and the planned comparisons were large. Information level had no significant effect on operator trust, except in Function A (collecting and/or filtering information). Although the test did not show statistical significance, the effect sizes indicate that deeper-level information does bolster trust, particularly when the ASM is unreliable. ASM reliability had a

significant effect on operator trust across all automation functions (Table 6). Operator trust in the ASM was consistently lower when the ASM was unreliable.

Table 5 One-way ANOVA results for participant scores on the Jian Trust in Automation by automation function, across all conditions

Automation function	*F(3, 52)*	*p*	ω^2
A: Collecting and/or filtering information	3.91	0.014	0.13
B: Integrating and displaying analyzed information	1.92	0.138	0.05
C: Making decisions and/or selecting actions	5.21	0.003	0.18
D: Executing actions	5.98	0.001	0.21

Table 6 Between-condition comparisons of participant trust scores on the Jian Trust in Automation survey, by automation function

Function	Paired comparison	Δ*M*	*SE*	*p*	d_s	*% difference*
A	SR > SU	11.92	5.68	0.048	0.82	17.53
	SR ≈ DR	–3.13	3.93	0.434	–0.31	–4.61
	SU ≈ DU	–7.39	5.46	0.192	–0.53	–13.18
	DR > DU	7.67	3.61	0.042	0.78	10.78
B	SR > SU	8.31	5.11	0.110	0.58	12.12
	SR ≈ DR	–1.33	4.94	0.789	–0.13	–1.94
	SU ≈ DU	–1.44	4.94	0.772	–0.09	–2.38
	DR > DU	8.20	4.76	0.091	0.70	11.74
C	SR > SU	14.00	5.54	0.020	0.99	21.64
	SR ≈ DR	–2.91	4.41	0.515	–0.25	–4.49
	SU ≈ DU	–3.84	5.61	0.501	–0.26	–7.58
	DR > DU	13.07	4.50	0.007	1.06	19.33
D	SR > SU	14.62	5.42	0.013	1.06	21.99
	SR ≈ DR	–1.41	4.46	0.755	–0.12	–2.11
	SU ≈ DU	–2.15	5.23	0.685	–0.16	–4.15
	DR > DU	13.87	4.23	0.003	1.20	20.43

3.4.1.2 Trust Perception Survey–HRI

Questions 13–16 of the Functional Trust Survey are from the Trust Perception Scale –HRI (Schaefer 2016). Participant trust was evaluated both as an overall measure by condition, and then along the four automation functional stages outlined by Parasuraman et al. (2000).

There was concern that participant fatigue affected outcomes in later scenarios, therefore only data for the first scenario were assessed. Overall participant trust in the ASM, as well as trust along the four functions of automation factors, was assessed via one-way ANOVAs with planned comparisons.

Results from the Schaefer Trust Survey were similar to those found using the Jian Trust in Automation Scale. Analysis results for these findings are shown in Appendix O, Tables O-15–O-18.

3.4.2 Usability (Technology Acceptance Measure)

The perceived usability of the ASM was evaluated via items from the technology acceptance measure (TAM; Davis 1989) included in the Functional Trust Survey. The aggregate TAM score indicates overall perceived usability, while the subscores indicate perceived ease of use, usefulness, and intent to use in the future. Scores were assessed via repeated-measures ANOVAs, with paired t-tests used for inter-condition planned comparisons to evaluate differences due to information level or agent reliability. The perceived usefulness evaluation failed the sphericity test, so the Greenhouse–Geisser correction for sphericity is reported.

There was a significant difference in participant's perceived usability between the experimental conditions for all assessments (Table 7). Descriptive statistics by the experimental conditions are shown in Table O-8. There was a significant difference in the perceived usefulness, intent to use, and overall TAM scores due to agent reliability, but not due to information level (Table 8). When the ASM was unreliable, the scores were lower than when the ASM was reliable.

However, it appears that the depth of information may have mitigated the effect of ASM reliability for the perceived ease of use results in that there was no statistically significant difference between the DR and DU conditions. As the mean scores in the in-depth information conditions were very similar, yet lower than in the SR condition, this could indicate that participants found the additional information helpful even when the ASM was unreliable.

Table 7 Repeated-measures ANOVA results for TAM scores

Measure	df_1	df_2	F	p	ω^2
Overall TAM	3	165	17.47	<0.001	0.24
Perceived ease of use	3	165	3.51	0.017	0.04
Intent to use	3	165	13.34	<0.001	0.17
Perceived usefulness	2.5	138.4	23.42	<0.001	0.27

Table 8 Between-condition comparisons of participant scores on the TAM survey

Measure	Paired comparison	Δ*M*	*SE*	*p*	d_s	*% difference*
Overall TAM	SR > SU	1.10	0.18	<0.001	0.99	19.34
	SR ≈ DR	0.10	0.18	0.591	0.09	1.71
	SU ≈ DU	–0.01	0.22	0.963	–0.01	–0.22
	DR > DU	0.99	0.22	<0.001	0.90	17.76
Perceived ease of use	SR > SU	0.61	0.21	0.006	0.51	10.49
	SR ≈ DR	0.17	0.22	0.433	0.15	2.98
	SU ≈ DU	–0.04	0.23	0.875	–0.03	–0.69
	DR ≈ DU	0.40	0.25	0.115	0.33	7.10
Intent to use	SR > SU	1.48	0.25	<0.001	0.87	28.82
	SR ≈ DR	0.09	0.31	0.776	0.05	1.74
	SU ≈ DU	–0.09	0.34	0.793	–0.05	–2.44
	DR > DU	1.30	0.31	<0.001	0.75	25.80
Perceived usefulness	SR > SU	1.47	0.23	<0.001	1.05	25.49
	SR ≈ DR	–0.02	0.19	0.924	–0.02	–0.31
	SU ≈ DU	0.04	0.29	0.888	0.03	0.97
	DR > DU	1.53	0.27	<0.001	1.10	26.44

3.5 Anthropomorphic Measures

Anthropomorphic perceptions were assessed using the Godspeed measure. ID factors were examined for correlations with anthropomorphic perceptions (Appendix P, Table P-1). There were no meaningful correlations between any ID factor and the Godspeed measures.

Scores on the Godspeed measures were assessed via repeated-measures ANOVAs, with paired t-tests used for inter-condition planned comparisons to evaluate differences in participant perceptions of the ASM due to information level or agent reliability. All measures failed the sphericity test, so the Greenhouse–Geisser correction for sphericity is reported.

There was a significant difference in participant's perceptions of the ASM between the experimental conditions for all assessments (Table 9). Descriptive statistics by the experimental conditions are shown in Table P-2. There was a significant difference in participant's perceptions of the ASM due to agent reliability, but not due to information level (Table 10). When the ASM was reliable, participants anthropomorphized the agent more than when it was unreliable. Participants rated the ASM as more animate, likable, intelligent, and safer to work with when the agent was reliable than when it was not.

Table 9 Repeated-measures ANOVA results for Godspeed measure scores

Measure	df_1	df_2	F	p	ω^2
Anthropomorphism	2.55	140.28	16.42	<0.001	0.15
Animacy	2.41	132.69	20.91	<0.001	0.21
Likeability	2.39	131.42	26.68	<0.001	0.27
Perceived intelligence	2.27	124.77	62.75	<0.001	0.50
Perceived safety	2.58	141.93	11.30	<0.001	0.13

Table 10 Between-condition comparisons of participant scores on the Godspeed survey

Measure	Paired comparison	ΔM	SE	p	d_s	% difference
Anthropomorphism	SR > SU	0.40	0.10	<0.001	0.51	14.04
	SR ≈ DR	–0.09	0.08	0.262	–0.11	–3.23
	SU ≈ DU	0.04	0.10	0.685	0.05	1.59
	DR > DU	0.54	0.09	<0.001	0.64	18.05
Animacy	SR > SU	0.60	0.10	<0.001	0.84	18.07
	SR ≈ DR	0.00	0.07	1.000	0.00	0.00
	SU ≈ DU	–0.13	0.09	0.178	–0.19	–4.63
	DR > DU	0.47	0.10	<0.001	0.64	14.27
Likeability	SR > SU	0.63	0.10	<0.001	0.93	16.81
	SR ≈ DR	0.00	0.08	0.965	–0.01	–0.10
	SU ≈ DU	–0.03	0.09	0.772	–0.04	–0.81
	DR > DU	0.60	0.10	<0.001	0.86	16.22
Perceived intelligence	SR > SU	1.19	0.13	<0.001	1.77	27.63
	SR ≈ DR	0.01	0.07	0.881	0.02	0.25
	SU ≈ DU	–0.02	0.12	0.831	–0.03	–0.80
	DR > DU	1.15	0.13	<0.001	1.53	26.87
Perceived safety	SR > SU	0.45	0.09	<0.001	0.78	12.67
	SR ≈ DR	0.05	0.09	0.565	0.09	1.52
	SU ≈ DU	–0.05	0.08	0.528	–0.09	–1.74
	DR > DU	0.34	0.11	0.004	0.53	9.78

4. Discussion

4.1 Synopsis and Review

The goal of this study was to examine how the transparency and reliability of an ASM affected a human observer's task performance, workload, SA, trust in the agent, perceived usability, and perceptions of the agent. Agent transparency typically is examined within perfectly reliable systems; however, it is known that unreliability severely impacts operator performance and perceptions of a system. Herein we examined whether increased agent transparency would mitigate the undesirable effects of unreliable automation, thus supporting the human's trust calibration and understanding of the agent.

Participants monitored a simulated Soldier squad that included an ASM as it traversed a simulated training environment, while concurrently monitoring the environment for targets and identifying events that the squad encountered. In this within-subjects design study, participants completed four scenarios, one in each of the agent transparency (surface-only vs. deep-level information) and reliability (reliable vs. unreliable) combinations. Each scenario contained six events, and in two of the events the ASM would incorrectly identify the event. The ASM's actions were always appropriate for the event as identified. During each scenario, participants were periodically asked questions to gauge their SA, perceived accuracy, and perceived reliability of the ASM, as well as their confidence in their assessments.

The main purpose of the target detection task was to encourage participant engagement and enable the investigators to assess whether the participant was correctly identifying the events and elements in the environment. As such, these tasks were not particularly challenging, and the event rate for the target presentation was quite low. Participant tasks were to monitor the environment for threats (target detection task) and identify the events as the squad encountered them.

Performance on the target detection task was evaluated by how efficiently participants identified the targets. Although the primary purpose of the task was to maintain participant engagement, it was still expected that performance would be negatively impacted in the unreliable conditions, and even more so when in-depth information was presented during the unreliable condition. These hypotheses (i.e., H1 and H2) were not supported; there was no difference in target detection performance regardless of agent reliability or transparency level. After each scenario, the participant's perceived cognitive workload was evaluated. It was expected that the workload would increase in the unreliable conditions, and the addition of in-depth information in the unreliable condition would further increase workload. However, these hypotheses (i.e., H3 and H4) were not supported. Overall, these findings indicate that all participants were fully engaged, had no difficulty accurately identifying the events as presented, and were not overtaxed by the demands of the simulation.

The participants were also tasked with monitoring ASM communications and actions, assessing its accuracy in identifying events, and reliability in its responses to those events. All participants had very high SA1 (perception) and SA2 (comprehension) scores throughout the scenarios, which did not support the expectation (i.e., H5 and H6) that SA would be negatively affected when the agent was unreliable.

However, participant SA3 (projection of future state) scores were quite different from their SA1 and SA2 scores. Not altogether surprising, agent reliability influenced participant projection of ASM future accuracy and reliability, in that participants in the unreliable conditions rated the ASM's reliability and accuracy lower than their reliable-condition counterparts, thus supporting H6. Furthermore, there appeared to be carryover effects from witnessing the ASM error in event identification. Once an error occurred, those participants in the unreliable condition continued to rate the ASM as less accurate and reliable than those in the reliable condition, even when the ASM did not make any additional errors. These lowered ratings did improve as the ASM continued to identify events accurately, and appeared to be trending to a point where they would once again be similar to those in the reliable conditions. It was expected (i.e., H5) that in-depth information from the ASM might mitigate these effects. However, that was not the case, as information level appeared to not influence participants' projections of ASM accuracy and reliability. Prior research showed users may begin their work with an unfamiliar system with overly high expectations of the systems' performance (positivity bias; Cacioppo and Berntson 1994; Cacioppo et al. 1997), and witnessing an error causes them to overcorrect their expectations, resulting in lower assessments of reliability than warranted (Dzindolet et al. 1999). However, it appears as though continued experience with the ASM allowed the participants to adjust their assessments of the system's reliability over time. This is similar to findings that users will correct their low-reliability assessments of a system over time, so long as the user witnesses no further errors (Dzindolet et al. 2003).

Confidence in SA has been identified as a crucial element in effective decision-making and performance (Endsley and Jones 1997; McGuinness 2004). While there was no difference in individual self-confidence ratings in SA1 and SA2 responses due to information level or ASM reliability, there were differences in confidence in SA3 assessments due to ASM reliability. At the beginning of the scenarios, when no ASM error had occurred, there was no difference in reported confidence between participants in the two conditions. Once the ASM made an error in identification, those in the unreliable conditions reported significantly lower confidence in their assessment of ASM accuracy and reliability than their reliable-condition counterparts. Surprisingly, this difference in confidence was consistent throughout the remainder of the scenario, that is, even when no further errors occurred, participants who had witnessed an ASM error continued to report low confidence in their ability to assess and predict agent reliability.

Prior research has shown that participants report higher confidence in their responses when agreeing with a decision aid, rather than when they disagree (Wiegmann 2002) because consistently agreeing with the automation allows

participants to assess the automation's reliability apart from their own. When the ASM initially erred, the participants reported lower reliability. However, on subsequent events, when the ASM did not err, and they continued to rate its reliability as lowered, they were aware of their error in rating the reduced reliability. Even though their assessment of the ASM's future reliability continued to increase as the error event became more distant, their continued awareness of their own inability to predict accurately the ASM's reliability affected their subjective confidence. While subjective confidence is not a reliable indicator of the validity of one's judgments (Kahneman and Klein 2009), as Endsley and Jones pointed out, "SA is attributed with a certain degree of confidence based on the source of the information upon which it is founded and their confidence in their ability to process that information into comprehension and projections" (1997, p. 36). Hence, the lowered confidence ratings are most likely a result of the participant's awareness of their own inability to predict the ASM's future behavior accurately. Reduced confidence in decision-making could have deleterious results in a military setting, as persons with lower confidence tend to be hesitant, overly cautious, and slow to action (Lichacz et al. 2003).

Participants did not appear to distinguish between agent accuracy in identifying events and agent reliability in responding to the events as identified. Their training emphasized the distinction between the two, participants were evaluated on how well they understood the differences, and yet the scoring for agent accuracy and reliability were nearly identical throughout all scenarios and conditions. This could be a result of participants deciding that once the agent made a mistake, all following actions were inherently flawed.

However, the ability for the human teammate to distinguish between agent errors (i.e., due to misidentification, algorithms, or sensory information [e.g., errors in judgment], errors due to equipment malfunctions [e.g., unreliable automation]), or recognizing that the agent has revised its goals or objectives could be crucial to effective human–agent teaming as agents become capable of learning from their environment and revising their goals autonomously.

Findings from the Trust in Automation Survey (Jian et al. 2000) indicate that overall participant trust was 15% to 18% higher when the ASM was reliable, regardless of information level. Participants also reported higher trust in the ASM when it was reliable for each of the four automation functions (Parasuraman et al. 2000) than when it was unreliable. Information level did not affect participant trust in the ASM for the automation functions, except for the "Collecting and/or Filtering Information" function, where participants with in-depth information reported 4.6% and 13% higher trust than those with only surface-level information, in the (respectively) reliable and unreliable conditions. This suggests that in-depth

knowledge of the underlying reasons for an agent's actions and reasoning may bolster human trust in the agent and mitigate the effects of unreliable automation, at least so far as specific tasks are concerned (Muir 1994; Dzindolet et al. 2003). Consistent with the trust findings, perceived usability was higher in the reliable conditions than in the unreliable conditions, regardless of the information level.

Agent reliability influenced participant perceptions of the agent, irrespective of the information level. Participants rated the ASM as more animate, likable, intelligent, and safer to work with when the agent was reliable than when it was not. Prior research has shown that participants rate an autonomous agent as more animate, likable, intelligent, and safer when it is more transparent as to its reasoning (Selkowitz et al. 2016); these results indicate that agent reliability also influences perceptions of the agent's anthropomorphic characteristics.

4.2 Limitations and Future Directions

The environmental monitoring tasks (i.e., target detection and event identification) were very low effort, and this most likely contributed to the lack of differences in task performance, workload, and SA. Future research should explore how agent transparency interacts with agent reliability in a high workload setting and could potentially be used to mitigate the undesirable effects of unreliable automation.

Surprisingly, agent transparency had a little-to-no influence on the study findings. Eye-tracker data indicated that participants did use the additional information when it was present. This could indicate that the additional information provided to the participants was not needed to complete the tasks, or that the ASM's responses to events were clearly understood, rendering the additional information superfluous. Clearly, improving understanding as to which tasks require a more in-depth understanding of the agent's reasoning, and how to discern what that depth would entail, is also needed.

5. Conclusions

Agent transparency requirements in simple, low-workload environments may not be the same as those in higher-demand settings. In this study, agent transparency was not essential for task completion, and contrary to previous findings (Mercado et al. 2016; Wright et al. 2017), increased agent transparency did not improve task performance, SA1, SA2, trust in the agent, or anthropomorphism of the agent. However, increased agent transparency also did not increase workload, which agrees with previous research (Mercado et al. 2016). Even though transparency information did not directly support task performance, it influenced the human's perceptions of the robot. Participants trusted in the agent's abilities to collect and

filter information when given evidence of that activity. This trust suggests that future attempts to facilitate transparent human–agent interaction may benefit from reminding participants of the work the agent is doing "under the hood", even when such transparency may not directly benefit task performance.

Agent reliability may be a stronger influence on the human's perceptions of the agent than agent transparency, although both are important to effective human–agent teaming (Chen et al. 2018). Agent errors had a profound and lasting effect on the human teammates' perception of the agent's reliability, causing the human to rate the agent as less reliable even when it did not commit any errors. This effect appeared to diminish as the agent continued to display reliable behavior. Human perceptions of the ASM were influenced by the ASM's reliability, with unreliable behavior resulting in reduced trust and lower anthropomorphic evaluations. Surprisingly, agent error also resulted in participants' reduced confidence in their assessment of the agent's reliability, regardless of the agent's continued error-free behavior, and this effect was persistent over time. Methods to restore the human teammates' confidence in their assessments of the robot should be explored, as it is crucial to appropriate continued use of the robot teammate when the user is aware of system errors.

6. References

Ahlstrom U, Friedman-Berg FJ. Using eye movement activity as a correlate of cognitive workload. Int J Ind Ergon. 2006;36(7):623–636.

Ahmed N, de Visser E, Shaw T, Mohamed-Ameen A, Campbell M, Parasuraman R. Statistical modelling of networked human–automation performance using working memory capacity. Ergonomics. 2014;57(3):295–318.

Barrett GV, Alexander RA, Doverspike D, Cellar D, Thomas JC. The development and application of a computerized information-processing test battery. Appl Meas Educ. 1982;6(1):13–29.

Bartneck C, Kulić D, Croft E, Zoghbi S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Social Robotics. 2009;1(1):71–81.

Beatty J. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychol Bull. 1982;91(2):276–292.

Boyce MW, Chen JY, Selkowitz AR, Lakhmani SG. Agent transparency for an autonomous squad member. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2015. Report No.: ARL-TR-7298.

Cacioppo JT, Berntson GG. Relationship between attitudes and evaluative space: a critical review with emphasis on the separability of positive and negative substrates. Psychol Bull. 1994;115:401–423.

Cacioppo JT, Gardner WL, Berntson GG. Beyond bipolar conceptualizations and measures: the case of attitudes and evaluative space. Pers Soc Psychol Rev. 1997;1:3–25.

Chen JY, Barnes MJ. Supervisory control of multiple robots effects of imperfect automation and individual differences. Hum Factors. 2012;54(2):157–174.

Chen JY, Barnes MJ, Qu Z. RoboLeader: a surrogate for enhancing the human control of a team of robots. Adelphi (MD): Army Research Laboratory (US); 2010. Report No.: ARL-MR-0735.

Chen JY, Durlach PJ, Sloan JA, Bowens LD. Human–robot interaction in the context of simulated route reconnaissance missions. Mil Psychol. 2008;20(3):135.

Chen JY, Lakhmani SG, Stowers K, Selkowitz AR, Wright JL, Barnes MJ. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theoretical Issues in Ergonomics Science. 2018;19(3):259–282.

Chen JY, Procci K, Boyce M, Wright JL, Garcia A, Barnes MJ. Situation awareness-based agent transparency. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2014. Report No.: ARL-TR-6905.

Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly. 1989;13(3):319–340.

Derryberry D, Reed MA. Anxiety-related attentional biases and their regulation by attentional control. J Abnorm Psychol. 2002;111(2):225–236.

Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The role of trust in automation reliance. Int J Hum Comput Stud. 2003;58:697–719.

Dzindolet MT, Pierce LG, Beck HP, Dawe LA. Misuse and disuse of automated aids. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 1999 Sep; Sage, CA. Los Angeles (CA): SAGE Publications. 1999;43(3):339–339.

Endsley M, Jones WM. Situation awareness information dominance and information warfare. Dayton (OH): Logicon Technical Services Inc.; 1997.

Endsley MR. Toward a theory of situation awareness in dynamic systems. Hum Factors. 1995;37(1):32–64.

Evans AW. Safe operations of unmanned systems for reconnaissance in complex environments-army technology objective (SOURCE ATO) field experimentation observations and soldier feedback. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2012. Report No.: ARL-TN-0488.

Goldberg JH, Kotval XP. Computer interface evaluation using eye movements: methods and constructs. Int J Ind Ergon. 1999;24(6):631–645.

Greenwald AG, Nosek BA, Banaji MR. Understanding and using the implicit association test: I. An improved scoring algorithm. J Perso Soc Psychol. 2003 Aug;85(2):197.

Gugerty L, Brooks J. Reference-frame misalignment and cardinal direction judgments: group differences and strategies. J Exp Psychol Appl. 2004;10(2):75.

Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Adv Psychol. 1988;52:139–183.

Hinds PJ, Roberts TL, Jones H. Whose job is it anyway? A study of human–robot interaction in a collaborative task. Hum Comput Interact. 2004;19(1):151–181.

Ishihara S. Tests for color-blindness. Handaya, Tokyo: Hongo Harukicho; 1917.

Jacob R, Karn KS. Eye tracking in human–computer interaction and usability research: ready to deliver the promises. Mind. 2003;2(3):573–605.

Jian J-Y, Bisantz AM, Drury CG. Foundations for an empirically determined scale of trust in automated systems. Int J Cogn Ergon. 2000;4(1):53–71.

Jones DG, Kaber DB. Situation awareness measurement and the situation awareness global assessment technique. In: Stanton N, Hedge H, Brookhuis K, Salas E, editors. Handbook of Human Factors and Ergonomics Methods. Boca Raton (FL): CRC Press; 2004. p. 42.1–42.8.

Kahneman D, Klein G. Conditions for intuitive expertise: a failure to disagree. Am Psychol. 2009;64(6):515.

Kroft P, Wickens CD. Displaying multi-domain graphical databases: an evaluation of scanning clutter, display size, and user activity. Inf Des J. 2002;11(1):44–52.

Lakhmani SG, Chen JYC, Wright JL, Selkowitz AR, Schwartz M. Agent transparency for an autonomous squad member: uncertainty and projected outcomes. Aberdeen Proving Ground (MD): CCDC Army Research Laboratory (US). Forthcoming 2020.

Lathan CE, Tracey M. The effects of operator spatial perception and sensory feedback on human–robot teleoperation performance. Presence: Teleoperators Virtual Environment. 2002;11(4):368–377.

Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors. 2004;46(1):50–80.

Lewis M. Designing for human–agent interaction. AI Magazine. 1998;19(2):67.

Lichacz FM, Cain B, Patel S. Calibration of confidence in situation awareness queries. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 2003 Oct; Sage, CA. Los Angeles (CA): SAGE Publications. 2003;47(1):222–226.

Lohrenz MC, Trafton JG, Beck MR, Gendron ML. A model of clutter for complex, multivariate geospatial displays. Hum Factors. 2009;51(1):90–101.

Lyons JB. Being transparent about transparency: a model for human–robot interaction. Paper presented at the 2013 AAAI Spring Symposium Series; 2013 Mar 25–27; Stanford, CA. Palo Alto (CA): The AAAI Press. p. 48–53.

Lyons JB, Havig PR. Transparency in a human–machine context: interface approaches for fostering shared awareness/intent. In: Shumaker R, Lackey S,

editors. VAMR 2014. Paper presented at the 6th International Conference on Virtual, Augmented, and Mixed Reality Designing and Developing Virtual and Augmented Environments; 2014 June 22–27; Heraklion, Crete, Greece. Basel, Switzerland: Springer, Cham. Lecture Notes in Computer Science, Vol 8525.

McBride SE, Rogers WA, Fisk AD. Understanding human management of automation errors. Theor Issues Ergon Sci. 2014;15(6):545–577.

McGuinness B. Quantitative analysis of situational awareness (QUASA): applying signal detection theory to true/false probes and self-ratings. Bristol (UK): BAE Systems, Advanced Technology Center; 2004.

Mercado JE, Rupp MA, Chen JY, Barnes MJ, Barber D, Procci K. Intelligent agent transparency in human–agent teaming for Multi-UxV management. Hum Factors. 2016;58(3):401–415.

Merritt SM, Heimbaugh H, LaChapell J, Lee D. I trust it, but I don't know why effects of implicit attitudes toward automation on trust in an automated system. Hum Factors. 2013;55(3):520–534.

Military.com. Pentagon project seeks to build autonomous robots. [accessed April 2020]. https://www.military.com/defensetech/2013/06/13/pentagon-launches-pilot-to-build-autonomous-robots.

Muir BM. Trust in automation: part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics. 1994;37:1905–1922.

Parasuraman R, Molloy R, Singh IL. Performance consequences of automation-induced "complacency". Int J Aviat Psycho. 1993;3(1):1–23.

Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. Hum Factors. 1997;39(2):230–253.

Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern Syst Humans. 2000;30(3):286–297.

Parasuraman R, Sheridan TB, Wickens CD. Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J Cogn Eng Decis Mak. 2008;2(2):140–160.

Peavler WS. Pupil size, information overload, and performance differences. Psychophysiology. 1974;11(5):559–566.

Redick TS, Broadway JM, Meier ME, Kuriakose PS, Unsworth N, Kane MJ, Engle RW. Measuring working memory capacity with automated complex span tasks. European J Psychol Assess. 2012;28(3):164–171.

Salem M, Lakatos G, Amirabdollahian F, Dautenhahn K. Would you trust a (faulty) robot?: effects of error, task type and personality on human–robot cooperation and trust. In: Adams JA, editor. HRI 2015. Proceedings of the Tenth Annual ACM/IEEE International Conference on Human–Robot Interaction; 2015 Mar 2–5; Portland, OR. New York (NY): Association for Computing Machinery. p. 141–148.

Salmon PM, Stanton NA, Walker GH Jenkins D, Ladva D, Rafferty L, Young M. Measuring situation awareness in complex systems: comparison of measures study. Int J Ind Ergon. 2009;39(3):490–500.

Schaefer, KE. Measuring trust in human robot interactions: development of the "Trust Perception Scale–HRI". In: Lawless WF, Mittu R, Wagner A, Sofge D, editors. The Intersection of Robust Intelligence and Trust in Autonomous Systems. Boston (MA): Springer; 2016. p. 191–218.

Selkowitz AR, Lakhmani SG, Larios CN, Chen JY. Agent transparency and the autonomous squad member. Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 2016 Sep 19–23; Washington, DC. 2016;60(1):1319–1323.

Unsworth N, Heitz RP, Schrock JC, Engle RW. An automated version of the operation span task. Behav Res Methods. 2005;37(3):498–505.

Van Orden KF, Limbert W, Makeig S, Jung TP. Eye activity correlates of workload during a visuospatial memory task. Hum Factors. 2001;43(1):111–121.

Wickens CD, Dixon SR. The benefits of imperfect diagnostic automation: a synthesis of the literature. Theor Issues Ergon Sci. 2007;8(3):201–212.

Wiegmann DA. Agreeing with automated diagnostic aids: a study of users' concurrence strategies. Hum Factors. 2002;44(1):44–50.

Wright JL, Chen JY, Quinn SA, Barnes MJ. The effects of level of autonomy on human–agent teaming for multi-robot control and local security maintenance. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2013. Report No.: ARL-TR-6724.

Wright JL, Quinn SA, Chen JY, Barnes MJ. Individual differences in human-agent teaming: an analysis of workload and situation awareness through eye movements. In: Proceedings of the Human Factors and Ergonomics Society

Annual Meeting; 2014 Sep; Sage, CA. Los Angeles (CA): SAGE Publications. 2014;58(1):1410–1414.

Wright JL, Chen JY, Barnes MJ, Hancock PA. Agent reasoning transparency: the influence of information level on automation induced complacency. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2017 Jun 1. Report No.: ARL-TR-8044.

Appendix A. Demographics Questionnaire

This appendix appears in its original form, without editorial change.

Demographic Questionnaire

Date: _____ Participant ID: _____

1. General Information

a. Age: _____ Gender: M F Handedness: L R

b. How long ago did you have an eye exam? Within the last (Circle one):
6 months 1 year 2 years 4 years or more

c. Do you have any of the following (Circle all that apply):
Astigmatism Near-sightedness Far-sightedness Other (explain): _____

d. Do you have corrected vision (Circle one)? Yes No Glasses Contact Lenses
If so, are you wearing them today? Yes No

e. Are you in your good/ comfortable state of health physically? YES NO
If NO, please briefly explain:

f. How many hours of sleep did you get last night? _____ hours

2. Military Experience

a. Do you have prior military service? YES NO If Yes, how long _____

3. Educational Data

a. What is your highest level of education completed? Select one.

____ GED ____ Bachelor's Degree

____ High School ____ M.S/M.A

____ Some College ____ Ph.D.

____ Associates or Technical Degree

What subject is your degree in (for example, Engineering)? _____

4. Computer Experience

a. How long have you been using a computer?
__Less than 1 year ___1-3 years ___4-6 years ___7-10 years ___10 years or more

b. How often do you play computer/video games? (Circle one)
Daily 3-4X/ Week Weekly Monthly Once or twice a year Never

c. Enter the names of the games you play most frequently:
_____

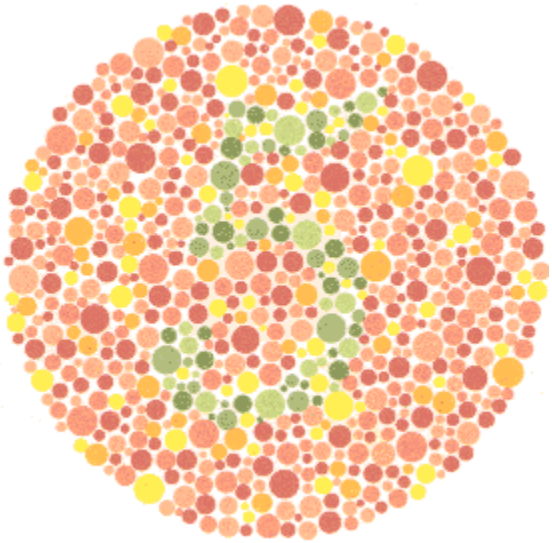
d. How often do you operate a radio-controlled vehicle (car, boat, or plane)?
Daily Weekly Monthly Once or twice a year Never

e. How often do you use graphics/drawing features in software packages?
Daily Weekly Monthly Once or twice a year Never

Appendix B. Ishihara Color Vision Test

Ishihara Color Vision Test

Below is an example of one of the screens that the participant will see during the color vision test: a series of dots that compose the letter 5 among other dots of differing colors.



Appendix C. Implicit Association Test

Implicit Association Test (IAT)

This implicit trust measure was adapted from Merritt et al.'s Implicit Association Test (IAT).[1] The evaluative category (i.e., good/bad) words were adopted from Project Implicit's race IAT (words used: joy, love, peace, wonderful, pleasure, glorious, laughter, happy; agony, terrible, horrible, nasty, evil, awful, failure, hurt). Focus groups identified two strongly related words for the human category (human and person) and the automation category (automation and machine) by consulting a thesaurus and generating synonyms. The more positive an individual's implicit attitude toward automation, the more quickly he or she should be able to complete the task when "automation" and "good" are paired together, and the more difficulty he or she should have when "automation" and "bad" are paired together (Fig. C-1).

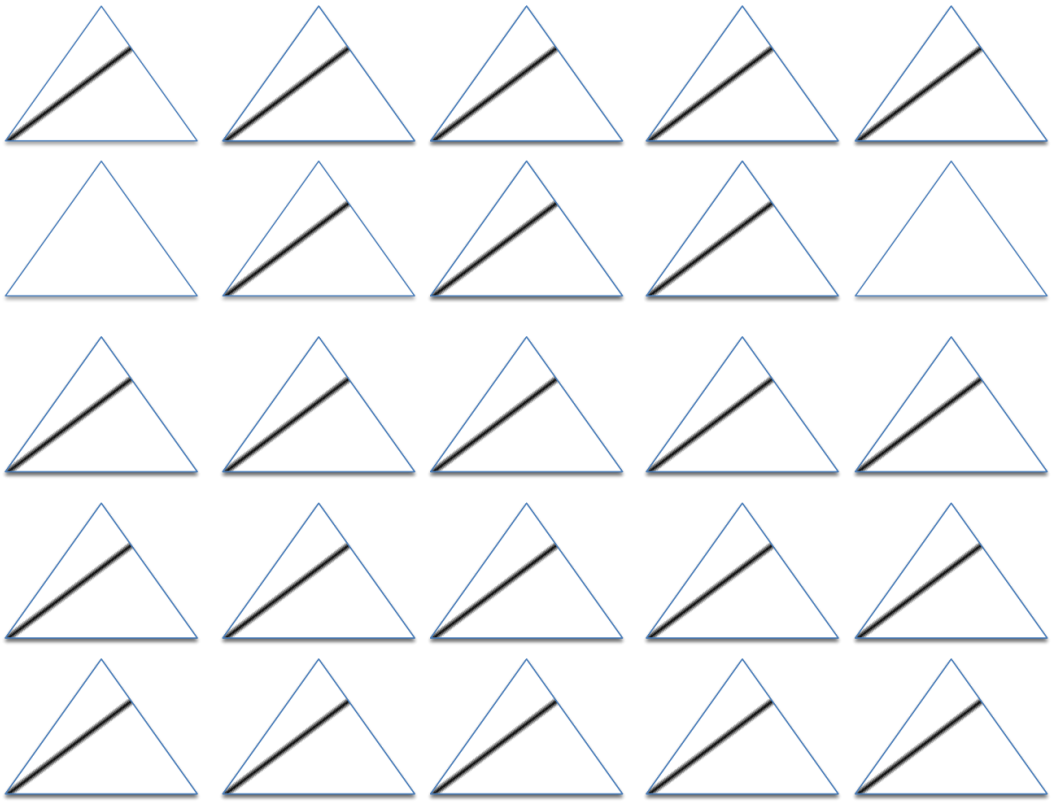
Fig. C-1 Example IAT screen shown to participants

[1] Merritt SM, Heimbaugh H, LaChapell J, Lee D. I trust it, but I don't know why effects of implicit attitudes toward automation on trust in an automated system. Hum Factors. 2013;55(3):520–534.

Appendix D. Matrix Scanning Task

Matrix Scanning Task

Participants will be instructed to say how many triangles do not have stripes through them. The answer will be 1, 2, 3, or 4. This spatial scanning task has been shown to be a good indicator of the person's ability to perform visual scanning.[1] Participant's score will be assessed on reaction time and accuracy.



[1] Barrett GV, Alexander RA, Doverspike D, Cellar D, Thomas JC. The development and application of a computerized information-processing test battery. Appl Meas Educ. 1982;6(1):13–29.

Appendix E. Spatial Orientation Test

The Spatial Orientation Test, modeled after the cardinal direction test developed by Gugerty and his colleagues,[1] is a computerized test consisting of a brief training segment and 32 test questions. The program automatically captures both accuracy and response time. Participants are shown the image in Fig. E-1.

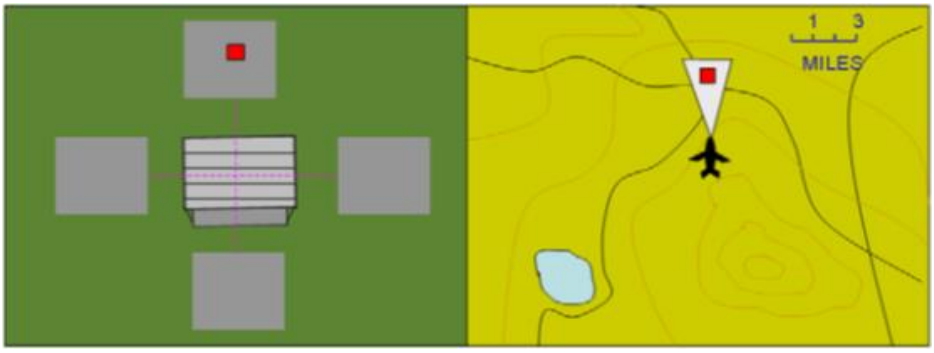


Fig. E-1 Example item from the Spatial Orientation Test

The right-side image is a map showing a plane flying. The left side of the display is the pilot's view (from the cockpit of the plane) of several parking lots surrounding a building. The participants' task is to use the right side of the display to learn which direction the plane is flying. They then use this information to identify which parking lot (north, south, east, or west) in the left-side image has the dot. In the example in Fig. E-1, the plane is heading north and so the dot appears in the north parking lot. In the example shown in Fig. E-2, the plane is heading south and so the dot appears in the east parking lot.

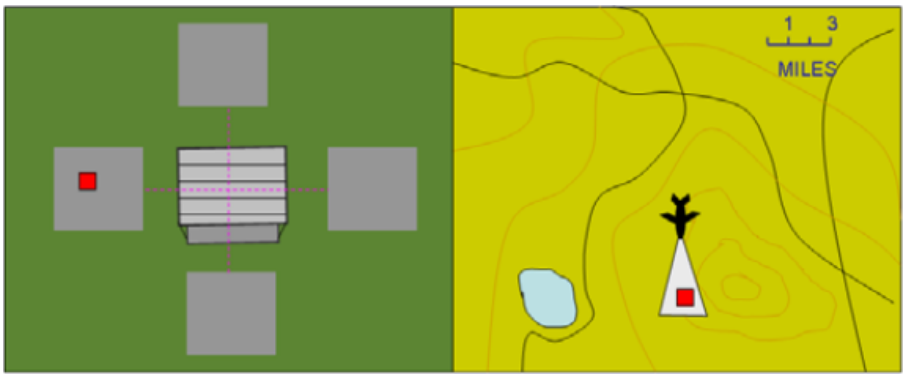


Fig. E-2 Example item from the Spatial Orientation Test

Participants are shown 32 of these images in succession; each time the direction the plane is flying and the location of the dot are randomized. Participants answer by clicking on one of four buttons (north, south, east, or west). This test is self-paced;

[1] Gugerty L, Brooks J. Reference-frame misalignment and cardinal direction judgments: group differences and strategies. J Exp Psych: App. 2004;10(2):75–88.

the participants may take as long as they wish to answer; when they answer one question, the next question automatically appears. No questions can be skipped, and the order of images is randomized among participants.

Appendix F. National Aeronautics and Space Administration–Task Load Index (NASA–TLX)

This appendix appears in its original form, without editorial change.

NASA TLX Workload Assessment

Instructions: Ratings Scales

We are interested in the "workload" you experienced during this scenario. Workload is something experienced individually by each person. One way to find out about workload is to ask people to describe what they experienced. Workload may be caused by many different factors and we would like you to evaluate them individually. The set of six workload rating factors was developed for you to use in evaluating your experiences during different tasks. Please read them. If you have a question about any of the scales in the table, please ask about it. It is extremely important that they be clear to you.

Definitions

Title	Endpoints	Descriptions
MENTAL DEMAND	Low / High	How much mental and perceptual activity was required (that is, thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
PHYSICAL DEMAND	Low / High	How much physical activity was required (that is, pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
TEMPORAL DEMAND	Low / High	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
PERFORMANCE	Poor / Good	How successful do you think you were in accomplishing the goals of the task? How satisfied were you with your performance in accomplishing these goals?
EFFORT	Low / High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
FRUSTRATION LEVEL	Low / High	How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

We want you to evaluate workload. Rate the workload on each factor on a scale. Each scale has two end descriptions, and 20 slots (hash marks) between the end descriptions. Place an "x" in the slot (between the hash marks) that you feel most accurately reflects your workload.

After you have finished the entire series, we will be able to use the pattern of your choices to create a weighted combination of ratings into a summary workload score.

We ask you to evaluate your workload for this scenario. This includes all the duties involved in your job (e.g., detecting targets and using display).

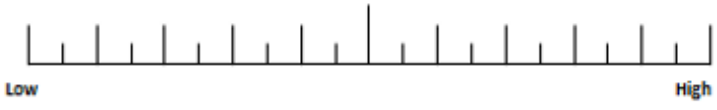
Participant ID: _____

TLX Workload Scale

Please rate your workload by putting a mark on each of the six scales at the point which matches your experience.

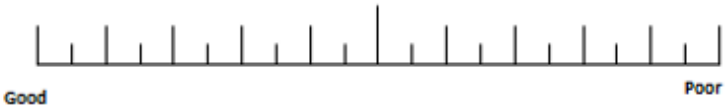
Mental Demand

Physical Demand



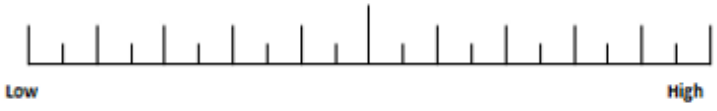
Temporal Demand

Performance



Effort

Frustration



Appendix G. Godspeed Measure[1]

This appendix appears in its original form, without editorial change.

1. Bartneck C, Kulić D, Croft E, Zoghbi S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Social Robotics. 2009;1(1):71–81.

Appendix

GODSPEED I: ANTHROPOMORPHISM

Please rate your impression of the robot on these scales:

Fake	1	2	3	4	5	Natural
Machinelike	1	2	3	4	5	Humanlike
Unconscious	1	2	3	4	5	Conscious
Artificial	1	2	3	4	5	Lifelike
Moving rigidly	1	2	3	4	5	Moving elegantly

GODSPEED II: ANIMACY

Please rate your impression of the robot on these scales:

Dead	1	2	3	4	5	Alive
Stagnant	1	2	3	4	5	Lively
Mechanical	1	2	3	4	5	Organic
Artificial	1	2	3	4	5	Lifelike
Inert	1	2	3	4	5	Interactive
Apathetic	1	2	3	4	5	Responsive

GODSPEED III: LIKEABILITY

Please rate your impression of the robot on these scales:

Dislike	1	2	3	4	5	Like
Unfriendly	1	2	3	4	5	Friendly
Unkind	1	2	3	4	5	Kind
Unpleasant	1	2	3	4	5	Pleasant
Awful	1	2	3	4	5	Nice

GODSPEED IV: PERCEIVED INTELLIGENCE

Please rate your impression of the robot on these scales:

Incompetent	1	2	3	4	5	Competent
Ignorant	1	2	3	4	5	Knowledgeable
Irresponsible	1	2	3	4	5	Responsible
Unintelligent	1	2	3	4	5	Intelligent
Foolish	1	2	3	4	5	Sensible

GODSPEED V: PERCEIVED SAFETY

Please rate your emotional state on these scales:

Anxious	1	2	3	4	5	Relaxed
Agitated	1	2	3	4	5	Calm
Quiescent	1	2	3	4	5	Surprised

Appendix H. Functional Trust Survey

This appendix appears in its original form, without editorial change.

For each of the following items and situations, circle the number which best describes your feeling or your impression based on the system you just used. For each item, consider the following situations:

- A: When the system is collecting and/or highlighting/filtering information.
- B: When the system is integrating information, generating predictive displays, and/or presenting its analysis.
- C: When the system is making decisions and/or selecting actions.
- D: When the system is executing actions.

1. The system is deceptive when…

	not at all			*neutral*			*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

2. The system behaves in an underhanded manner when…

	not at all			*neutral*			*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

3. I am suspicious of the system's intent, action, or outputs when…

	not at all			*neutral*			*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

4. I am wary of the system when…

	not at all			*neutral*		*extremely*	
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

5. The system's actions will have a harmful or injurious outcome when…

	not at all			*neutral*			*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

6. I am confident in the system when…

	not at all			*neutral*			*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

7. The system provides security when…

	not at all			*neutral*			*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

8. The system has integrity when…

	not at all			*neutral*			*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

9. The system is dependable when…

	not at all			*neutral*			*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

10. The system is reliable when…

	not at all			*neutral*			*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

11. I can trust the system when…

	not at all						*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

12. I am familiar with the system when…

	not at all						*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

13. The system is predictable when...

	not at all						*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

14. The system meets the needs of the mission when...

	not at all						*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

15. The system provides appropriate information when...

	not at all						*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

16. The system malfunctions when...

	not at all						*extremely*
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

Now imagine that you are employed as an unmanned vehicle operator to complete missions. Reflecting on the experience with the system you just used, please rate the extent to which you agree with each of these items by circling a value from **1** (**strongly disagree**) to **7** (**strongly agree**), where **4** is **neutral**.

	Strongly Disagree			**Neutral**			**Strongly Agree**
17. Using the system would improve my job performance.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
18. Using the system would make it easier to do my job.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
19. I would find the system useful in my job.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
20. Learning to operate the system is easy for me.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
21. It is easy for me to become skillful at using the system.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
22. I find the system easy to use.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
23. I intend to use this system for my job.	**1**	**2**	**3**	**4**	**5**	**6**	**7**

Appendix I. Reading Span Task (RSPAN)

Participants will be administered a computerized version of the reading span (RSPAN) task[1,2] to evaluate their working memory capacity as well as remove participants with potential reading-comprehension issues.

RSPAN Instructions for Automated Presentation

The experiment is separated into two sections: 1) participants receive practice, and 2) the participants perform the actual experiment. The practice sessions are further broken down into three sections.

The first practice is simple letter span. Participants see letters appear on the screen one at a time and then must recall these letters in the same order they saw them. In all experimental levels, letters remain on the screen for 800 ms. Recall consists of filling in boxes with the appropriate letters. Entering a letter or space in a box should advance the cursor to the next box. At the final box, hitting the spacebar will advance to the next slide. After each recall slide, the computer provides feedback about the number of letters correctly recalled.

Next, participants practice the sentence portion of the experiment. Participants first see a sentence (e.g., "Andy was stopped by the policeman because he crossed the yellow heaven."). Once the participant has read the sentence, they are required to answer YES or NO (did the sentence make sense?). After each sentence's sense verification, participants are given feedback. The reading practice serves to familiarize participants with the sentence portion of the experiment as well as calculate how long it takes a given person to solve the sentence problems. Thus, it attempts to account for individual differences in the time it takes to solve reading problems. After the reading practice, the program calculates the individual's mean time required to solve the problems. This time (plus 2.5 standard deviations [SDs]) is then used as a time limit for the reading portion of the experimental session.

The final practice session has participants perform both the letter recall and reading portions together, just as they will do in the experimental block. As with traditional RSPAN, participants first see the sentence and after verifying whether it makes sense or not, they see the letter to be recalled. If participants take more time to verify the sentence than their average time plus 2.5 SDs, the program automatically moves on. This serves to prevent participants from rehearsing the letters when they

[1] Unsworth N, Heitz RP, Schrock JC, Engle RW. An automated version of the operation span task. Behav Res Meth. 2005;37:498–505.

[2] Daneman M, Carpenter PA. Individual differences in working memory and reading. J Verbal Learning Verbal Behav. 1980; 19(4):450–466.

should be verifying the sense of the sentences. After the participant completes all of the practice sessions, the program moves them to the real trials.

The experimental trials consist of three trials of each set size with the set sizes ranging from 3 to 6. This makes for a total of 54 letters and 54 sentence problems. Subjects are instructed to keep their reading accuracy at or above 80% at all times. During recall, a percentage in red is presented in the upper right-hand corner. Subjects are instructed to keep a careful watch on the percentage to keep it above 80%. Subjects get feedback at the end of each trial. Subjects who do not finish with a reading accuracy score of 80% or better will be excused from continuing with the study.

RSPAN Timing (*may be adjusted after review)*

Sentence-verification screen: Min = none, Max = mean of practice trials +2.5 SD.

Letter presentation: 800 ms.

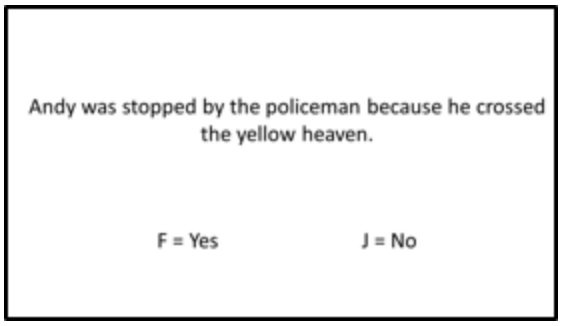
Recall screen: Min = none, Max = 2 min (there is a "Continue" button to move forward faster).

READY screen: 3 s (no keys active, cannot skip this screen).

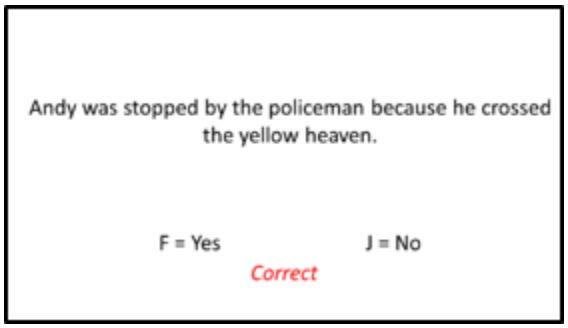
Slide Examples

Ready screen

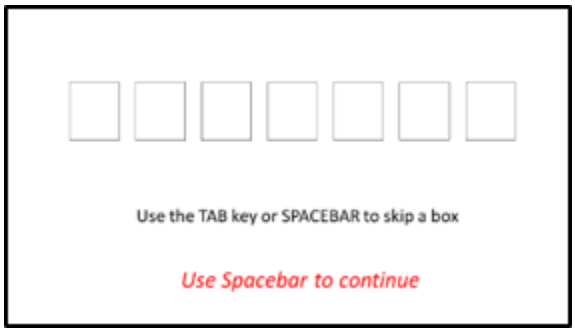
Letter screen



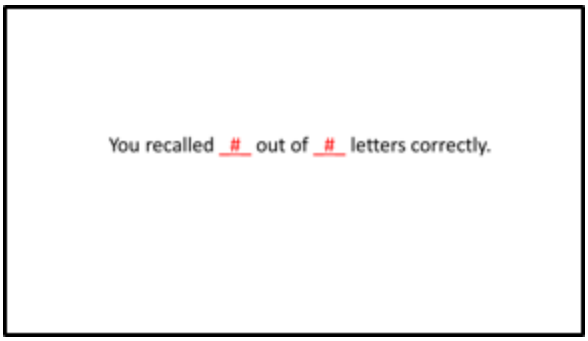
Sentence screen



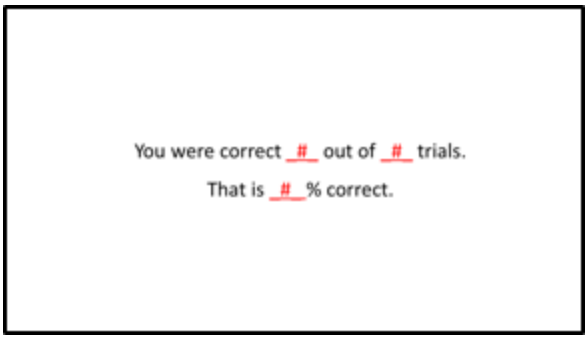
Sentence screen with feedback (for sentence practice only)



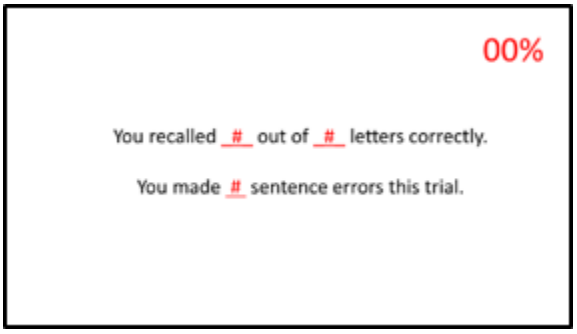
Recall screen; always 7 boxes shown



Feedback screen, letter practice



Feedback screen, sentence practice



Feedback screen, final practice and main experiment

Appendix J. Attentional Control Survey

This appendix appears in its original form, without editorial change.

Attentional Control Survey **Participant #_____** **Date _____**

For each of the following questions, circle the response that best describes you.

It is very hard for me to concentrate on a difficult task when there are noises around.
Almost never, Sometimes, Often, Always

When I need to concentrate and solve a problem, I have trouble focusing my attention.
Almost never, Sometimes, Often, Always

When I am working hard on something, I still get distracted by events around me.
Almost never, Sometimes, Often, Always

My concentration is good even if there is music in the room around me.
Almost never, Sometimes, Often, Always

When concentrating, I can focus my attention so that I become unaware of what's going on in the room around me.
Almost never, Sometimes, Often, Always

When I am reading or studying, I am easily distracted if there are people talking in the same room.
Almost never, Sometimes, Often, Always

When trying to focus my attention on something, I have difficulty blocking out distracting thoughts.
Almost never, Sometimes, Often, Always

I have a hard time concentrating when I'm excited about something.
Almost never, Sometimes, Often, Always

When concentrating, I ignore feelings of hunger or thirst. Almost never, Sometimes, Often, Always

I can quickly switch from one task to another. Almost never, Sometimes, Often, Always

It takes me a while to get really involved in a new task. Almost never, Sometimes, Often, Always

It is difficult for me to coordinate my attention between the listening and writing required when taking notes during lectures. Almost never, Sometimes, Often, Always

I can become interested in a new topic very quickly when I need to.
Almost never, Sometimes, Often, Always

It is easy for me to read or write while I'm also talking on the phone.
Almost never, Sometimes, Often, Always

I have trouble carrying on two conversations at once. Almost never, Sometimes, Often, Always

I have a hard time coming up with new ideas quickly. Almost never, Sometimes, Often, Always

After being interrupted or distracted, I can easily shift my attention back to what I was doing before.
Almost never, Sometimes, Often, Always

When a distracting thought comes to mind, it is easy for me to shift my attention away from it.
Almost never, Sometimes, Often, Always

It is easy for me to alternate between two different tasks. Almost never, Sometimes, Often, Always

It is hard for me to break from one way of thinking about something and look at it from another point of view.
Almost never, Sometimes, Often, Always

Appendix K. Situation Awareness Questions

After each event, the simulation paused, the screens would blank, and the following questions were presented to the participants. Confidence ratings are not shown, but were included with every question.

SCREEN 1

What did the squad encounter?

A. Single shooter	C. Obstruction	E. IED
B. Ambush	D. Flooding	F. Civilians

What action did the Autonomous Squad Member (ASM) take?

A. Duck and cover	C. Avoid	E. Follow
B. Reroute	D. Monitor	F. Smoke grenade

What resource did the ASM predict would be spent/lost?

A. Energy/fuel	C. Time
B. Mechanical	D. Signal strength

SCREEN 2

Did the ASM make an error?

A. Yes
B. No

What was the correct action for the ASM to take?

A. Duck and cover	C. Avoid	E. Follow
B. Reroute	D. Monitor	F. Smoke grenade

What resource should the ASM have tried to conserve?

A. Energy/fuel	C. Time
B. Mechanical	D. Signal strength

SCREEN 3

Based on this encounter, how reliable will the ASM be in the next encounter?

A. Reliable	C. Somewhat unreliable
B. Somewhat reliable	D. Unreliable

Based on this encounter, how accurate will the ASM's assessments be in the next encounter?

A. Accurate	C. Somewhat inaccurate
B. Somewhat accurate	D. Inaccurate

Appendix L. Task Performance Results Tables

Table L-1 Individual difference (ID) factor correlations with threat detection correctness efficiency (TDCE) scores. Correlations are reported using Pearson's *r*. A Mann-Whitney *U* test was used to evaluate potential differences in threat identification due to gaming experience (GE).

Condition	PAC		VS		SOT		WMC		IAT		GE		
	r(54)	*p*	*r(54)*	*p*	*r(54)*	*p*	*r(54)*	*p*	*r(54)*	*p*	*U*	*Z*	*p*
SR	0.109	0.426	0.087	0.525	–0.270[a]	0.044	0.106	0.436	0.011	0.938	353.50	–0.43	0.665
SU	0.116	0.396	0.117	0.389	–0.280[a]	0.037	0.156	0.251	–0.167	0.219	349.00	–0.51	0.611
DR	–0.071	0.605	0.174	0.200	–0.109	0.422	0.012	0.932	0.027	0.844	331.50	–0.80	0.424
DU	0.163	0.231	–0.005	0.971	0.008	0.952	0.128	0.347	0.069	0.613	265.50	–1.90	0.057

[a] Correlation is significant at the 0.05 level (2-tailed).

Note: PAC = perceived attentional control; VS = visual scanning; SOT = spatial orientation test; WMC = working memory capacity; IAT = implicit association test; GE = gaming experience.

Table L-2 Regression analysis for spatial orientation (SO) scores on TDCE scores by experimental condition

ID factor	Condition	R	R^2	*Adj* R^2	B	t(54)	*p*
SO	SR	0.27	0.07	0.06	–0.01	–2.06	0.044
	SU	0.28	0.08	0.06	–0.01	–2.14	0.037
	DR	0.11	0.01	–0.01	0.00	–0.81	0.422
	DU	0.01	0.00	–0.02	0.00	0.06	0.952

Table L-3 Descriptive statistics for TDCE scores by experimental condition (N = 56)

Condition	*M*	*SD*	*SE*
SR	0.667	0.124	0.017
SU	0.657	0.125	0.017
DR	0.659	0.134	0.018
DU	0.662	0.143	0.019

Table L-4 Between-condition comparisons of TDCE scores

Planned comparison	Δ*M*	*SE*	*p*	d_s	*% difference*
SR ≈ SU	0.010	0.018	0.583	0.08	1.47
SR ≈ DR	0.007	0.017	0.667	0.06	1.12
SU ≈ DU	–0.005	0.018	0.768	–0.04	–0.81
DR ≈ DU	–0.003	0.016	0.853	–0.02	–0.46

Table L-5 ID Factor correlations with response time (RT) and event task (ET) scores. Correlations are reported using Pearson's *r*. A Mann–Whitney *U* test was used to evaluate potential differences in threat identification due to GE.

Measure	Condition	PAC		VS		SOT		WMC		IAT		GE		
		r(54)	*p*	*r(54)*	*p*	*r(54)*	*p*	*r(54)*	*p*	*r(54)*	*p*	*U*	*Z*	*p*
RT	SR	0.006	0.968	0.009	0.947	0.244	0.070	–0.254	0.059	–0.160	0.238	315.00	–1.07	0.283
	SU	–0.040	0.770	–0.175	0.198	–0.177	0.191	–0.014	0.917	–0.019	0.887	304.00	–1.26	0.209
	DR	–0.020	0.882	0.023	0.864	0.003	0.982	–0.081	0.553	–0.133	0.327	366.00	–0.22	0.822
	DU	–0.234	0.083	–0.076	0.576	0.063	0.645	–0.078	0.570	0.039	0.773	264.00	–1.92	0.054
ET	SR	0.025	0.853	–0.048	0.725	0.276[a]	0.040	–0.269[a]	0.045	–0.170	0.211	348.00	–0.52	0.600
	SU	–0.030	0.825	–0.175	0.197	–0.207	0.125	–0.060	0.663	–0.085	0.533	325.00	–0.91	0.364
	DR	–0.019	0.892	–0.112	0.412	–0.060	0.659	–0.088	0.521	–0.143	0.292	323.00	–0.94	0.347
	DU	–0.173	0.203	–0.126	0.353	0.012	0.932	–0.120	0.379	0.047	0.731	292.00	–1.46	0.145

[a] Correlation is significant at the 0.05 level (2-tailed).

Table L-6 Descriptive statistics for RT and ET scores by experimental condition (N = 56)

Measure	Condition	*M*	*SD*	*SE*
RT	SR	26.50	4.86	0.65
	SU	26.75	4.36	0.58
	DR	25.65	3.75	0.50
	DU	25.93	5.15	0.69
ET	SR	26.61	4.59	0.61
	SU	27.27	4.94	0.66
	DR	26.06	3.81	0.51
	DU	26.69	5.96	0.80

Table L-7 Between-condition comparisons of RT and ET scores

Measure	Planned comparison	Δ*M*	*SE*	*p*	d_s	*% difference*
RT	SR ≈ SU	–0.26	0.71	0.718	–0.06	–0.98
	SR ≈ DR	0.84	0.70	0.237	0.19	3.18
	SU ≈ DU	0.82	0.69	0.238	0.17	3.07
	DR ≈ DU	–0.28	0.78	0.721	–0.06	–1.10
ET	SR ≈ SU	–0.67	0.74	0.373	–0.14	–2.50
	SR ≈ DR	0.55	0.67	0.412	0.13	2.07
	SU ≈ DU	0.58	0.82	0.479	0.11	2.14
	DR ≈ DU	–0.63	0.85	0.462	–0.13	–2.42

Appendix M. National Aeronautics and Space Administration–Task Load Index (NASA–TLX) Results Tables

Table M-1 ID Factor correlations with National Aeronautics and Space Administration–Task Load Index (NASA–TLX) scores. Correlations are reported using Pearson's *r*. A Mann–Whitney *U* test was used to evaluate potential differences in perceived workload due to gaming experience.

Measure	Condition	PAC		VS		SOT		WMC		IAT		GE		
		r(54)	*p*	*r(54)*	*p*	*r(54)*	*p*	*r(54)*	*p*	*r(54)*	*p*	*U*	*Z*	*p*
Global	SR	–0.05	0.696	0.09	0.530	–0.17	0.213	–0.08	0.566	0.03	0.816	322.50	–0.95	0.342
	SU	–0.19	0.162	–0.05	0.739	0.17	0.200	0.02	0.881	0.03	0.798	210.00	–2.82	0.005
	DR	0.03	0.801	–0.15	0.272	0.10	0.462	–0.15	0.258	–0.09	0.531	379.50	0.00	1.000
	DU	0.12	0.359	0.07	0.612	–0.14	0.292	0.11	0.428	–0.04	0.751	361.50	–0.30	0.764
Condition	**Subscale**	**PAC**		**VS**		**SOT**		**WMC**		**IAT**		**GE**		
		r(54)	*p*	*r(54)*	*p*	*r(54)*	*p*	*r(54)*	*p*	*r(54)*	*p*	*U*	*Z*	*p*
SR	MD	0.13	0.354	0.11	0.429	–0.21	0.116	–0.01	0.957	–0.09	0.493	358.00	–0.36	0.720
	PD	0.08	0.536	0.01	0.963	0.07	0.587	–0.04	0.743	–0.14	0.300	371.50	–0.16	0.872
	TD	0.10	0.453	0.23	0.087	0.00	0.974	0.03	0.828	0.01	0.931	297.00	–1.38	0.169
	Perf	–0.39[a]	0.003	–0.24	0.077	–0.32[b]	0.016	–0.09	0.502	0.28[b]	0.035	351.00	–0.48	0.635
	Effort	0.04	0.785	0.26	0.055	–0.11	0.436	–0.341[b]	0.010	–0.05	0.711	339.50	–0.67	0.505
	Frust	–0.08	0.563	0.00	0.972	–0.09	0.523	0.18	0.186	0.06	0.635	333.00	–0.81	0.417
SU	MD	–0.12	0.372	–0.19	0.170	0.16	0.227	–0.02	0.894	0.02	0.863	278.50	–1.68	0.092
	PD	0.04	0.778	0.16	0.232	–0.09	0.505	–0.08	0.544	–0.15	0.270	359.50	–0.43	0.669
	TD	–0.22	0.104	–0.12	0.386	0.28[b]	0.036	–0.08	0.581	0.04	0.768	207.50	–2.87	0.004
	Perf	–0.12	0.364	–0.16	0.228	0.03	0.854	0.05	0.726	–0.04	0.743	201.50	–2.97	0.003
	Effort	–0.06	0.681	0.05	0.740	0.11	0.424	0.21	0.122	0.01	0.965	295.00	–1.41	0.159
	Frust	–0.03	0.817	0.08	0.534	0.06	0.664	–0.06	0.679	0.30	0.024	362.50	–0.30	0.768
DR	MD	0.07	0.594	–0.07	0.591	0.13	0.343	0.05	0.732	–0.11	0.428	330.50	–0.82	0.414
	PD	0.18	0.189	0.09	0.500	0.12	0.386	0.07	0.606	0.04	0.791	346.50	–0.62	0.532
	TD	–0.10	0.478	–0.05	0.729	0.10	0.471	–0.01	0.929	–0.15	0.280	367.50	–0.20	0.841
	Perf	0.01	0.944	–0.21	0.122	–0.16	0.235	–0.15	0.285	–0.08	0.561	379.00	–0.01	0.993
	Effort	0.19	0.157	–0.07	0.605	0.12	0.372	–0.25	0.063	0.04	0.784	349.50	–0.50	0.617
	Frust	–0.20	0.148	–0.310[b]	0.020	0.22	0.097	–0.11	0.423	0.02	0.884	341.00	–0.67	0.501
DU	MD	–0.09	0.492	0.10	0.472	–0.04	0.764	0.02	0.900	–0.13	0.346	293.50	–1.43	0.152
	PD	–0.04	0.771	0.06	0.678	–0.28[b]	0.039	–0.04	0.784	0.08	0.578	290.00	–1.83	0.067
	TD	0.10	0.480	0.07	0.607	–0.17	0.215	0.04	0.766	0.00	0.998	347.00	–0.54	0.588
	Perf	0.10	0.478	0.03	0.855	0.04	0.751	0.04	0.765	–0.11	0.399	348.00	–0.53	0.599
	Effort	0.10	0.466	0.00	0.994	–0.03	0.849	0.14	0.288	0.02	0.891	371.50	–0.13	0.894
	Frust	0.08	0.554	0.08	0.555	0.01	0.958	0.02	0.874	–0.10	0.474	321.00	–1.01	0.315

[a] Correlation is significant at the 0.01 level (2-tailed).

[b] Correlation is significant at the 0.05 level (2-tailed).

Note: MD = mental demand; PD = pupil diameter; TD = temporal demand; Perf = performance; frust = frustration; PAC = perceived attentional control; VS = visual scanning; SOT = spatial orientation test; WMC = working memory capacity; IAT = implicit association test; GE = gaming experience.

Table M-2 Repeated-measures analysis of variance (ANOVA) results for participant NASA–TLX scores across conditions

NASA–TLX scale	df_1	df_2	F	p	ω^2
Global workload	3	165	0.57	0.635	0.00
Mental demand (MD)	3	165	0.23	0.877	0.00
Effort	3	165	0.08	0.970	0.00
Performance (Perf)	3	165	1.18	0.319	0.00
Temporal demand (TD)	3	165	0.83	0.478	0.00

Table M-3 Descriptive statistics for NASA–TLX and subscale scores (N = 56) by experimental condition

Measure	Condition	*M*	*SD*	*SE*
Global workload	SR	39.87	2.22	16.62
	SU	36.09	2.18	16.28
	DR	38.38	2.17	16.25
	DU	37.28	2.03	15.21
MD	SR	17.01	1.15	8.58
	SU	15.80	1.16	8.67
	DR	16.58	1.16	8.70
	DU	16.88	1.18	8.84
Effort	SR	10.58	0.82	6.12
	SU	10.39	0.88	6.59
	DR	10.86	0.81	6.06
	DU	10.94	0.97	7.25
Perf	SR	7.93	0.85	6.34
	SU	6.76	0.78	5.85
	DR	6.85	0.73	5.44
	DU	5.99	0.69	5.17
TD	SR	7.17	0.67	4.99
	SU	5.79	0.58	4.30
	DR	6.92	0.66	4.92
	DU	6.54	0.68	5.10

Table M-4 Between-condition comparisons of NASA–TLX scores

Measure	Paired comparison	Δ*M*	SE	*p*	d_s	*% difference*
Global workload	SR – SU	3.78	3.08	0.225	0.23	9.48
	SR – DR	1.49	3.05	0.627	0.09	3.73
	SU – DU	–1.19	2.85	0.678	–0.08	–3.30
	DR – DU	1.10	3.00	0.715	0.07	2.87
MD	SR – SU	1.21	1.42	0.399	0.14	7.10
	SR – DR	0.43	1.54	0.782	0.05	2.52
	SU – DU	–1.07	1.57	0.497	–0.12	–6.78
	DR – DU	–0.29	1.71	0.865	–0.03	–1.76
Effort	SR – SU	0.20	1.27	0.878	0.03	1.86
	SR – DR	–0.28	1.20	0.816	–0.05	–2.64
	SU – DU	–0.55	1.27	0.664	–0.08	–5.33
	DR – DU	–0.08	1.23	0.950	–0.01	–0.71
Perf	SR – SU	1.17	1.17	0.319	0.19	14.78
	SR – DR	1.08	1.01	0.287	0.18	13.65
	SU – DU	0.77	0.95	0.418	0.14	11.45
	DR – DU	0.86	0.98	0.381	0.16	12.61
TD	SR – SU	1.38	0.88	0.121	0.30	19.26
	SR – DR	0.25	0.98	0.800	0.05	3.49
	SU – DU	–0.76	0.93	0.420	–0.16	–13.06
	DR – DU	0.38	0.97	0.702	0.07	5.42

Table M-5 Repeated-measures ANOVA results for participant ocular indices across conditions

Measure	df_1	df_2	*F*	*p*	ω^2
Fixation count	3	150	1.55	0.205	0.01
Fixation duration (s)	3	150	0.85	0.468	0.00
Pupil diameter	3	150	0.60	0.614	0.00

Table M-6 Descriptive statistics for participant ocular metrics, by experimental condition

Measure	Condition	*M*	*SD*	*SE*
Fixation count	SR	1376.76	193.34	27.07
	SU	1416.98	204.80	28.68
	DR	1381.69	183.10	25.64
	DU	1412.61	201.93	28.28
Fixation duration (s)	SR	0.55	0.10	0.01
	SU	0.53	0.09	0.01
	DR	0.54	0.11	0.02
	DU	0.53	0.11	0.01
Pupil diameter	SR	3.22	0.50	0.07
	SU	3.19	0.49	0.07
	DR	3.20	0.48	0.07
	DU	3.19	0.52	0.07

Table M-7 Between-condition comparisons of participant ocular indices

Measure	Paired comparison	Δ*M*	SE	*p*	d_s	*% difference*
Fixation count	SR ≈ SU	–40.22	25.65	0.123	–0.20	–2.92
	SR ≈ DR	–4.92	21.88	0.823	–0.03	–0.36
	SU ≈ DU	4.37	23.82	0.855	0.02	0.31
	DR ≈ DU	–30.92	20.80	0.143	–0.16	–2.24
Fixation duration (s)	SR ≈ SU	0.014	0.010	0.173	0.14	2.52
	SR ≈ DR	0.004	0.009	0.632	0.04	0.81
	SU ≈ DU	–0.001	0.011	0.925	–0.01	–0.19
	DR ≈ DU	0.008	0.011	0.442	0.08	1.54
Pupil diameter	SR ≈ SU	0.027	0.021	0.208	0.05	0.83
	SR ≈ DR	0.014	0.027	0.605	0.03	0.44
	SU ≈ DU	0.000	0.022	0.999	0.00	0.00
	DR ≈ DU	0.013	0.021	0.552	0.03	0.40

Table M-8 Descriptive statistics for participant fixation count in the logic factors area of interest (AOI), by experimental condition

Measure	Condition	*M*	*SD*	*SE*
Logic Factor AOI Fixation Count	SR	38.26	31.84	4.46
	SU	37.11	27.34	3.83
	DR	56.86	45.60	6.39
	DU	57.77	43.52	6.09

Table M-9 Between-condition comparisons of participant fixation count in the ASM logic factor AOI

Measure	Paired comparison	Δ*M*	SE	*p*	d_s	*% difference*
Logic factor AOI fixation Count	SR ≈ SU	1.15	5.34	0.830	0.04	3.01
	SR < DR	–18.60	5.71	0.002	–0.47	–48.61
	SU < DU	–20.66	6.33	0.002	–0.57	–55.66
	DR ≈ DU	–0.91	6.46	0.889	–0.02	–1.59

Appendix N. Situation Awareness Results Tables

Table N-1 Individual difference (ID) factor correlations with SA1 and SA2 scores. Correlations are reported using Pearson's *r*. A Mann–Whitney *U* test was used to evaluate potential differences in situation awareness (SA) due to gaming experience (GE).

Measure	Condition	PAC		VS		SOT		WMC		IAT		GE		
		r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*U*	*Z*	*p*
SA1 % correct	SR	0.13	0.337	0.03	0.813	–0.05	0.690	0.33[a]	0.012	–0.01	0.957	369.00	–0.18	0.856
	SU	–0.32[a]	0.018	0.18	0.196	–0.28[a]	0.035	–0.06	0.636	0.00	0.988	360.50	–0.33	0.743
	DR	0.13	0.335	–0.02	0.873	0.11	0.435	0.27[a]	0.044	0.00	0.988	338.00	–0.70	0.483
	DU	–0.03	0.803	–0.11	0.426	–0.06	0.680	0.07	0.585	0.08	0.582	349.50	–0.51	0.608
SA1 confidence	SR	0.24	0.075	–0.06	0.661	–0.07	0.604	–0.16	0.236	0.01	0.915	349.00	–0.51	0.610
	SU	0.05	0.732	0.19	0.161	0.03	0.852	–0.10	0.476	–0.14	0.310	315.50	–1.07	0.284
	DR	0.13	0.352	0.06	0.646	–0.05	0.736	0.08	0.563	–0.20	0.135	375.50	–0.07	0.947
	DU	0.25	0.064	–0.09	0.510	–0.01	0.932	–0.04	0.791	0.04	0.796	367.50	–0.20	0.841
SA2 %correct	SR	0.03	0.837	0.12	0.375	–0.04	0.792	0.07	0.620	–0.03	0.824	335.00	–0.76	0.446
	SU	–0.22	0.107	–0.12	0.377	–0.03	0.826	0.10	0.448	0.06	0.682	317.50	–1.07	0.286
	DR	0.16	0.238	0.02	0.909	–0.12	0.376	0.29[a]	0.033	–0.01	0.943	303.50	–1.29	0.196
	DU	0.08	0.543	–0.05	0.701	–0.04	0.749	0.13	0.331	0.04	0.755	341.50	–0.65	0.518
SA2 confidence	SR	0.27[a]	0.043	–0.06	0.642	0.11	0.407	–0.08	0.538	0.03	0.798	272.50	–1.79	0.074
	SU	0.11	0.437	0.05	0.724	0.23	0.094	–0.12	0.397	–0.21	0.120	317.00	–1.05	0.296
	DR	0.00	0.973	0.01	0.929	–0.05	0.726	0.21	0.121	–0.04	0.763	316.50	–1.05	0.292
	DU	0.23	0.091	–0.15	0.255	0.20	0.135	–0.18	0.182	0.10	0.458	361.50	–0.30	0.764

[a] Correlation is significant at the 0.05 level (2-tailed).

Note: SA = situation awareness; PAC = perceived attentional control; VS = visual scanning; SOT = spatial orientation test; WMC = working memory capacity; IAT = implicit association test; GE = gaming experience.

Table N-2 Descriptive statistics for SA1 and SA2 measures, by experimental condition (N = 56)

Measure		Condition	*M*	*SD*	*SE*
SA1	% correct	SR	0.845	0.073	0.010
		SU	0.841	0.072	0.009
		DR	0.831	0.096	0.013
		DU	0.833	0.088	0.012
	Confidence	SR	4.629	0.249	0.033
		SU	4.634	0.223	0.030
		DR	4.627	0.238	0.032
		DU	4.575	0.263	0.035
SA2	% correct	SR	0.866	0.084	0.011
		SU	0.873	0.094	0.012
		DR	0.875	0.098	0.013
		DU	0.852	0.092	0.012
	Confidence	SR	4.647	0.284	0.038
		SU	4.651	0.276	0.037
		DR	4.616	0.316	0.042
		DU	4.598	0.314	0.042

Table N-3 Between-condition comparisons of SA1 and SA2 results

Measure		Paired comparison	Δ*M*	*SE*	*p*	d_s	*% difference*
SA1	% correct	SR ≈ SU	0.004	0.011	0.708	0.06	0.50
		SR ≈ DR	0.014	0.014	0.326	0.16	1.60
		SU ≈ DU	0.007	0.015	0.623	0.09	0.86
		DR ≈ DU	–0.002	0.017	0.902	–0.02	–0.25
	Confidence	SR ≈ SU	–0.005	0.032	0.875	–0.02	–0.11
		SR ≈ DR	0.002	0.031	0.951	0.01	0.04
		SU ≈ DU	0.059	0.037	0.123	0.24	1.26
		DR ≈ DU	0.052	0.038	0.185	0.21	1.11
SA2	% correct	SR ≈ SU	–0.007	0.012	0.587	–0.07	–0.77
		SR ≈ DR	–0.009	0.013	0.486	–0.10	–1.01
		SU ≈ DU	0.021	0.012	0.094	0.22	2.39
		DR ≈ DU	0.023	0.011	0.049	0.24	2.63
	Confidence	SR ≈ SU	–0.004	0.036	0.915	–0.01	–0.08
		SR ≈ DR	0.031	0.042	0.472	0.10	0.66
		SU ≈ DU	0.053	0.035	0.142	0.18	1.14
		DR ≈ DU	0.018	0.047	0.697	0.06	0.40

Table N-4 ID factor correlations with SA3 (projected reliability and projected accuracy) scores. Correlations are reported using Pearson's *r*. A Mann–Whitney *U* test was used to evaluate potential differences in SA due to GE.

Measure	Condition	PAC		VS		SOT		WMC		IAT		GE		
		r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*U*	*Z*	*p*
Projected reliability	SR	0.15	0.262	–0.02	0.908	0.01	0.920	0.09	0.532	0.15	0.271	334.50	–0.84	0.398
	SU	0.03	0.822	–0.11	0.423	–0.12	0.360	–0.14	0.302	0.14	0.316	324.50	–0.93	0.351
	DR	0.04	0.773	0.20	0.134	–0.06	0.647	0.26[a]	0.049	0.12	0.377	338.50	–0.81	0.418
	DU	0.14	0.319	0.00	0.984	–0.01	0.961	0.02	0.908	0.26	0.053	324.00	–0.94	0.349
Projected accuracy	SR	0.20	0.149	0.09	0.500	–0.18	0.185	0.21	0.117	0.15	0.277	345.50	–0.67	0.501
	SU	–0.08	0.556	–0.02	0.899	0.02	0.907	–0.27[a]	0.044	0.09	0.513	367.50	–0.20	0.840
	DR	0.14	0.312	0.25	0.062	–0.03	0.806	0.22	0.104	–0.07	0.606	366.00	–0.26	0.792
	DU	–0.01	0.935	–0.13	0.357	0.17	0.216	–0.11	0.425	0.16	0.232	378.50	–0.02	0.986
Confidence in reliability assessment	SR	0.32[a]	0.016	–0.04	0.776	0.10	0.483	–0.02	0.858	0.08	0.561	353.00	–0.46	0.649
	SU	0.04	0.747	0.09	0.522	0.12	0.396	–0.14	0.313	–0.21	0.113	374.50	–0.08	0.933
	DR	0.14	0.303	0.09	0.497	0.04	0.764	0.17	0.208	–0.02	0.893	291.50	–1.51	0.132
	DU	0.11	0.403	0.00	0.977	0.23	0.082	–0.24	0.078	–0.10	0.473	315.00	–1.09	0.278
Confidence in accuracy assessment	SR	0.32[a]	0.018	–0.01	0.926	0.10	0.457	–0.01	0.944	0.11	0.435	358.00	–0.37	0.711
	SU	0.05	0.716	0.01	0.958	0.12	0.369	–0.15	0.258	–0.22	0.109	345.50	–0.57	0.568
	DR	0.15	0.276	0.12	0.382	0.14	0.318	0.09	0.525	–0.04	0.783	296.50	–1.41	0.158
	DU	0.16	0.226	–0.08	0.557	0.25	0.067	–0.20	0.135	–0.07	0.628	352.00	–0.46	0.645

[a] Correlation is significant at the 0.05 level (2-tailed).

Table N-5 Descriptive statistics for SA3 measures, by experimental condition (N = 56)

Measure		Condition	*M*	*SD*	*SE*
Reliability	Score	SR	3.869	0.189	0.025
		SU	3.075	0.416	0.056
		DR	3.845	0.271	0.036
		DU	3.081	0.407	0.054
	Confidence	SR	4.604	0.468	0.063
		SU	4.324	0.585	0.078
		DR	4.622	0.401	0.054
		DU	4.369	0.548	0.073
Accuracy	Score	SR	3.860	0.267	0.036
		SU	3.128	0.393	0.052
		DR	3.881	0.184	0.025
		DU	3.095	0.384	0.051
	Confidence	SR	4.639	0.468	0.063
		SU	4.313	0.586	0.078
		DR	4.628	0.379	0.051
		DU	4.336	0.545	0.073

Table N-6 Between-condition comparisons of SA3 results

Measure		Paired comparison	Δ*M*	*SE*	*p*	d_s	*% difference*
Reliability	Score	SR > SU	0.794	0.062	<0.001	2.46	20.53
		SR ≈ DR	0.024	0.037	0.524	0.10	0.61
		SU ≈ DU	–0.006	0.065	0.924	–0.02	–0.20
		DR > DU	0.764	0.062	<0.001	2.21	19.88
	Confidence	SR > SU	0.280	0.068	<0.001	0.53	6.09
		SR ≈ DR	–0.017	0.056	0.755	–0.04	–0.38
		SU ≈ DU	–0.045	0.040	0.264	–0.08	–1.04
		DR > DU	0.253	0.063	<0.001	0.53	5.47
Accuracy	Score	SR > SU	0.732	0.059	<0.001	2.18	18.95
		SR ≈ DR	–0.021	0.030	0.485	–0.09	–0.55
		SU ≈ DU	0.033	0.057	0.567	0.08	1.05
		DR > DU	0.786	0.058	<0.001	2.61	20.24
	Confidence	SR > SU	0.327	0.068	<0.001	0.62	7.04
		SR ≈ DR	0.012	0.053	0.828	0.03	0.25
		SU ≈ DU	–0.023	0.042	0.583	–0.04	–0.54
		DR > DU	0.292	0.059	<0.001	0.62	6.30

Table N-7 Repeated-measures ANOVA results for ASM reliability and accuracy projections by event type

Measure		*F (1,219)*	*p*	ω^2
Reliability	Beginning	0.03	0.872	0.000
	Error	364.53	<0.001	0.094
	First following	11.84	<0.001	0.001
	Second following	6.64	0.011	0.000
Accuracy	Beginning	0.35	0.555	0.000
	Error	375.78	<0.001	0.087
	First following	13.70	<0.001	0.002
	Second following	5.91	0.016	0.000

Table N-8 Descriptive statistics and between-condition comparison results for reliability and accuracy projections, shown by event type

Measure	Event	Condition	*N*	*M*	*SD*	*SE*	Between-condition comparison		
							p	d_s	*% difference*
Reliability	Beginning	NE	112	3.768	0.569	0.052	0.872	–0.02	–0.32
		E	109	3.780	0.533	0.053			
	Error	NE	112	3.875	0.358	0.071	<0.001	2.57	50.04
		E	109	1.936	1.012	0.072			
	First following	NE	112	3.795	0.556	0.061	0.001	0.46	7.89
		E	109	3.495	0.728	0.062			
	Second following	NE	112	3.929	0.259	0.036	0.011	0.35	3.32
		E	109	3.798	0.467	0.036			
Accuracy	Beginning	NE	112	3.839	0.494	0.049	0.555	0.08	1.07
		E	109	3.798	0.541	0.050			
	Error	NE	112	3.929	0.291	0.069	<0.001	2.61	48.39
		E	109	2.028	0.995	0.070			
	First following	NE	112	3.804	0.583	0.060	<0.001	0.50	8.34
		E	109	3.486	0.689	0.061			
	Second following	NE	112	3.902	0.299	0.038	0.016	0.33	3.36
		E	109	3.771	0.484	0.038			

Note: NE = no error witnessed; E = error witnessed; *N* = number; *M* = mean; *SD* = standard deviation; *SE* = standard error of the mean.

Table N-9 Repeated-measures ANOVA results for participant confidence in their reliability and accuracy projections by event type

Measure		*F (1,219)*	*p*	ω^2
Confidence in Rel assessment	Beginning	0.06	0.807	0.000
	Error	10.47	0.001	0.001
	First following	10.47	0.001	0.001
	Second following	10.16	0.002	0.001
Confidence in Acc assessment	Beginning	0.64	0.425	0.000
	Error	15.52	<0.001	0.001
	First following	15.24	<0.001	0.002
	Second following	14.90	<0.001	0.001

Table N-10 Descriptive statistics and between-condition comparison results for participant confidence in their reliability and accuracy projections shown by event type

Measure	Event	Condition	*N*	*M*	*SD*	*SE*	Between-condition comparison		
							p	d_s	*% difference*
Confidence in Rel assessment	Beginning	NE	112	4.482	0.710	0.067	0.807	0.03	0.52
		E	109	4.459	0.714	0.068			
	Error	NE	112	4.625	0.555	0.064	0.001	0.44	6.37
		E	109	4.330	0.782	0.065			
	First following	NE	112	4.527	0.747	0.077	0.001	0.44	7.79
		E	109	4.174	0.870	0.078			
	Second following	NE	112	4.732	0.537	0.060	0.002	0.43	5.78
		E	109	4.459	0.727	0.061			
Confidence in Acc assessment	Beginning	NE	112	4.527	0.735	0.068	0.425	0.11	1.71
		E	109	4.450	0.700	0.069			
	Error	NE	112	4.652	0.581	0.064	<0.001	0.53	7.70
		E	109	4.294	0.761	0.065			
	First following	NE	112	4.563	0.708	0.075	<0.001	0.53	9.11
		E	109	4.147	0.870	0.076			
	Second following	NE	112	4.759	0.541	0.060	<0.001	0.52	6.89
		E	109	4.431	0.712	0.060			

Appendix O. Functional Trust Tables

Table O-1 Individual difference (ID) factor correlations with Jian et al. (2000) trust in automation scores. Correlations are reported using Pearson's *r*. A Mann–Whitney *U* test was used to evaluate potential differences in trust due to gaming experience (GE).

ID factor		Trust A				Trust B				Trust C				Trust D				Overall Trust			
		SR	SU	DR	DU	SR	SU	DR	DU	SR	SU	DR	DU	SR	SU	DR	DU	SR	SU	DR	DU
PAC	*r*	0.326[a]	–0.019	0.301[a]	0.064	0.267[a]	0.002	0.310[a]	0.124	0.401[b]	–0.038	0.238	0.141	0.380[b]	0.026	0.209	0.072	0.378[b]	–0.007	0.271[a]	0.111
	p	0.014	0.892	0.024	0.641	0.047	0.986	0.020	0.361	0.002	0.780	0.078	0.300	0.004	0.847	0.122	0.599	0.004	0.959	0.044	0.414
VS	*r*	0.216	0.149	0.248	0.150	0.298[a]	0.191	0.177	0.066	0.240	0.130	0.191	0.233	0.214	–0.046	0.140	0.084	0.268[a]	0.116	0.194	0.147
	p	0.110	0.272	0.065	0.270	0.026	0.158	0.192	0.631	0.075	0.341	0.159	0.084	0.113	0.735	0.303	0.539	0.046	0.394	0.152	0.279
SOT	*r*	0.082	0.141	0.094	0.111	0.062	0.075	0.055	0.126	0.167	–0.024	0.085	0.102	0.185	–0.068	0.069	0.037	0.136	0.033	0.079	0.105
	p	0.549	0.302	0.492	0.413	0.652	0.582	0.686	0.356	0.220	0.858	0.533	0.454	0.173	0.616	0.616	0.789	0.316	0.809	0.564	0.439
WMC	*r*	–0.119	–0.136	0.280[a]	0.111	–0.148	–0.122	0.300[a]	0.146	0.045	–0.112	0.400[b]	0.114	–0.107	–0.024	0.396[b]	0.156	–0.090	–0.108	0.363[b]	0.148
	p	0.383	0.317	0.037	0.417	0.277	0.370	0.025	0.283	0.740	0.410	0.002	0.404	0.431	0.862	0.003	0.251	0.508	0.426	0.006	0.277
IAT	*r*	–0.128	0.048	–0.194	–0.064	–0.039	0.076	–0.137	0.000	–0.001	–0.012	–0.185	–0.080	–0.041	0.037	–0.194	0.027	–0.056	0.042	–0.185	–0.032
	p	0.348	0.728	0.153	0.641	0.777	0.577	0.313	0.999	0.997	0.928	0.173	0.557	0.762	0.788	0.152	0.844	0.684	0.761	0.171	0.816
GE	*U*	371.5	306.5	331.5	358.0	369.50	376.00	325.50	342.00	368.00	372.50	349.00	357.50	354.50	376.00	359.00	355.50	363.00	376.00	342.00	368.5
	Z	–0.13	–1.22	–0.80	–0.36	–0.17	–0.06	–0.90	–0.63	–0.19	–0.12	–0.51	–0.37	–0.42	–0.06	–0.34	–0.40	–0.27	–0.06	–0.62	–0.18
	p	0.894	0.224	0.423	0.720	0.868	0.954	0.367	0.532	0.848	0.907	0.611	.714	.676	.954	.732	.689	.783	.954	.532	.855

[a] Correlation is significant at the 0.05 level (2-tailed).

[b] Correlation is significant at the 0.01 level (2-tailed).

Note: ID = individual difference; PAC = perceived attentional control; VS = visual scanning; SOT = spatial orientation test; WMC = working memory capacity; IAT = implicit association test; GE = gaming experience.

Table O-2 Regression analysis for PAC and WMC on overall Jian Trust in Automation scores by experimental condition

ID factor	Condition	*R*	R^2	*Adj* R^2	*B*	*t(54)*	*p*
PAC	SR	0.38	0.14	0.13	0.48	3.00	0.004
	SU	0.01	0.00	–0.02	–0.01	–0.05	0.959
	DR	0.27	0.07	0.06	0.33	2.07	0.044
	DU	0.11	0.01	–0.01	0.18	0.82	0.414
WMC	SR	0.09	0.01	–0.01	–0.09	–0.67	0.508
	SU	0.11	0.01	–0.01	–0.14	–0.80	0.426
	DR	0.36	0.13	0.12	0.33	2.87	0.006
	DU	0.15	0.02	0.00	0.18	1.10	0.277

Table O-3 Descriptive statistics for overall Jian Trust in Automation scores by experimental condition

Condition	*N*	*M*	*SD*	*SE*
SR	13	66.92	10.21	2.83
SU	13	54.71	16.12	4.47
DR	15	69.12	10.84	2.80
DU	15	58.42	10.87	2.81

Table O-4 Between-condition comparisons of participant trust scores on the Schaefer Trust Survey items

Planned comparison	Δ*M*	*SE*	*p*	d_s	*% difference*
SR > SU	5.13	1.77	0.009	1.14	22.03
SR ≈ DR	–0.74	1.20	0.543	–0.24	–3.18
SU ≈ DU	–1.16	1.83	0.533	–0.25	–6.38
DR > DU	4.72	1.29	0.001	1.33	19.61

Table O-5 Descriptive statistics for overall Schaefer Trust Survey scores by experimental condition

Condition	*N*	*M*	*SD*	*SE*
SR	13	23.31	3.36	0.93
SU	13	18.17	5.43	1.50
DR	15	24.05	2.95	0.76
DU	15	19.33	4.04	1.04

Table O-6 One-way ANOVA results for participant scores on the Schaefer Trust Survey by automation function, across all conditions.

Automation function	*F(3, 52)*	*p*	ω^2
A: Collecting and/or filtering information	4.28	0.009	0.15
B: Integrating and displaying analyzed information	2.44	0.075	0.07
C: Making decisions and/or selecting actions	8.11	0.000	0.28
D: Executing actions	11.00	0.000	0.35

Table O-7 Between-condition comparisons of participant trust scores on the Schaefer Trust Survey items, by automation function

Function	Paired comparison	Δ*M*	*SE*	*p*	d_s	*% difference*
A	SR > SU	4.85	1.86	0.018	1.02	20.26
	SR ≈ DR	–0.34	1.10	0.759	–0.12	–1.44
	SU ≈ DU	–2.72	2.03	0.193	–0.52	–14.27
	DR > DU	2.47	1.37	0.086	0.66	10.16
B	SR > SU	3.62	1.78	0.048	0.70	14.87
	SR ≈ DR	–0.23	1.72	0.896	–0.08	–0.93
	SU ≈ DU	–0.97	1.72	0.574	–0.17	–4.71
	DR > DU	2.87	1.66	0.090	0.73	11.68
C	SR > SU	6.38	1.95	0.003	1.29	28.62
	SR ≈ DR	–0.89	1.59	0.580	–0.21	–4.00
	SU ≈ DU	–1.28	2.01	0.532	–0.24	–8.02
	DR > DU	6.00	1.67	0.001	1.31	25.86
D	SR > SU	6.77	1.84	0.002	1.44	30.34
	SR ≈ DR	–0.43	1.42	0.764	–0.11	–1.93
	SU ≈ DU	0.33	2.07	0.873	0.06	2.09
	DR > DU	7.53	1.70	0.000	1.61	32.47

Table O-8 Descriptive statistics for overall TAM and subscale scores (N = 56) by experimental condition

Measure	Condition	*M*	*SD*	*SE*
Overall TAM	SR	5.68	1.00	0.13
	SU	4.58	1.21	0.16
	DR	5.59	1.04	0.14
	DU	4.59	1.16	0.16
Perceived ease of use	SR	5.79	1.08	0.14
	SU	5.18	1.29	0.17
	DR	5.61	1.26	0.17
	DU	5.21	1.19	0.16
Intent to use	SR	5.14	1.70	0.23
	SU	3.66	1.71	0.23
	DR	5.05	1.81	0.24
	DU	3.75	1.64	0.22
Perceived usefulness	SR	5.77	1.16	0.15
	SU	4.30	1.60	0.21
	DR	5.79	1.12	0.15
	DU	4.26	1.61	0.21

Appendix P. Anthropomorphic Measures Tables

Table P-1 ID factor correlations with anthropomorphic measure scores. Correlations are reported using Pearson's *r*. A Mann–Whitney *U* test was used to evaluate potential differences in trust due to gaming experience (GE).

ID Factor	Anthropomorphism				Animacy				Likeability				Perceived intelligence				Perceived safety			
	SR	**SU**	**DR**	**DU**	**SR**	**SU**	**DR**	**DU**	**SR**	**SU**	**DR**	**DU**	**SR**	**SU**	**DR**	**DU**	**SR**	**SU**	**DR**	**DU**
PAC	–0.08	–0.05	0.02	–0.04	–0.08	–0.13	–0.04	–0.05	–0.05	–0.14	0.02	0.05	0.11	–0.05	0.06	0.04	0.07	–0.07	–0.04	0.08
	0.56	0.72	0.87	0.76	0.55	0.33	0.77	0.71	0.74	0.29	0.87	0.73	0.40	0.70	0.68	0.77	0.62	0.60	0.77	0.54
VS	–0.06	0.05	–0.07	0.04	–0.03	0.15	–0.04	–0.12	0.06	0.11	0.00	–0.10	0.04	0.05	0.17	–0.13	–0.13	0.10	–0.07	–0.17
	0.64	0.69	0.62	0.78	0.81	0.26	0.77	0.38	0.68	0.42	0.98	0.46	0.77	0.74	0.21	0.36	0.33	0.44	0.61	0.20
SOT	–0.02	–0.03	0.09	0.08	0.02	–0.13	0.14	0.06	0.06	–0.24	0.12	–0.04	0.04	–0.06	0.15	0.14	0.12	–0.08	0.22	–0.03
	0.89	0.80	0.51	0.56	0.90	0.33	0.30	0.65	0.64	0.08	0.37	0.75	0.74	0.65	0.26	0.30	0.37	0.54	0.10	0.83
WMC	0.00	–0.10	0.14	0.05	0.02	–0.04	0.16	0.06	–0.13	–0.13	0.17	0.19	0.02	–0.10	0.289[a]	0.17	0.04	0.07	0.16	0.18
	0.99	0.47	0.31	0.72	0.87	0.77	0.25	0.66	0.35	0.34	0.20	0.16	0.90	0.47	0.03	0.21	0.79	0.61	0.25	0.18
IAT	0.04	0.08	–0.12	0.00	–0.03	0.03	–0.18	0.07	0.03	0.13	–0.06	0.15	–0.10	0.18	–0.04	0.14	–0.05	0.12	–0.11	–0.03
	0.78	0.54	0.39	0.99	0.82	0.85	0.19	0.63	0.82	0.35	0.64	0.28	0.47	0.17	0.75	0.32	0.74	0.37	0.42	0.82
GE	378.0	359.0	322.0	374.0	347.0	335.5	323.0	376.0	370.0	286.5	295.5	368.0	333.0	339.5	357.5	355.0	338.0	287.5	300.0	347.0
	–0.03	–0.34	–0.96	–0.09	–0.54	–0.74	–0.94	–0.06	–0.16	–1.56	–1.41	–0.19	–0.78	–0.67	–0.37	–0.41	–0.71	–1.57	–1.35	–0.55
	0.98	0.73	0.34	0.93	0.59	0.46	0.34	0.95	0.87	0.12	0.16	0.85	0.44	0.50	0.71	0.68	0.47	0.12	0.18	0.58

[a] Correlation is significant at the 0.05 level (2-tailed).

Table P-2 Descriptive statistics for Godspeed measure scores (N = 56) by experimental condition

Measure	Condition	*M*	*SD*	*SE*
Anthropomorphism	SR	2.88	0.86	0.12
	SU	2.47	0.70	0.09
	DR	2.97	0.90	0.12
	DU	2.43	0.77	0.10
Animacy	SR	3.29	0.76	0.10
	SU	2.70	0.65	0.09
	DR	3.29	0.77	0.10
	DU	2.82	0.69	0.09
Likeability	SR	3.72	0.65	0.09
	SU	3.09	0.69	0.09
	DR	3.72	0.68	0.09
	DU	3.12	0.72	0.10
Perceived intelligence	SR	4.30	0.52	0.07
	SU	3.11	0.79	0.11
	DR	4.29	0.61	0.08
	DU	3.14	0.87	0.12
Perceived safety	SR	3.52	0.56	0.07
	SU	3.08	0.58	0.08
	DR	3.47	0.65	0.09
	DU	3.13	0.62	0.08

List of Symbols, Abbreviations, and Acronyms

3-D	three-dimensional
ANOVA	analysis of variance
AOI	area of interest
ARL	Army Research Laboratory
ARPI	Autonomy Research Pilot Initiative
ASM	Autonomous Squad Member
CCDC	US Army Combat Capabilities Development Command
CI	confidence interval
d_s	Cohen's *d*, standardized (uses pooled variance)
DOD	Department of Defense
DR	100% reliable/surface-level plus in-depth information (experimental condition)
DU	67% reliable/surface-level plus in-depth information (experimental condition)
ET	event task
Frust	frustration level
GE	gaming experience
IAT	implicit association test
ID	individual difference
IED	improvised explosive device
IR	infrared
M	mean
MD	mental demand
Mdn	median
N	number
NASA–TLX	National Aeronautics and Space Administration–Task Load Index
PAC	perceived attentional control

PD	pupil diameter
Perf	performance
RT	response time
RSPAN	reading span task
SA	situation awareness
SAGAT	Situation Awareness Global Assessment Technique
SAT	Situation-awareness-based Agent Transparency
SD	standard deviation
SE	standard error of the mean
SO	spatial orientation
SOT	spatial orientation test
SPA	spatial ability
SR	100% reliable/surface-level information (experimental condition)
SU	67% reliable/surface-level information (experimental condition)
TAM	technology acceptance measure
TD	temporal demand
TDCE	threat detection correctness efficiency
UCF	University of Central Florida
VS	visual scanning
WMC	working memory capacity

1 (PDF) DEFENSE TECHNICAL INFORMATION CTR DTIC OCA

1 (PDF) CCDC ARL FCDD RLD CL TECH LIB

1 (PDF) CCDC ARL FCDD RLH B T DAVIS BLDG 5400 RM C242 REDSTONE ARSENAL AL 35898–7290

1 (PDF) CCDC ARL FCDD HSI J THOMAS 6662 GUNNER CIRCLE ABERDEEN PROVING GROUND MD 21005–5201

1 (PDF) USAF 711 HPW 711 HPW/RH K GEISS 2698 G ST BLDG 190 WRIGHT PATTERSON AFB OH 45433–7604

1 (PDF) USN ONR ONR CODE 341 J TANGNEY 875 N RANDOLPH STREET BLDG 87 ARLINGTON VA 22203–1986

1 (PDF) USA NSRDEC RDNS D D TAMILIO 10 GENERAL GREENE AVE NATICK MA 01760–2642

1 (PDF) OSD OUSD ATL HPT&B B PETRO 4800 MARK CENTER DRIVE SUITE 17E08 ALEXANDRIA VA 22350

ABERDEEN PROVING GROUND

12 (PDF) CCDC ARL FCDD RLH J LANE Y CHEN P FRANASZCZUK K MCDOWELL K OIE FCDD RLH BD D HEADLEY FCDD RLH FA A DECOSTANZA FCDD RLH FB A EVANS FCDD RLH FC J GASTON FCDD RLH FD S G LAKHMANI A MARATHE J L WRIGHT