ARL-TR-9458 • MAY 2022

DEVCOM ARMY RESEARCH LABORATORY

Bidirectional Communications in Human–Agent Teaming: The Effect of Communication Style

by Julia L Wright, Shan G Lakhmani, Michael R Schwartz, and Jessie YC Chen

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 manufacturers 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-9458 ● MAY 2022

Bidirectional Communications in Human–Agent Teaming: The Effect of Communication Style

Julia L Wright, Shan G Lakhmani, and Jessie YC Chen
DEVCOM Army Research Laboratory

Michael R Schwartz
University of Central Florida

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
May 2022	Technical Report	2 Aug 2018–1 Aug 2020

4. TITLE AND SUBTITLE
Bidirectional Communications in Human–Agent Teaming: The Effect of Communication Style

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Julia L Wright, Shan G Lakhmani, Michael R Schwartz, and Jessie YC Chen

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-9458

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 unlimited.

13. SUPPLEMENTARY NOTES
ORCID IDs:
Julia L Wright, 0000-0003-3026-1538; Shan Lakhmani, 0000-0001-6052-439X; Jessie YC Chen 0000-0003-0557-9042

14. ABSTRACT
This study examined the effects of communication style on human performance, trust, situation awareness, and perceptions of a robot in a human–robot team. In a 2 × 2 mixed-factor study, 32 participants conducted a simulated cordon-and-search-style task while teamed with a robot. Participants were assigned to a communication style (directive vs. nondirective; within) and both groups experienced periods of high versus low task load (among subjects). Results indicate task load was a greater influence on the participants' task performance than communication style, although there were some differential effects on response time and workload due to communication style. This may be due to a difference in feedback inherent in the differing communication styles.

15. SUBJECT TERMS
communication style, human–agent teaming, bidirectional communications, feedback, agent transparency, workload, individual differences, Humans in Complex Systems

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	103	19b. TELEPHONE NUMBER (Include area code) 407-208-3348

Contents

List of Figures

List of Tables

1. Introduction

In a unidirectional communication model, neither side is aware of or capable of addressing the needs of the other (Héder 2014). Bidirectional communication is thought to be less cognitively demanding than unidirectional communication (Héder 2014) as well as inherently more accurate (as the recipient can ask for more information as needed) and timely (as the recipient does not have to wonder if the information is no longer relevant). In human–agent teams, knowledge transfer through communication supports shared awareness (Lyons 2013; Sycara and Sukthankar 2006). When considering the combined effects of these advantages, it is reasonable to expect that a human–agent team using bidirectional communication would have improved performance outcomes compared with a team using unidirectional communication methods. This study proposes to investigate the impact of bidirectional communications in human–agent teams using a series of squad-level, cordon-and-search-like tasks.

Research in human–agent communications has largely focused on the agent's ability to understand the human. Researchers have theorized on the need for robots to understand natural language (Lueth et al. 1994; Mavridis 2015), semantic modelling (Labrou et al. 1999; Yi and Goodrich 2014), gesture recognition (Calinon and Billard 2007; Fiore et al. 2011; Mavridis 2015), and intent recognition (Hayes and Scassellati 2013), and many have begun developing these capabilities and exploring their associated issues through experimentation (Calinon and Billard 2007; Kaupp et al. 2010). It is clear many researchers consider robots that understand human language, context, and intent to be the next step in the evolution of machines. Enabling the agent to gain information by communicating with human teammates has also been shown to improve the agents' performance (Breazeal and Thomaz 2008; Cakmak and Thomaz 2012). While many researchers have investigated the effects of communications within human–agent teams on the human teammates' performance and perceptions (Rau et al. 2009; Selkowitz et al. 2016; Wright et al. 2017; Lakhmani et al. 2019a; Stowers et al. 2020; Wright et al. 2020), relatively few have extended this research to examine the effects of the robots' communication style.

The purpose of this study is to examine to what extent the robot's communication style influences the human teammate's perceptions of an autonomous robotic partner. Prior work has explored how within-team communications influence the human teammate with unidirectional communications, to wit the agent supplying information to the human (without input from the human) regarding its perceptions, goals, and actions (Selkowitz et al. 2016; Lakhmani et al. 2019a; Wright et al. 2020). Evidence shows that within this unidirectional communication setting, the

greater the transparency of the agent in communicating goals, motivations, projected outcomes, and uncertainty information, the more the human teammate trusts the agent, anthropomorphizes the agent, and perceives it to be more intelligent and animate (Lakhmani et al. 2019a; Wright et al. 2020). In addition, increased transparency of the agent better supports the human teammate's situation awareness (SA; Selkowitz et al. 2016).

What has yet to be explored is how the human's ability to communicate with the agent (i.e., to change goals and motivations and preserve resources) will affect the human's perceptions of said agent, trust in the agent, and SA of the agent. Invariably, when agents are deployed with dismounted squads, the squad leader will have the ability and necessity to communicate changing goals and directives to the agent. It is imperative the outcome of this bidirectional communication is understood a priori to understand and avoid (when able) potential difficulties that could be encountered on the battlefield.

1.1 Communication Styles

Human–robot dialogue may affect the human's perceptions of the robot (Kaupp et al. 2010), although the perceptions the human develops regarding the robot could be inaccurate or incorrect. In the teleoperation study from Fong et al. (2003), the human guided a robot through a congested area while maintaining communication with the robot. The robot could question the human, and the human could question the robot as to its status, progress, and current state. Most participants responded when queried by the robot, although some delayed until they were finished with their current task. However, all participants declined to initiate questioning of the robot, indicating that they could infer the robot state by its performance (Fong et al. 2003). This response indicates the human participants were not attributing a very high level of animacy to the robot. When asked why they did not question the robot, participants revealed misconceptions as to how the robot worked or the importance of robot-initiated communications. It is possible the participants were trying to understand communication with the robot using human communication schemas, which proved to be inadequate.

Person-to-person communications tend to be nuanced, occurring for many more reasons than simply information exchange or gathering. The Interpersonal Communication Motive (ICM) model (Rubin et al. 1988) outlines six factors influencing the motives behind why people communicate. People communicate for pleasure, to express affection, to feel included, for escape or relaxation, and to exert control. Communications are made up of three facets: whom we talk to, how we talk to them, and what we talk about (Graham et al. 1993). The "who" could be an

intimate, acquaintance, or coworker and will determine to a large extent how we talk to them and what we talk about, as the relationship among communicators focuses and shapes the interaction (Rubin 1977). However, in a two-member team conducting joint tasks, the "who" is predetermined, and the "what" that is discussed while conducting the task will mostly be limited to task-relevant information (Klein et al. 2005). That leaves the "how" to shape the communications, and that will be dependent on the communicator's communication style.

Norton's communicator style (1978) consists of two dimensions (i.e., directive vs. nondirective) that can be either active or inactive, and is based on interpersonal motives, functions, and one's personal need satisfaction. The directive (DIR) style is dominant, precise, and often contentious, while the nondirective (NDIR) style is friendly, attentive, tactful, and encouraging of others' ideas. The active style is dramatic and animated, while inactive style is relaxed and calm. Norton contested that one's communicator style carries meaning and structures communications. DIR style has been found to be positively correlated with the ICM communication motives of control, inclusion, escape, and pleasure, while the nondirective is positively correlated with the motives of pleasure, affection, inclusion, and relaxation (Graham et al. 1993). Human teammates are sensitive to the robot's communication style, which has been shown to influence their acceptance and perceptions of the robot (Rau et al. 2009). In a cross-cultural study, Chinese participants were more likely to accept a robot's suggestions, and report greater trust, likeability, and credibility, when the robot communicated in an implicit (i.e., nondirective) communication manner rather than an explicit (i.e., directive) manner. However, German participants rated the robot using the implicit communication manner much lower than the explicit robot and were less likely to follow its suggestions (Rau et al. 2009). This indicates it may be important to match the human's preferred communication style to improve interaction efficiency in human–agent teams (Chien et al. 2020; Matthews et al. 2019).

In human–agent teams, the manner in which information is shared is determined by the interface design (Kilgore and Voshell 2014), one aspect of which would be the communication style of the team members. Whether the robot simply shares information about its status and beliefs about its surroundings (unidirectional communication) or the team members have the ability to query one another, updating goals and correcting misinformation (bidirectional communication), is determined not by the team but by the capabilities built into the interface. Hence, to an extent, human perceptions of the agent may be determined not by the agent's task performance or abilities but instead by a design decision made long before the team was deployed. In this work, the impact of communication style on the human's task performance, SA, perceived workload, trust, and perceptions of an autonomous

agent will be assessed. In addition, several individual difference factors that may influence the findings will also be evaluated.

1.2 Situation Awareness

Developing appropriate SA has been shown to be a mission-critical goal for human–robot teams (Evans 2012). Several conceptions of SA exist; the most popular is Endsley's (1995) information-processing-based model. The information-processing-based model suggests an individual's SA comprises three levels, each distinct from the others, yet cumulative in nature. These are Level 1: perception of elements within the environment; Level 2: comprehension of their meaning; and Level 3: projection of their status in the near future (Endsley 1995).

The SA-based Agent Transparency (SAT) model (Chen et al. 2014) provides a framework for the information an agent should provide that supports an individual's SA. Similar to Endsley's model, it also has three levels, each outlining the type of information needed to support the related level in the Endsley model. However, maintaining SA is an ongoing, interactive process between an individual and their environment (Smith and Hancock 1995). When a human is teaming with an agent on a shared task, each must maintain their own SA of their environment, as well as their SA of the other's knowledge, understanding, and abilities, to be effective (Bradshaw et al. 2011). The dynamic SAT model (Chen et al. 2014) represents the continuously updating interactions between the human and agent engaged in a shared task. By comparing performance during the unidirectional communication condition with the bidirectional communications conditions, we can explore the relative utility of the two SAT models.

To assess an individual's current level of SA, we will be using a query method similar to the Situation Awareness Global Assessment Technique (SAGAT). SAGAT is a method where SA-related queries are administered to participants during predetermined pauses of the simulation during the task under analysis (Jones and Kaber 2004; Salmon et al. 2009; Stanton et al. 2012). We will also assess the related concept of confidence in one's SA (Endsley and Jones 1997) using a five-point Likert scale included with each SA probe (McGuinness 2004). In addition to SA, we will measure the participant's perceived workload in communicating with the agent.

1.3 Workload

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." As such, cognitive workload is determined not by the demands of the task but by the capabilities of the operator given particular task load demands. A priority in the proposed study is to see if there is a relationship between workload and the communication style used to relay information between team members. To that end, each participant will complete two scenarios, one at each task load level (high vs. low).

Two different measures of workload will be used. The first measure of participants' perceived workload is the NASA task load index (TLX) (Hart and Staveland 1988). The NASA-TLX asks the participant to rate their level of subjective workload during the experiment. The NASA-TLX is composed of six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration. This measure will be administered after each scenario.

The second is ocular workload measures. These will be recorded using an eye tracker connected to the computer monitor on which the task is displayed. Ocular measures have been shown to be an effective way of measuring workload (Ahlstrom and Friedman-Berg 2006). Blink duration and mean pupil diameter have been shown to positively correlate with cognitive workload (Ahlstrom and Friedman-Berg 2006). The number of fixations positively correlate with task difficulty (Ehmke and Wilson 2007). The proposed study will use these workload measures to assess any differences in cognitive workload induced by the different communication styles.

1.4 Trust

Another research question for the study is how the participant's trust in the agent will be affected by the style in which the teammates communicate. Operator trust is 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). To properly calibrate an operator's trust, a robot needs to meaningfully provide insight on its actions and why it is performing those actions (Chen et al. 2014). Too much trust will cause complacency or misuse, while under-trust causes disuse (Parasuraman and Riley 1997). In the proposed study, trust will be measured using a modified trust in automation scale (Jian et al. 2000). The scale was modified to assess trust within the four classes of automation functions as described in Parasuraman et al. (2000). 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 (Merritt et al. 2013; Krausman et al. 2022). Unlike their explicit counterparts, implicit attitudes are determined by the "strengths of association between concepts (e.g., black people, gay people) and evaluations (e.g., good, bad)" (Project Implicit 2017).

A positive implicit attitude towards automation can result in both good and bad consequences in human-autonomy teaming; it may support user trust in systems that are not reliable; however, it could also cause users to be more likely to demonstrate complacent behavior when teaming with automation (Singh et al. 1993; Merritt et al. 2013). In the current study, explicit trust was assessed using a modified version of the trust in automation scale (Jian et al. 2000), while implicit attitudes towards automation were evaluated using a computer-based Implicit Association Test (IAT) measuring attitude toward automation (Merritt et al. 2013; Project Implicit 2017).

1.5 Robot Intelligence and Humanness

Both the extent to which the robot is perceived as being intelligent and its anthropomorphic tendencies can also influence trust (Ososky et al. 2014; Lee and See 2004). Hinds et al. (2004) found that in human–robot teaming tasks, the human operator felt less responsible for the task when they collaborated with a human-like robot than with a machine-like robot. This finding suggests that when the robot is more human-like, the operator is more willing to cede responsibility for the task outcomes to their robot teammate. In a previous study (Selkowitz et al. 2016), it was shown that when the robot conveyed information regarding its uncertainty and projected outcomes, the operator rated the robot as being more human-like compared to a robot that only conveyed its reasoning and current understanding of its environment. In addition, the robot was rated as more trustworthy, even though its reliability remained unchanged (Selkowitz et al. 2016). The Godspeed questionnaire (Bartneck et al. 2009) will be used to assess participant perceptions of humanness, animacy, likeability, and intelligence of the agent.

1.6 Individual Differences

Additionally, the effects of several individual difference (ID) variables that affect an operator's performance in a multitasking environment will be investigated. These include perceived attentional control (PAC) and working memory capacity (WMC). Previous studies have shown that high PAC and WMC contribute to

performance in simulated environments and robot supervisory tasks (Chen and Terrence 2009; Wright et al. 2018). PAC has been shown to relate to operator performance and SA on tasks that require attention focus and shifting of attention (Chen and Barnes 2012) and will be assessed using the Derryberry and Reed (2002) self-report survey. WMC differences have been shown to affect performance in multirobot supervisory tasks (Ahmed et al. 2014) and SA (Endsley 1995; Wickens and Holland 2000) and will be assessed using the automated reading span task (Redick et al. 2012; Unsworth et al. 2005).

1.7 Current Study

This study explored how bidirectional communication styles interact with task load to affect operator performance, trust, workload, and perceptions of the agent in a multitasking, dynamic environment. The experiment was a mixed-factor design, with communication style (i.e., DIR vs. NDIR) as the between-subjects factor and task load (i.e., low vs. high) as the within-subject variable.

In a simulated multitasking environment, participants conducted a cordon-and-search-type task with a robotic teammate near a busy roadway. While the robot was responsible for searching and securing the rear of the building, the participant was responsible for monitoring the roadway for potential threats (threat detection) and warning the robot of incoming insurgents. Task load (low vs. high) was manipulated by increasing the event rate of the threat-detection task.

Each participant was assigned to a style of communicating with the robot and then completed two trials, one in each task load condition. In both communication conditions, the robot also monitors the roadway and attempts to identify persons who enter the area and determine their actions. The robot then queried the participant whether its assessment of the person is accurate. In the DIR condition, the participant concurred or corrected the agent with no further response from the agent. In the NDIR condition, after the participant concurred or corrected, the agent reviewed the information and notified the participant whether it agreed or disagreed with the participant's response.

1.8 Research Objective

The goal of this research is to understand how differing communication styles interact with task load, within a human–agent teaming context, to influence the human's performance, trust, workload, SA, and perceptions of the agent.

1.8.1 Hypothesis 1

Hypothesis 1 (H1) Task Performance:

H1a: Participants in the DIR condition will perform better on the communications task than those in the NDIR condition. Communication task performance will be assessed by correct responses and how quickly responses are made.

H1b: Within each communication style condition, communication task performance will be higher in the low task load (LTL) condition than in the high task-load condition.

H1c: Participants in the DIR condition will perform better on the target-identification task than those in the NDIR condition. Target-identification task performance will be assessed by correct identifications and how quickly the targets are identified.

H1d: Within each communication style condition, target-identification performance will be higher in the low task-load condition than in the high task-load condition.

1.8.2 Hypothesis 2

Hypothesis 2 (H2) Trust:

H2a: Participants in the DIR condition will have higher trust in the robot than those in the NDIR condition.

H2b: Within each communication style condition, participant trust in the robot will be greater in the high task-load condition than in the low task-load condition.

1.8.3 Hypothesis 3

Hypothesis 3 (H3) Workload:

H3a: Participants in the NDIR condition will have greater workload than those in the DIR condition.

H3b: Within each communication style condition, participant-perceived cognitive workload will be greater in the high task-load condition than in the low task condition.

1.8.4 Hypothesis 4

Hypothesis 4 (H4) SA:

H4a: Participants in the DIR condition will have higher SA than those in the NDIR condition.

H4b: Within each communication style condition, participant SA will be higher in the high task-load condition than in the low task-load condition.

1.8.5 Hypothesis 5

Hypothesis 5 (H5) Perception of the robot:

H5a: Participants in the DIR condition will perceive the robot to have lower animacy, be less likable, have lower intelligence, and be less safe than those in the NDIR condition.

H5b: Within each communication style condition, task load will affect participant perceptions of the agent, which will be higher (i.e., greater animacy, more likeable, higher intelligence, and safer) in the low task-load condition than in the high task-load condition.

1.8.6 Hypothesis 6

Hypothesis 6 (H6) Individual Differences:

H6: There will be differential results on all dependent measures (i.e., target-detection performance, trust, workload, SA, and perceptions of the agent) due to ID (i.e., IAT, WMC, and PAC).

2. Method

2.1 Participants

Forty-five participants (ages 18–40) 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. Thirteen potential participants were dismissed from the study or the data from their sessions was deemed useless and had to be replaced: one was given incorrect condition sequences that rendered that data useless, eight experienced equipment malfunctions, and four failed the Ishihara Color Vision Test. Those who were dismissed received payment for the time they participated, at a minimum for 1 h. The 32 remaining participants (12 males, 20 females; Min_{age} = 18 years, Max_{age} = 26 years, M_{age} = 20.25 years)* successfully completed the experiment, and their data was used in the analysis. Sample size calculations using G*Power (Faul et. al 2007) indicate that for a medium effect size (Cohen's f = .35), a minimum 30 participants was required.

* *Min* = minimum, *Max* = maximum, and *M* = median.

2.2 Apparatus

2.2.1 Simulator

A custom software application capable of showing images and video was used to present the experimental simulation to the participant. The simulator was coded in the HAVOK simulation engine. The simulation will be delivered via a commercial desktop computer system, two 22-inch monitors, standard keyboard, and 3-button mouse. The left-side monitor displays the Soldier's point of view (POV) of the task environment (Fig. 1). The right-side monitor displays the robot's communication interface (Fig. 2).

Fig. 1 Left-side screen shows the virtual environment from a squad member's POV. Ahead is the building the robot teammate is searching; traffic crosses the area between, and the participant uses the buttons at the bottom of the screen to identify and report to the robot the nature and behavior of the traffic.

Fig. 2 Right-hand screen shows the autonomous agent's interface. Left half shows an overhead view of the area and displays real-time movement in the monitored area. In the upper left-hand corner the robot transparency modules are displayed: the upper shows the robot's current goal, priority, and projected loss; the lower shows the robot's understanding of the human's current goal, priority, and potential loss. Right half of screen has a camera feed showing the robot's POV (upper), communications windows (center), and robot's resources usage (lower).

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 (Fig. 3). 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. The eye-tracker system comprises two pairs of cameras with IR lights placed under the computer monitor. The Smart Eye system uses reflections of IR flashes on the cornea to determine gaze direction, as well as recording iris and pupil information.

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

2.2.3 Facilities

The study was conducted in an indoor, climate-controlled laboratory/office space with the participant seated at a typical office desk.

2.3 Surveys and Tests

2.3.1 Demographic Questionnaire

A demographics questionnaire was administered at the beginning of the experimental session (Appendix A). Information on participant's age, gender, education level, computer familiarity, and gaming experience (GE) was collected to rule out any effects due to these differences. Participants who played action video games at least weekly were classified as gamers (Gamers N = 1, NonGamer N = 31); but as the majority were classified as nongamers this measure was removed from further evaluation.

2.3.2 Ishihara Color Vision Test

An Ishihara Color Vision Test (1972) using nine test plates (Appendix B) was administered via PowerPoint slide presentation. Since the interface employs several colors to display the robot dialogue and plans, normal color vision is required to effectively interact with the system. Four potential participants failed to correctly identify at least seven of the plates, so they were paid for 1 h ($15) and dismissed.

2.3.3 Implicit Association Test

Implicit trust was measured using the modified 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 imply 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, and then reversed so that higher scores indicate greater implicit trust in automation: Min_{IAT} = –1.67, Max_{IAT} = 1.81, Mdn_{IAT} = –0.24, SD_{IAT} = 0.98, IAT_{LOW} N = 16, and IAT_{HIGH} N = 16.*

2.3.4 NASA-TLX

Participants' perceived workload was evaluated with the computerized version of the NASA-TLX questionnaire (Appendix D), 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.

2.3.5 Godspeed Questionnaire Series (GQS)

The GQS (Appendix E) assesses 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 the full GQS after each experimental trial. Questions were presented in random order.

2.3.6 Functional Trust Survey

The Functional Trust survey (Appendix F) was developed to further distinguish the basis of an individual's trust in an autonomous agent. It comprises the Trust in Automation survey (Jian et al. 2000) modified to include trust assessment along the four functions of automation use identified by Parasuraman, Sheridan, and Wickens (2000): A—gathering and filtering information, B—integrating and displaying

* *Mdn* = median and *SD* = standard deviation

information, C—suggesting or making decisions, and D—executing actions. Participants completed a Functional Trust survey after each experimental trial.

2.3.7 Reading Span Task (RSPAN)

Verbal working memory capacity (WMC) was assessed using the automated RSPAN (Appendix G), 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 indicated greater WMC: $Min_{RSPAN} = 11$, $Max_{RSPAN} = 54$, $Mdn_{RSPAN} = 30.00$, $M_{RSPAN} = 32.099$, and $SD_{RSPAN} = 11.02$. High/low group membership was determined by median split of all participants' scores: $\text{RSPAN}_{\text{LOW}}$ $N = 16$ and $\text{RSPAN}_{\text{HIGH}}$ $N = 16$.

2.3.8 Attentional Control Survey

A questionnaire on Attentional Control (Appendix H) 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. 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: $Min_{PAC} = 40$, $Max_{PAC} = 54$, $Mdn_{PAC} = 44.50$, $M_{PAC} = 44.63$, $SD_{PAC} = 3.46$; PAC_{LOW} $N = 16$, and PAC_{HIGH} $N = 16$.

2.4 Experimental Design and Performance Measures

2.4.1 Experimental Design

This study was a 2 × 2 mixed-factor design. Within-subjects evaluations compared differences in participant behavior and attributions regarding the robot across varying task-load conditions. Between-subjects evaluations assessed how differences in communication style contributed to these differences.

2.4.2 Independent Variables

The independent variables were communication style and task load. Communication style was manipulated by assigning participants to one of two conditions: directive (DIR) or nondirective (NDIR). Task load was manipulated by varying the event rate for traffic in the monitored area between High Traffic and Low Traffic. Participants completed two trials in their assigned communication

style, one high traffic and one low traffic. Task-load condition sequence was counterbalanced across participants.

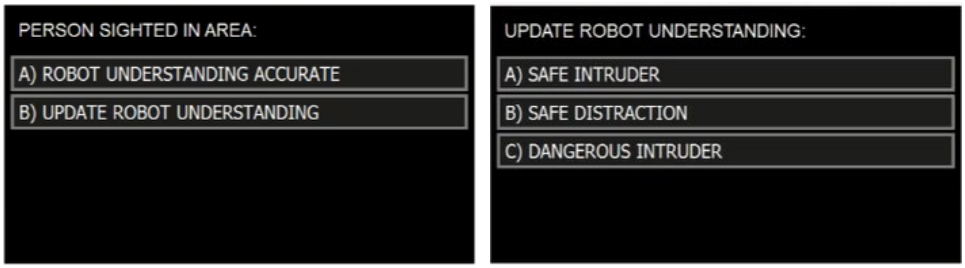
In the LTL condition, traffic in the monitored area appeared at an approximate rate of 1 signal every 10 s. In the HTL condition, this rate was increased to approximately 1 signal every 5 s. The 1:2 event ratio has been shown to result in distinct level differences in participant performance and workload (Abich et al. 2013). Traffic (signals) consisted of pedestrians and a variety of vehicles. On average, it took 10 s for a signal to enter, cross the monitored area, and exit on the opposite side. Signals nearer the participant could obstruct the view of signals further from the participant; however, all signals were clear and unobstructed for at least 6 s.

2.4.3 Dependent Measures

2.4.3.1 Communications

Central to the team's cordon-and-search task is bidirectional team communications—the robot kept the human apprised of its understanding of the current situation, and the human assisted the robot by either verifying or correcting the robot's understanding. When a person entered the cordoned area, the robot asked the human teammate to verify that its understanding of the type of person (dangerous or not) and their behavior (passing through or attempting to enter the building) was accurate. In the communications area of the display, the robot first asks if its understanding is accurate (as indicated in the human transparency module, upper left-hand corner of the monitor, in Fig. 2). If the participant indicated the robot's understanding was not accurate, the robot would suggest three likely options. One offered suggestion was always correct.

Scoring is as follows: If the robot's understanding is accurate and the participant selects A, they receive 2 points. If the robot's understanding is incorrect and the participant selects B, they receive 1 point (Fig. 4a). There are no points awarded or penalties for incorrect responses. When the participant selects B, the second dialogue screen is displayed (Fig. 4b) showing three potential descriptions of the person type and behavior. If the participant selects the correct option, they receive 1 point.



a b

Fig. 4 (a) Bidirectional communications dialogue windows: when the robot identifies a person in the cordoned area, it asks if its assessment of the person is accurate; (b) if the participant indicates the robot's understanding is inaccurate, a selection of potential explanations is displayed

2.4.3.2 Target Identification

During the cordon-and-search task, the human sends reports to the robot identifying persons and vehicles in the monitored area. The accuracy and speed of these notifications were assessed.

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, pupil diameter (in millimeters), and blink-duration 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 robot through all three SA levels, Situation Awareness Global Assessment Technique-style SA queries (Jones and Kaber 2004) were employed. 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. SA queries were designed to assess the participants' SA at a specific SA level: SA1—Level 1 SA, perception; SA2—Level 2 SA, comprehension; SA3—Level 3 SA, the projection of future state (see Appendix I for example questions). SA queries were scored as correct (+1) or incorrect (–1), with higher scores indicating better SA.

2.4.3.5 Trust

After each mission, the Functional Trust survey was administered to assess the participants' trust and perceived usability of the robot.

2.4.3.6 Anthropomorphic Tendencies

The Godspeed measure was administered after each trial to assess the participant's attributions of anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of the robot.

2.4.3.7 Eye-Tracking Measures

Supporting information will also come from eye-tracking data:

- Pupil diameter
- Number of fixations
- Blink duration

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. They then completed the demographics questionnaire and the IAT. Participants were randomly assigned to a communication style (i.e., DIR vs. NDIR) and trained via a PowerPoint slideshow to familiarize themselves with the display characteristics and the expectations from a cordon-and-search-like task. During training, participants received a series of evaluations to confirm they understood the material. Following training, participants completed a series of computer-based evaluation exercises of the overall task to ensure they were knowledgeable of the display characteristics and able to apply them to the task. Participants who scored less than 80% on these evaluations were allowed to review the material and redo the assessment. All participants successfully completed the training. After training, participants were offered a short break. After the break, the eye tracker was calibrated to the participant, and they began the experimental session.

The experimental task required the participant to work with a simulated robot in a series of squad-level cordon-and-search-like tasks. The participant observed two monitors, one displaying a simulated environment and the other displaying a robot interface. The robot acted as a search element, exploring a building for high-value targets. During this scenario, the robot encountered events that affected its goals, rationale, and projected future state. Using the robot's interface, the human monitored the robot's actions while simultaneously acting as a cordon element,

identifying prespecified stimuli of interest in the simulated environment. Jointly, the human and the robot kept people out of the building; the participant alerted the robot when individuals approached the building's entrance, and the robot chased away any intruders who attempted to enter the building. The participant encountered events and the robot would communicate its understanding of the human's status. When this occurred, the robot would enquire of the participant if its understanding was accurate, and the participant communicated (according to their assigned communication style) this information to the robot. These communications did not influence the simulation's outcome or alter the participant's task.

Participants completed two scenarios, one scenario in each task load level (low vs. high). After each scenario, participants completed the NASA-TLX, the Functional Trust scale, and the GQS. After completing both scenarios, the participants completed a working memory test and the PAC questionnaire. After completing all scenarios and tests, participants were thanked for participation, and any questions they had pertaining to the study were answered before dismissal.

3. Results

3.1 Data Analysis

Data analysis was performed using SPSS V24 software. Data was examined using analysis of variances (ANOVAs) and t-tests ($\alpha = 0.05$). A Bonferroni correction for multiple comparisons was used when applicable. Planned comparisons were conducted to examine differences between conditions, specifically (DIR, L) to (NDIR, L); (DIR, H) to (NDIR, H); (DIR, L) to (DIR, H); and (NDIR, L) to (NDIR, H). These comparisons were evaluated for each dependent measure. Means, SD, and 95% confidence interval are reported for each measure.

Effect sizes are reported using standardized Cohen's d (d_s) for differences between means and omega-squared (ω^2) for population-based effects estimates. The population-based effect estimate (ω^2) is used as it is more conservative than the sample-based effect estimate eta-squared (η^2).

Individual difference (ID) factors (i.e., IAT, PAC, and WMC) were assessed for potential differential effects on the factors of interest. When an ID factor was revealed to be a significant predictor or correlate highly with the measure of interest, these results are reported.

3.2 Task Performance

3.2.1 Communications

The participants were tasked with ensuring the robot's understanding of the detected persons and their activity in the monitored area was accurate. As such, they were required to respond to robot queries, verifying when the robot's understanding was accurate, and correcting when it was not. Performance was evaluated via two methods: percent of correct responses (of answered queries) and how quickly participants responded to the query. (See Table J-1 in Appendix J for descriptive statistics.)

3.2.1.1 Correct Responses

It was expected that participants would have fewer correct responses in the NDIR condition than in the DIR condition, and fewer correct responses in the high task-load condition than in the low task-load condition.

Within each communication style condition there was no difference in the number of correct responses due to task load: NDIR: $t(16) = 0.40$, $p = 0.693$, $d_s = 0.12$, and DIR, $t(14) = 1.11$, $p = 0.284$, $d_s = 0.32$.

There was no difference in the number of correctly answered queries (as a percentage of answered queries) between communication conditions: $F(1, 31) = 0.01$, $p = 0.908$, $\omega^2 = 0.03$. These effects were also examined between communication conditions within each level of task load, but no difference was found in either the high task-load condition, $F(1, 31) = 0.20$, $p = .655$, $\omega^2 = .02$, or low task-load condition, $F(1,31) = 0.11$, $p = .917$, $\omega^2 = .03$.

Summary: Hypotheses H1a and H1b were not supported. Neither communication style nor task load affected the number of correct responses participants made in each scenario.

3.2.1.2 Response Times

It was expected that participants in the NDIR condition would have longer response times than those in the DIR condition, and response times would be longer in the high task-load condition than in the low task-load condition.

Within the NDIR condition there was a slight difference in overall response times between task load conditions, $t(16) = 1.68$, $p = 0.112$. Participants in the high task-load condition ($M = 4.0$ s, SD = 0.67) took longer to respond to agent queries than those in the low task-load condition ($M = 3.6$ s, SD = 0.73, $d_s = 0.46$). In the DIR condition there was no difference in overall response times due to task load, $t(14)$

= 0.05, p_{single} = 0.481, d_s = 0.01. When task load was high, participants in the NDIR condition (M = 4.0 s, SD = 0.67) took longer to respond than those in the DIR condition (M = 3.6 s, SD = 0.62, d_s = 0.55). See Fig. 5.

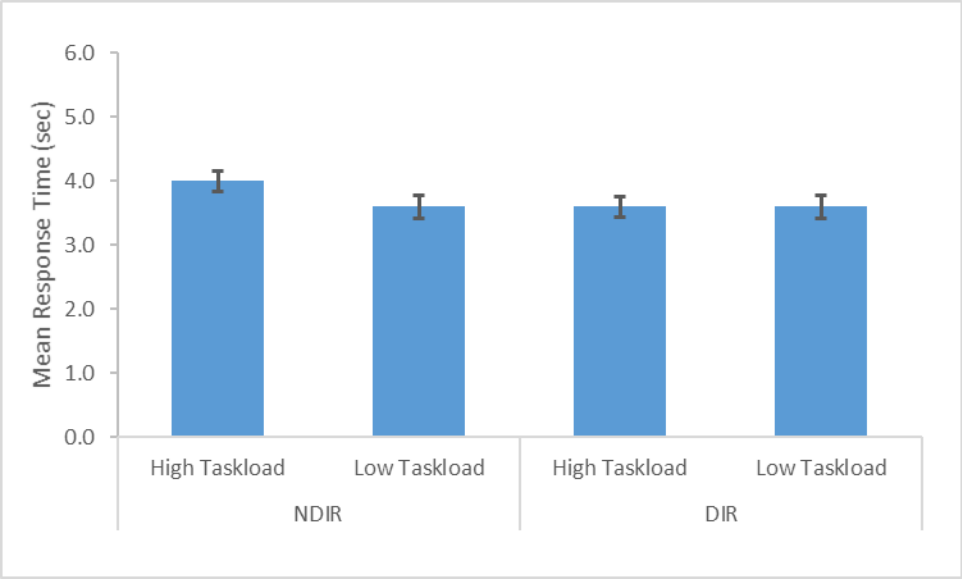


Fig. 5 Response time by task load

It was also expected participants would take longer to respond when their responses were correct rather than incorrect, indicating that lack of time would be the main contributor to incorrect responses.

There was a significant difference in the mean response time for correct and incorrect responses in both the low, $t(30) = 3.65$, $p < 0.001$, $d_s = 0.65$, and high, $t(26) = 2.16$, $p = 0.040$, $d_s = 0.32$, task-load conditions regardless of communication style. As expected, participants' mean response time for correct responses (low: M = 3.8 s, SD = 0.73; high: M = 3.9 s, SD = 0.64) were longer than those for incorrect responses (low: M = 3.2 s, SD = 1.09; high: M = 3.6 s, SD = 1.15). See Fig. 6.

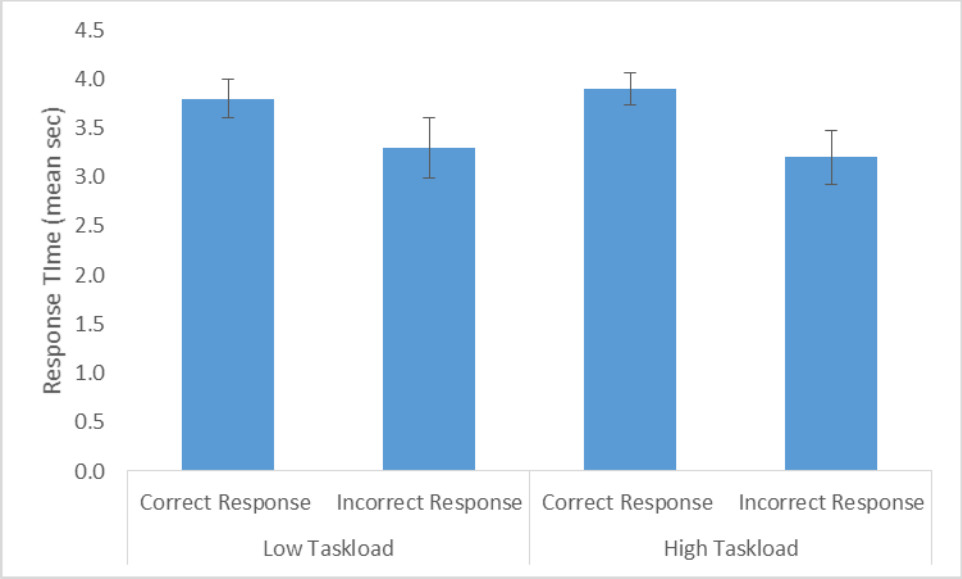


Fig. 6 Overall correct vs. incorrect response times by task load

This difference in response times was also found within each communication style condition for the low task-load scenarios, but not the high task load (HTL) in the NDIR condition.

In the DIR condition, there was a significant difference in the mean response time for correct and incorrect responses in both the low, $t(14) = 1.81$, $p = 0.092$, $d_s = 0.51$, and high, $t(13) = 3.29$, $p = 0.006$, $d_s = 0.84$, task load conditions. Participants' mean response time for correct responses (low: $M = 3.8$ s, SD = 0.73; high: M = 3.9 s, SD = 0.59) were longer than those for incorrect responses (low: M = 3.3 s, SD = 1.2; high: M = 3.2 s, SD = 1.02). See Fig. 7.

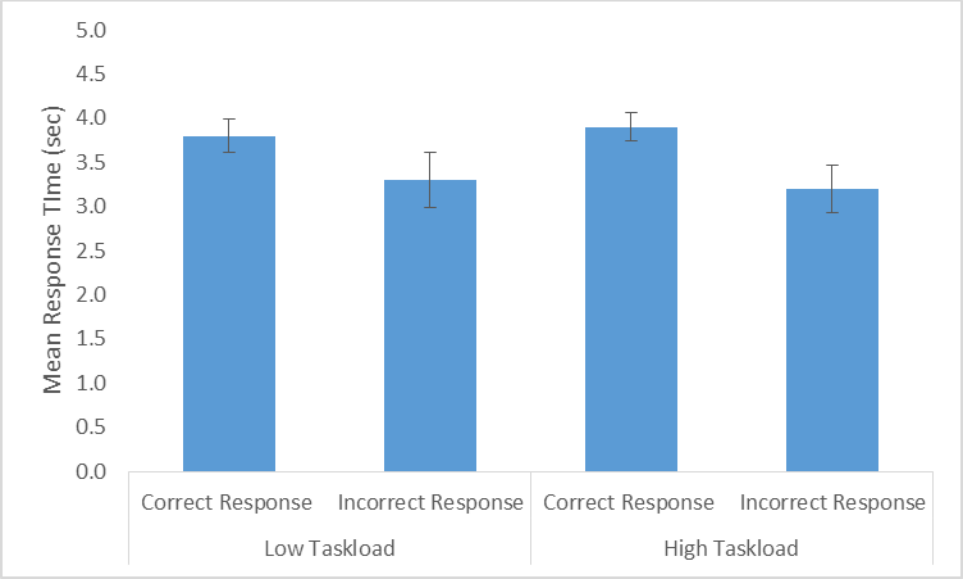


Fig. 7 DIR condition response times for correct vs incorrect responses by task load

In the NDIR condition there was a significant difference in the mean response time for correct and incorrect responses in the LTL, $t(15) = 3.52$, $p = 0.003$, $d_s = 0.91$, but not the HTL, $t(12) = 0.35$, $p = 0.732$, $d_s = 0.00$, condition. Participants' mean response time for correct responses in the low task-load condition (M = 3.9 s, SD = 0.75) were considerably longer than those for incorrect responses ($M = 3.1$ s, SD = 1.00), but not so in the high task-load condition (correct: $M = 4.0$ s, SD = 0.70; incorrect: $M = 4.0$ *sec*, SD = 1.20). See Fig. 8.

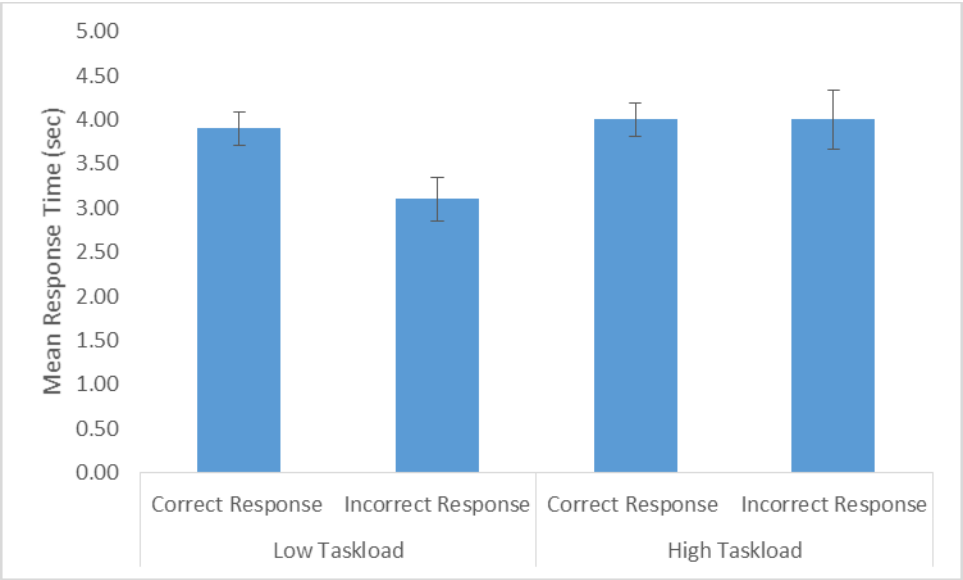


Fig. 8 NDIR condition response times for correct vs. incorrect responses by task load

Summary: There was partial support for H1a and H1b. There was no difference in overall response time due to communication style. Within each communication style condition, participants took longer to respond with correct responses than with incorrect responses in both task load conditions, which indicates that lack of time may have contributed to the incorrect responses. There was no difference in response time due to task load in the DIR; however, participants in the NDIR condition took longer to respond to robot queries when task load was high than in the low task-load condition, regardless of whether their response was correct or incorrect.

3.2.2 Target Identification

Participants were also tasked with monitoring the cordoned area and identifying persons and vehicles that enter the area, as well as indicate their behavior and potential for danger to themselves or the robot. Performance on this task was evaluated using ANOVAs and t-tests. Overall, observed power on these tests was quite low (< 0.3); however, sample effect sizes were consistently medium to large. This disparity indicates there most likely is a difference between the groups; however, there were not enough samples in the current tests to result in a significant p-value. As such, evaluations are conducted using the effect sizes, with differences of medium and above ($d_s > 0.5$; $\omega^2 > 0.04$) interpreted as a reportable difference. F and t-test results are reported for completeness, however, not used for interpretation of findings. See Table J-2 in Appendix J for descriptive statistics and Table J-3 for t-test results.

3.2.2.1 Persons in the Area—Correct Identifications

It was expected participants would have fewer correct identifications of persons in the cordoned area in the NDIR condition than in the DIR condition and fewer correct identifications in the high task-load condition than in the low task-load condition.

Within the NDIR condition there was a difference in the percent of correct identifications, $t(16) = 2.60$, $p = 0.019$, $d_s = 0.59$, due to task load. Participants had fewer correct identifications in the high task-load condition ($M = 49.4\%$, SD = 37.2) than in the low task-load condition ($M = 67.9\%$, SD = 24.5). In the DIR condition there was no difference in correct identifications, $t(14) = 0.45$, $p = 0.661$, $d_s = 0.10$, due to task load. See Fig. 9.

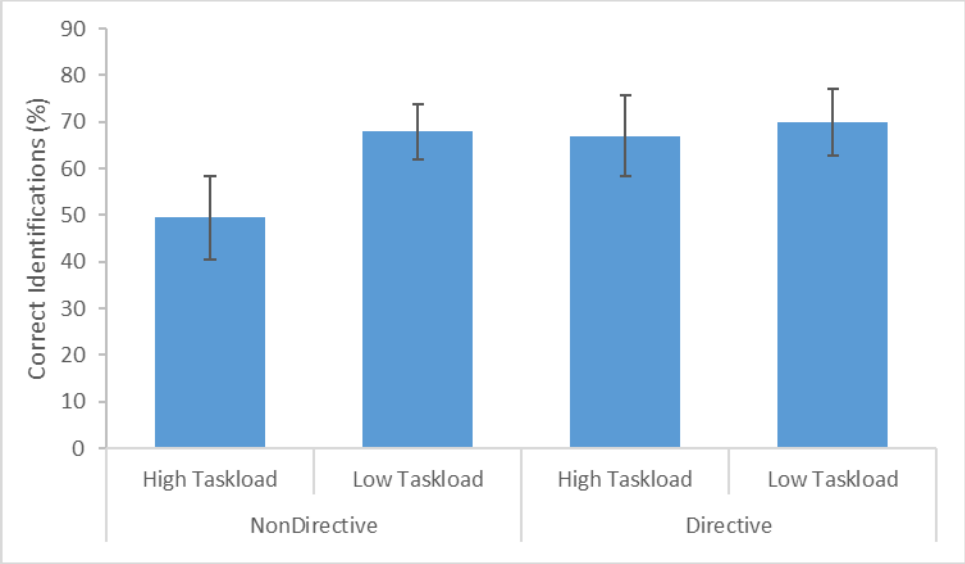


Fig. 9 Persons in cordoned area correctly identified, by communication style and task load

Between the communication styles, participants in the NDIR condition (M = 58.8%, SD = 27.4) had fewer correct identifications than those in the DIR condition (M = 73.0%, SD = 20.8, d_s = 0.58), $F(1,30) = 2.54$, $p = 0.122$, $\omega^2 = 0.05$, regardless of task load.

In the low task-load condition, there was no difference in the number of correct identifications between the NDIR condition (M = 67.9%, SD = 24.1) and the DIR condition (M = 75.0%, SD = 20.6), $t(30) = -0.22$, $p = 0.825$, $d_s = 0.31$. In the high task-load condition, participants in the NDIR condition (M = 60.0%, SD = 31.9%) had considerably fewer correct identifications than those in the DIR condition (M = 77.3%, SD = 21.7), $t(30) = -1.39$, $p = 0.174$, $d_s = 0.63$.

Summary: Hypothesis H1c was supported; participants in the NDIR communications style had fewer correct identifications of persons in the cordoned area than their DIR-style counterparts. There was partial support for H1d, as there was no difference in correct identifications in the DIR condition due to task load, but in the NDIR condition correct identifications were no better than chance when task load was high.

3.2.2.2 Persons in the Area—Response Time

It was expected participants in the NDIR condition would have longer response times than those in the DIR condition and response times would be longer in the high task-load condition than in the low task-load condition.

For the task of identifying persons in the cordoned area, there was no difference in response time within the NDIR condition, $t(13) = 1.28$, $p = 0.223$, $d_s = 0.39$, or the DIR condition, $t(12) = 0.65$, $p = 0.526$, $d_s = 0.21$, due to task load.

Between the communication styles, participants in the NDIR condition ($M = 4.2$ s, SD = 1.6) took slightly longer to identify persons than those in the DIR condition ($M = 3.5$ s, SD = 1.3, $d_s = 0.43$), regardless of task load, $F(1,30) = 1.43$, $p = 0.242$, $\omega^2 = .01$.

In the low task-load condition, participants in the NDIR condition ($M = 4.3$ s, SD = 1.2) had longer response times than those in the DIR condition ($M = 3.5$ s, SD = 1.3, $d_s = 0.64$), $F(1,30) = 3.16$, $p = 0.086$, $\omega^2 = 0.07$. In the high task-load condition, participants in the NDIR condition ($M = 4.9$ s, SD = 2.3) had longer response times than those in the DIR condition ($M = 3.9$ s, SD = 1.6, $d_s = 0.54$), $F(1,30) = 1.97$, $p = 0.173$, $\omega^2 = 0.03$. See Fig. 10.

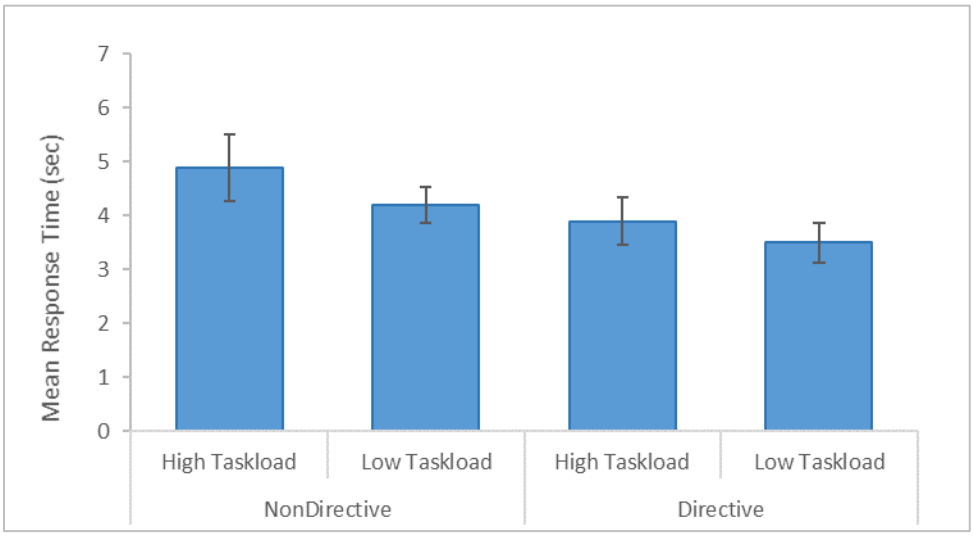


Fig. 10 Time to identify persons in the cordoned area, by communication style and task load

Summary: There was partial support for H1c and no support for H1d. There was no difference in correct-identification response times solely due to communication style or task load. However, when response times for each communication style were compared by task load, NDIR style had longer response times than DIR style.

3.2.2.3 Dangerous Persons in the Area—Correct Identifications

It was expected participants would have fewer correct identifications of dangerous persons in the cordoned area in the NDIR condition than in the DIR condition and fewer correct identifications in the high task-load condition than in the low task-load condition.

Within the NDIR condition there was a significant difference in the percent of correct identifications, t(16) = 3.73, p = 0.002, d_s = 0.59, due to task load. Participants had fewer correct identifications in the high task-load condition (M = 77.0%, SD = 22.5) than in the low task-load condition (M = 88.2%, SD = 14.7). Within the DIR condition there was a reportable difference in the percent of correct identifications, $t(31) = 3.40$, $p = 0.002$, $d_s = 0.50$, due to task load. Participants had fewer correct identifications in the high task-load condition (M = 81.3%, SD = 20.8) than in the low task-load condition (M = 90.4%, SD = 15.1).

Between the communication styles, participants had no reportable difference in correct identifications of dangerous persons in either the NDIR condition (M = 82.4%, SD = 17.9) or the DIR condition (M = 89.5%, SD = 14.8, d_s = 0.43), $F(1,31) = 1.51$, $p = 0.229$, $\omega^2 = 0.02$.

In the low task-load condition, there was no difference in the number of correct identifications of dangerous persons between the NDIR condition (M = 88.2%, SD = 14.7) and the DIR condition (M = 92.8%, SD = 15.7, d_s = 0.30), $F(1,31) = 0.71$, $p = 0.406$, $\omega^2 = 0.01$. However, in the high task-load condition there were fewer correct identifications of dangerous persons in the NDIR condition (M = 77.0%, SD = 22.6) than in the DIR condition (M = 86.1%, SD = 18.3, d_s = 0.44), $F(1,31) = 1.56$, $p = 0.221$, $\omega^2 = 0.02$.

Summary: Hypothesis H1c was supported; participants in the NDIR communications style had fewer correct identifications of dangerous persons in the cordoned area than their DIR counterparts. There was support for H1d, as there was a significant difference in correct identifications in both the NDIR and DIR conditions due to task load. See Fig. 11.

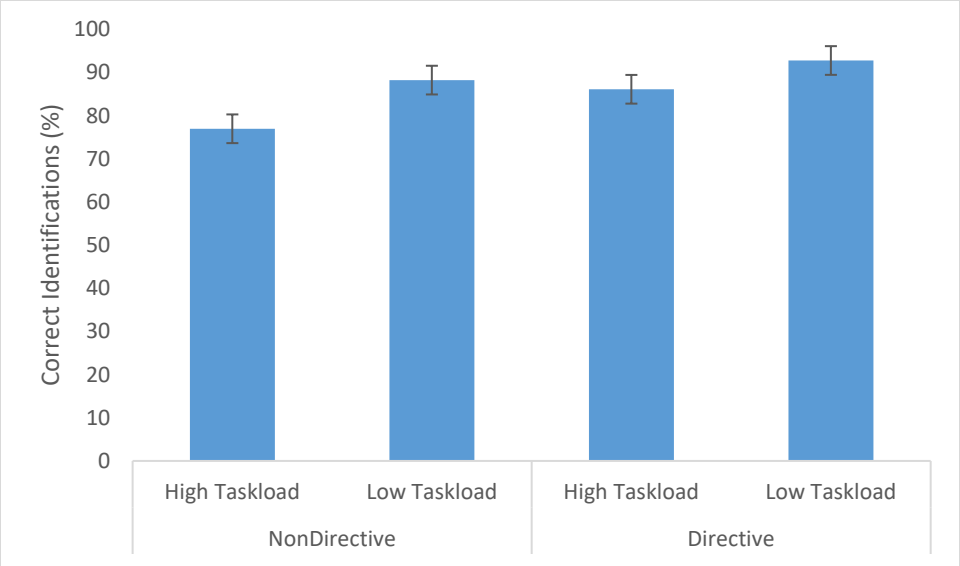


Fig. 11 Dangerous persons in the cordoned area correctly identified, by communication style and task load

3.2.2.4 Dangerous Persons in the Area—Response Time

It was expected participants in the NDIR condition would have longer response times than those in the DIR condition and response times would be longer in the high task-load condition than in the low task-load condition.

Within the NDIR condition there was a significant difference in response time due to task load condition, $t(16) = 2.34$, $p = 0.033$, $d_s = 0.88$. Participants in the high task-load condition took considerably longer to identify dangerous persons in the cordoned area ($M = 6.0$ s, SD = 2.6) than those in the low task-load condition ($M = 4.0$ s, SD = 1.9). Within the DIR condition there was no difference in response time to identify dangerous persons due to task load condition, $t(14) = 0.25$, $p = 0.810$, $d_s = 0.10$.

Between the communication styles, participants in the NDIR condition ($M = 5.1$ s, SD = 1.5) took longer to identify dangerous persons than those in the DIR condition ($M = 4.2$ s, SD = 1.3, $d_s = 0.68$), regardless of task load, $F(1,31) = 3.72$, $p = 0.063$, $\omega^2 = 0.08$.

In the low task-load condition participants in the NDIR condition ($M = 4.0$ s, SD = 1.9) had similar response times to those in the DIR condition (4.2 s, SD = 1.8, $d_s = 0.08$), $F(1,31) = 0.05$, $p = 0.832$, $\omega^2 = 0.03$. However, in the high task-load condition, participants in the NDIR condition ($M = 6.0$ s, SD = 2.6) had longer response times to those in the DIR (4.0 s, SD = 2.1, $d_s = 0.85$), $F(1,31) = 5.81$, $p = 0.022$, $\omega^2 = 0.13$.

Summary: There was support for H1c and partial support for H1d. Response times for correct identification of dangerous persons in the cordoned area were longer in the NDIR condition than the DIR condition. When task load was high, response times in the NDIR condition were longer than those in the DIR condition, but when task load was low there was no difference between the two conditions. See Fig. 12.

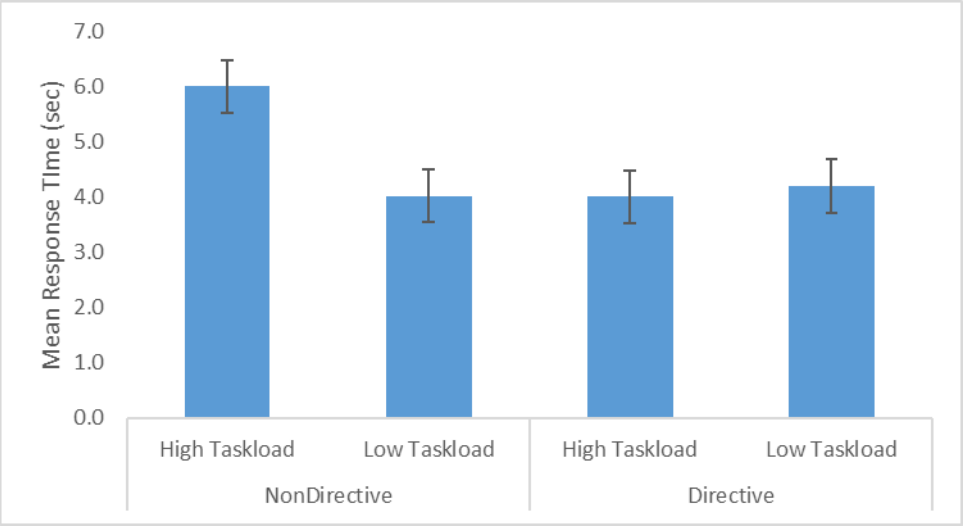


Fig. 12 Time to identify dangerous persons in the cordoned area, by communication style and task load

3.2.2.5 Suspicious Vehicles in the Area—Response Time and Correct Identifications

It was expected participants would have fewer correct identifications of suspicious vehicles in the cordoned area in the NDIR condition than in the DIR condition and fewer correct identifications in the high task-load condition than in the low task-load condition.

Within the NDIR condition there was a difference in response time to identify suspicious vehicles due to task load, $t(16) = 2.04$, $p = 0.058$, $d_s = 0.52$. In the high task-load condition, participants took longer to identify suspicious vehicles (M = 3.47 s, SD = 0.53) than in the low task-load condition (M = 3.13 s, SD = 0.78). There was no difference due to task load in the percentage of correct identifications of suspicious vehicles, $t(16) = 1.60$, $p = 0.130$, $d_s = 0.36$, or the percent of false positives, $t(16) = 1.42$, $p = 0.174$, $d_s = 0.35$.

Within the DIR condition there was no difference in response time to identify suspicious vehicles due to task load, $t(14) = 0.09$, $p = 0.930$, $d_s = 0.03$, the number of correct identifications, $t(14) = 0.65$, $p = 0.527$, $d_s = 0.21$, or the percent of false positives, $t(14) = 0.00$, $p = 1.000$, $d_s = 0.00$.

Between the communication styles there was no difference in response time or percent of correct identifications of suspicious vehicles.

In the low task-load condition, participants in the NDIR condition (M = 66.5%, SD = 14.1) had more false reports of suspicious vehicles than those in the DIR condition (M = 58.7%, SD = 11.9, d_s = 0.59), $F(1,31) = 2.82$, $p = 0.103$, $\omega^2 = 0.02$. There was no difference in either the percent of correct identifications of suspicious vehicles or the mean response time due to communication style. In the high task-load condition, there was no difference in the percent of correct identifications of suspicious vehicles, the percent of false positives, or the mean response time due to communication style.

Summary: There was no support for H1c and partial support for H1d. Correct identifications, reported false positives, and response times for correct identification of suspicious vehicles in the cordon area were not different between communication styles. Response times in the NDIR condition were longer when task load was high, although there was no difference in correct identifications or false positives and there was no difference due to task load in the DIR condition for any of the measures.

3.2.2.6 Obstacles in the Area—Response Time and Correct Identifications

It was also expected participants would have fewer correct identifications of obstacles in the cordoned area in the NDIR condition than in the DIR condition and fewer correct identifications in the high task-load condition than in the low task-load condition.

Within the NDIR condition there was no difference in response time to identify obstacles due to task load, $t(16) = 0.99$, $p = 0.333$, $d_s = 0.22$, or the percent correctly identified, $t(16) = 0.35$, $p = 0.734$, $d_s = 0.12$. There was a significant difference in the number of false positives reported, $t(16) = 6.53$, $p < 0.001$, $d_s = 2.36$, with more false positives occurring in the HTL (M = 0.40, SD = 0.15) than in the low task-load condition (M = 0.10), SD = 0.10).

Within the DIR condition there was no difference in response time to identify obstacles due to task load, $t(14) = 0.82$, $p = 0.043$, $d_s = 0.20$. There was a difference in the percent correctly identified, $t(14) = 1.64$, $p = 0.124$, $d_s = 0.51$, with a greater percentage of obstacles correctly identified in the low task-load condition (M = 0.94, SD = 0.10) than in the high task-load condition (M = 0.88, SD = 0.12). There was a significant difference in the number of false positives reported, $t(14) = 6.45$, $p = 0.000$, $d_s = 2.62$, with more false positives occurring in the HTL (M = 0.37, SD = 0.13) than in the low task-load condition (M = 0.08, SD = 0.09).

Between the communication styles, there was no difference in response time or percent of correct identifications of obstacles in the cordoned area.

In both task-load conditions there was no difference in the percent of correct identifications of obstacles, the percent of false positives, or the mean response time due to communication style.

Summary: There was no support for H1c and partial support for H1d. Correct identifications, reported false positives, and response times for correct identification of obstacles in the cordoned area were not different between communication styles. More false positives were reported in the NDIR condition when task load was high, although there was no difference in correct identifications or response times due to task load. In the DIR condition there were more false positives when task load was high and more correct identifications when task load was low, but no difference in response time.

3.3 Trust in the Agent

Upon completion of each scenario, participants assessed their trust in their robotic partner using the Functional Trust survey (Appendix F). It was expected participants in the DIR condition would report higher trust than those in the NDIR condition (H2a); further, within each communication style condition, reported trust in the robot would be higher when the task load is high (H2b). Tables for descriptive statistics and t-test results are in Appendix K.

Within the NDIR and DIR conditions there was no difference in trust, whether overall or any of the specific functions, due to task load. In both task-load conditions there was no difference in trust, whether overall or any of the specific functions, due to communication style.

Summary: There was no support for either H2a or H2b. Participants' self-reported trust in the robot was not affected by either communication style or task load.

3.4 Workload

3.4.1 NASA-TLX (Subjective Workload Assessment)

Upon completion of each scenario, participants assessed their cognitive workload using the NASA-TLX (Appendix D). It was expected participants in the NDIR condition would report higher workload than those in the DIR condition (H3a), and within each communication style condition, reported cognitive workload will be higher when the task load is high (H3b). Tables for NASA-TLX findings (i.e., descriptive statistics and t-test results) are in Appendix L.

3.4.1.1 Global (Unweighted) Score

Between communication styles there was no difference in cognitive workload global scores—$t(30) = –0.37$, $p = 0.711$; $M_{NDIR} = 51.62$, $SD_{NDIR} = 6.40$, $M_{DIR} = 52.99$, $SD_{DIR} = 13.43$, $d_s = 0.13$—nor was there a difference in cognitive workload global scores due to task load level.

Within each communication style condition, paired t-tests indicated there was no difference in global workload scores due to task load.

3.4.1.2 Mental Demand

Between communication styles there was no difference in mental-demand scores—$t(30) = 1.22$, $p = 0.232$; $M_{NDIR} = 76.32$, $SD_{NDIR} = 11.15$, $M_{DIR} = 70.67$, $SD_{DIR} = 15.01$, $d_s = 0.43$—nor was there a difference in mental-demand scores due to task load level.

Within each communication style condition, paired t-tests indicated there was no difference in mental-demand scores due to task load.

3.4.1.3 Physical Demand

Between communication styles there were significant differences in physical-demand scores, both overall—$t(30) = –1.98$, $p = 0.057$; $M_{NDIR} = 8.82$, $SD_{NDIR} = 11.49$, $M_{DIR} = 19.00$, $SD_{DIR} = 17.32$, $d_s = 0.70$—and within each task load level. Participants in the DIR condition reported greater physical demand than those in the NDIR condition (Fig. 13)

Within the NDIR style condition, paired t-tests indicated there was a moderately significant difference in physical-demand scores due to task load level. Participants in the high task-load condition reported greater physical demand than those in the low task-load condition. In the DIR condition there was no difference in physical demand scores due to task load.

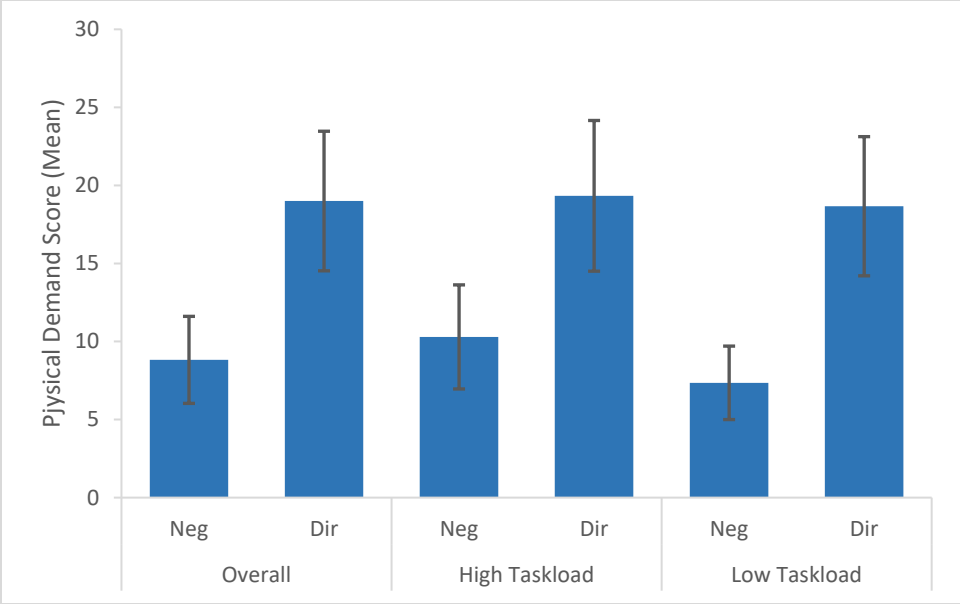


Fig. 13 NASA-TLX physical-demand scores by task load and communication style; bars denote standard error of the mean (SE)

3.4.1.4 Temporal Demand

Between communication styles there was no difference in temporal-demand scores—$t(30) = 0.31$, $p = 0.759$; $M_{NDIR} = 65.00$, $SD_{NDIR} = 18.75$, $M_{DIR} = 62.83$, $SD_{DIR} = 20.87$, $d_s = 0.11$—nor was there a difference in temporal-demand scores due to task load level.

Within each communication style condition, paired t-tests indicated there was no difference in temporal demand scores due to task load.

3.4.1.5 Effort

Between communication styles there was no difference in perceived-effort scores—$t(30) = 0.24$, $p = 0.815$; $M_{NDIR} = 64.85$, $SD_{NDIR} = 12.58$, $M_{DIR} = 63.83$, $SD_{DIR} = 11.80$, $d_s = 0.08$—nor was there a difference in perceived-effort scores due to task load level.

Within the NDR condition, paired t-tests indicated there was no difference in perceived-effort scores due to task load; however, in the DIR condition, perceived-effort scores in the high task-load condition were greater than those in the low task-load condition.

3.4.1.6 Frustration

Between communication styles there was no difference in frustration scores—$t(30) = 0.18$, $p = 0.861$; $M_{NDIR} = 48.82$, $SD_{NDIR} = 22.88$, $M_{DIR} = 50.17$, $SD_{DIR} = 19.74$, $d_s = 0.06$—nor was there a difference in frustration scores due to task load level.

Within each communication style condition, paired t-tests indicated there was no difference in frustration scores due to task load.

3.4.1.7 Performance

Between communication styles there was no difference in perceived performance scores—$t(30) = -1.13$, $p = 0.269$; $M_{NDIR} = 45.29$, $SD_{NDIR} = 17.74$, $M_{DIR} = 52.33$, $SD_{DIR} = 17.54$, $d_s = 0.40$—nor was there a difference in perceived performance scores due to task load level.

Within each communication style condition, paired t-tests indicated there was no difference in perceived performance scores due to task load.

Summary: The NASA-TLX results offered limited support for H3a and H3b. Participants reported greater physical demand in the DIR condition than in the NDIR condition, which is opposite of the predicted outcome for H3a. Within the communication style conditions, participants in the NDIR condition reported greater physical demand when task load was high, and participants in the DIR condition reported greater effort when task load was high, supporting H3b.

3.4.2 Eye-Tracking Measures (Objective Workload Assessment)

While completing each scenario, participant's eye movements and behaviors were recorded using a Smart Eye two-camera system. Ocular indices have been shown to indicate increased cognitive workload (i.e., blink duration and pupil diameter) and difficulty obtaining and/or understanding information (i.e., number of fixations) (Nakayama et al. 2002). Tables for Ocular Measure findings (i.e., descriptive statistics and t-test results) are in Appendix L.

Within each task load condition there was a significant difference in blink duration between the NDIR and DIR conditions. Participants in the NDIR condition had longer blink durations than those in the DIR condition (high $d_s = 0.95$ and low $d_s = 0.97$), indicating greater mental workload. Participants in the high task-load NDIR condition had more fixations than those in either the high task-load DIR ($d_s = 0.51$) or the low task-load NDIR ($d_s = 0.49$) conditions (Fig. 14). Within the NDIR condition, participants in the high task-load condition had larger pupil diameters than those in the low task-load condition ($d_s = 0.14$).

Summary: Evaluation of ocular measures supported both H3a and H3b. Blink duration, regardless of task load, was longer in the NDIR condition than in the DIR condition, supporting H3a. In the NDIR condition, fixation count and pupil diameter were greater in the high task-load condition than in the low task-load condition, giving partial support for H3b.

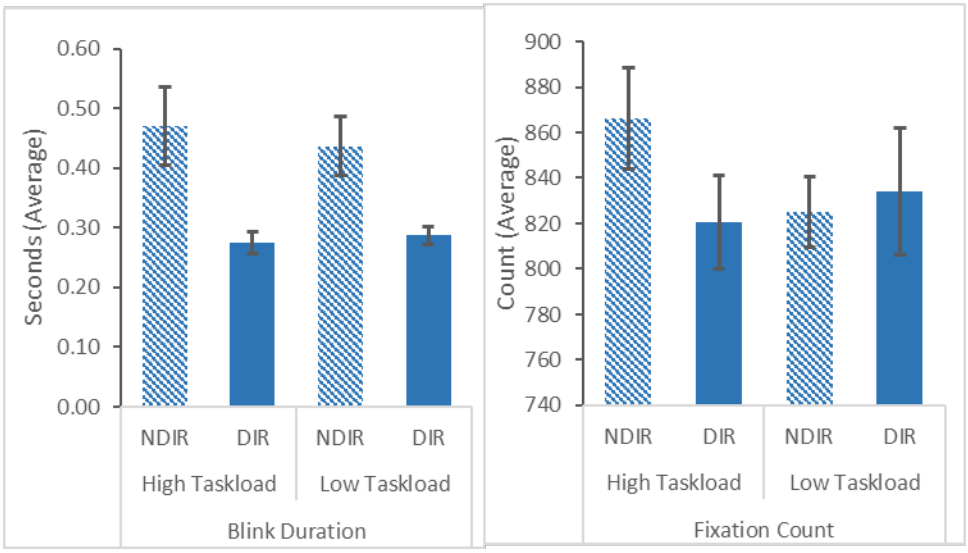


Fig. 14 Ocular indices' results; bars denote SE

3.5 Situation Awareness

During each scenario, there were several pauses wherein participants assessed the SA of their robotic partner, its reasoning, and the likely outcomes of its actions. It was expected that participants in the DIR condition would have better SA of their robotic partner than those in the NDIR condition (H4a), and within each communication style condition the participant's SA of the robot would be better when the task load is low (H4b). Tables for descriptive statistics and t-test results are in Appendix M.

Within the NDIR and DIR conditions there was no difference in participant SA, whether overall or any of the specific SA levels, due to task load. In both task load conditions there was no difference in participant SA, whether overall or any of the specific levels, due to communication style.

Summary: There was no support for either H4a or H4b. Participant-reported SA of the robot was not affected by either communication style or task load.

3.6 Godspeed Survey

Upon completion of each scenario, participants evaluated their robotic partner using the Godspeed Survey. It was expected that participants in the DIR condition would rate the robot as being less animate, less likeable, less intelligent, and less safe than those in the NDIR condition. It was also expected that these ratings would also be lower when task load was high, as compared with low task-load conditions. Tables for descriptive statistics and t-test results are in Appendix N.

Within each task load condition there was no difference in Godspeed evaluations between the NDIR and DIR communication styles.

Within each communication style there were significant differences due to task load for anthropomorphism, animacy, and likeability (Fig. 15). In the NDIR communication style condition, participants rated the robot in the low task-load condition as more anthropomorphic (d_s = 0.47), animate (d_s = 0.70), and likeable (d_s = 0.31) than in the high task-load condition. In the DIR communication style condition, participants rated the robot in the low task-load condition as more animate (d_s = 0.51) and Likeable (d_s = 0.32) than in the high task-load condition.

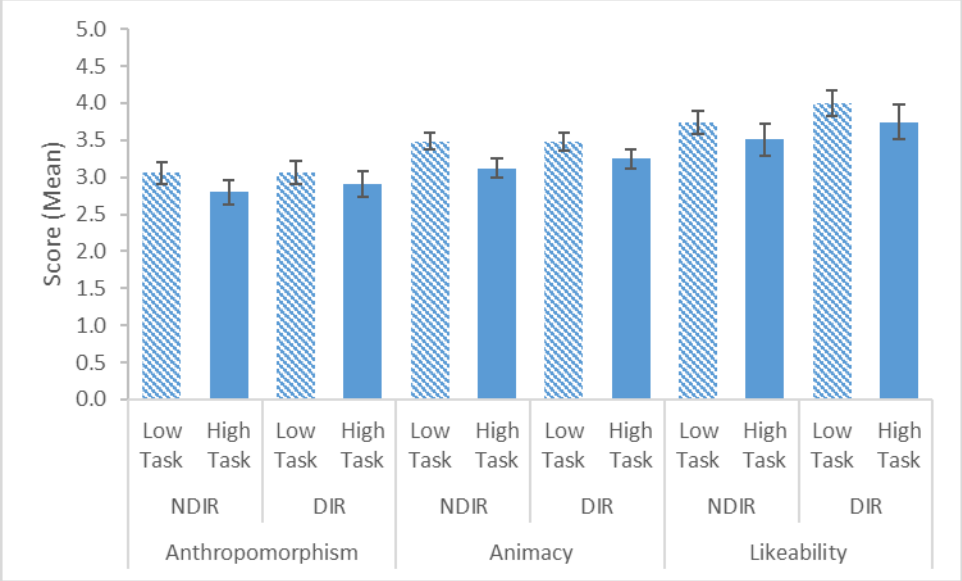


Fig. 15 Godspeed survey's paired t-test results, by task load within communication style

Summary: There was no support for H5a, as there were no differences in Godspeed evaluations due to communication style. There was partial support for H5b. Within each communication style, participants in the low task-load condition found the robot to be more animate and likeable and within the NDIR condition more anthropomorphic than in the high task-load condition. However, they did not note any difference in perceived intelligence or safety of the robot due to task load.

3.7 Individual Differences

It was hypothesized that there would be differential results on all dependent measures due to individual difference factors (H6). Correlations among the ID factors of implicit trust (shown via the Implicit Association Test), WMC, and PAC and each of the dependent variables were examined. The communication style groups were relatively small (NDIR N=17 and DIR N=15), as such correlations were examined at two-tailed significance, and moderately significant findings ($p < 0.08$) are described in this section (and listed in Appendix O).

3.7.1 Task Performance—Communications

There was a moderately significant negative correlation between IAT and correct responses for participants in the DIR style condition ($r = –0.48$ and $p = 0.067$). In the DIR condition, participants with lower implicit trust in the autonomy had more correct responses than those with higher implicit trust.

3.7.2 Task Performance—Identifications

Reported false positives for identifying persons in the cordoned area correlated with IAT scores in both NDIR ($r = 0.44$ and $p = 0.080$) and DIR ($r = 0.55$ and $p = 0.034$) conditions, indicating that regardless of communication style, participants with greater implicit trust in the autonomy reported more false positives than those with less implicit trust.

Implicit trust in the robot correlated positively with response time for identifying persons ($r = 0.60$ and $p = 0.019$) and suspicious vehicles ($r = 0.49$ and $p = 0.064$) in the cordoned area in the DIR condition. Participants with higher implicit trust took longer to identify persons and suspicious vehicles than those with lower implicit trust.

Greater working memory capacity was positively correlated with both the number of correct suspicious-vehicle identifications and the number of reported false positives of suspicious vehicles, indicating that regardless of communication style, participants with greater WMC reported more suspicious vehicles than those with lower WMC.

In the NDIR condition there was a negative correlation between PAC and the number of correctly identified suspicious vehicles. Participants with lower reported attentional control correctly identified more suspicious vehicles than those with higher attentional control, $r = –0.49$ and $p = 0.046$.

3.7.3 Trust Survey's Scores

There were no significant correlations between any individual-difference factors and any of the trust survey results.

3.7.4 Cognitive Workload

Perceived attentional control was negatively correlated with the global NASA-TLX score (r = –0.45 and p = 0.072) and the number of fixations (r = –0.57 and p = 0.017) in the NDIR condition. Participants with lower attentional control experienced greater cognitive workload than those with higher attentional control.

3.7.5 Situation Awareness

There were no significant correlations between any ID factors and any of the SA results.

3.7.6 Godspeed Survey

In the DIR condition, participants who were low in attentional control anthropomorphized the robot more (r = –0.49 and p = 0.062) than those with high attentional control.

Summary: There was partial support for H6. Implicit trust differences were apparent in both the communications and identifications task for the DIR-condition participants. PAC differences resulted in differential outcomes in the identification task and cognitive workload (NDIR condition) and anthropomorphism of the agent for those in DIR condition.

4. Discussion

4.1 Synopsis and Review

The goal of this study was to examine to what extent the style of intrateam communications influences a human's perceptions of an autonomous robotic teammate and performance on tasking. Previous research has examined the impact of directionality and content of human–robot communications on human perceptions and performance (Héder 2014; Lyons 2013; Chen et al. 2018). We examined how the robots' style of communications (i.e., active vs. inactive) affects the human in the human–robot team to better understand how to better support the human's task performance, workload, SA, trust, and perceptions of the robot.

Participants' primary task had two components: 1) maintain communications with the robot and 2) identify potential threats in the cordoned area. It was expected that

the additional attentional demands on those in the NDIR style condition would negatively affect their performance both in accuracy and response time.

On the communications task, neither communication style nor task load influenced the number of correct communications. However, when task load was high, participants in the NDIR condition took longer to respond to communications queries than in the LTL, as well as longer than their HTL DIR-condition counterparts. Participants also took longer to respond when their responses were correct than when they were incorrect, except in the HTL NDIR condition. It was suggested this indicated the incorrect responses might be due to participants feeling rushed, so they hurried their responses. If so, the lack of difference in the HTL NDIR condition would indicate the combined impact of HTL and the back-and-forth of the NDIR condition exacerbated this problem.

For the identification task, participants had to identify persons, dangerous (armed) persons, obstacles (stopped vehicles), and suspicious vehicles (large trucks) that entered the cordoned area. Task performance varied: When identifying persons and dangerous persons, those in the NDIR condition had fewer correct responses than those in DIR condition; in the NDIR condition, HTL performance was (overall) worse than LTL performance. Response times for identifying persons and dangerous persons were also longer in the NDIR condition than in the DIR. Identifying obstacles had somewhat similar results: While there were no differences in correct identifications or response time due to communication style, there were more reported false positives in the NDIR condition than in the DIR. When it came to identifying suspicious vehicles, the only difference in task performance was within the NDIR conditions due to task load. Understanding the subtle differences in identifying these threats helps us understand what these findings imply. Identifying suspicious vehicles took only a glance—there was one vehicle type and color that was considered suspicious, making it readily recognizable. Identifying obstacles took slightly more effort—the vehicle stopped as it entered the cordoned area and began to move again after 6 s. So, a second glance may have been needed to recognize this vehicle had stopped—slightly more effort than recognizing the suspicious vehicle. Identifying persons in the cordon area took even more effort—the persons' movements varied, some had weapons, and some approached the building. Participants needed to check closely on persons several times to ascertain whether they were merely a person in the area or if they posed a greater threat. These results suggest that when task load and task complexity are low there is little difference in performance due to communications style. However, as complexity and task load increase, the NDIR style of communicating requires more resources than the DIR style and, as a result, task performance suffers (Wickens 2002).

Communication style had no noticeable effect on participants' explicit trust in the robot; however, their implicit trust did correlate with task performance. Participants in both communication style conditions with greater implicit trust reported more false positives than those with lower implicit trust. Participants in the DIR condition with greater implicit trust had longer response times when identifying persons and vehicles in the cordon area; also, they had fewer correct communications responses than those with lower implicit trust, which could indicate they had difficulty with the communications task that their lower-implicit-trust counterparts did not. It is also possible those with lower implicit trust were better able to appropriately match their explicit trust in the robot with the robot's capabilities, thus properly supporting the task demands.

Subjective workload ratings indicate participants in the DIR condition reported higher physical demand than those in the NDIR condition, and they reported greater effort when task load was high than when it was low. However, optical measures indicate those in the NDIR condition had greater workload than those in the DIR condition as well as greater workload when task load was high than when it was low. While it is not uncommon for optical measures to be more sensitive to cognitive-workload differences than a subjective measure such as the NASA-TLX, it is unusual for the results to be completely different. It was expected that those in the NDIR condition would have greater workload than those in DIR, as their communications with the robot required them to monitor the communication window more closely to see the robot's final response—and optical measures showed this was in fact true. However, persons in the NDIR condition did not report any subjective differences in workload, while those in the DIR condition reported greater physical demand and effort.

The reason for difference may be found within the feedback literature. Not only has feedback style (i.e., informational vs. controlling) been found to affect a recipients' performance (Ryan 1982; Zhou 1998), so has feedback valence (i.e., positive vs. negative). Receiving positive feedback gives participants a boost in their perceived competence and motivation; however, receiving negative feedback (or no feedback) does not lead to this same boost and may in fact contribute to a performance decline (Zhou 1998). However, when the feedback is comprehensive (both positive and negative) and is delivered informally, the recipient may get both the boost of the positive and the performance-calibrating input of the negative feedback without suffering the detrimental consequences of negative feedback alone (Zhou 1998). In this study, the participants in the NDIR condition received comprehensive feedback in an informal manner, while those in the DIR condition received no feedback. As such, although those in the NDIR condition were expending more effort to conduct their tasks, their bolstered sense of competence

and performance feedback appeared to alleviate any subjective sense of increased workload (Becker et al. 1995). Alternatively, those in the DIR condition received no feedback; as such, they had no assistance in assessing their competence on the task and this in turn led to an increased subjective evaluation of their effort (Ryan 1982; Becker et al. 1995).

4.2 Limitations and Future Directions

While there has been research into the effects of feedback style and valence in human task performance, very little has been done in the area of human–autonomy teaming (HAT). The perception of high workload can be as damaging to task performance as actual workload, and the ability to incorporate methods to alleviate this perception for the human teammate has broad-reaching potential.

5. Conclusions

Understanding how differences in communication style influence human–robot interactions, teaming, and human perceptions of the robot is important to effective interface design. Prior work has demonstrated the benefits of bidirectional communications over unidirectional communications in HAT (Chen et al. 2018; Lakhmani et al. 2019b) this research explored how the style of communication content further affects that relationship.

This could be considered a study on who has the last word in a dialogue that influences team performance. Both styles were bidirectional; however, in the DIR condition there were no further communications after the participant responded to the robot, while in the NDIR condition the robot had a final response and the robot was able to disagree with the participant. This was expected to severely impact participant trust in the robot and perceptions of the robot, neither of which occurred. Performance in the identification task was worse in the NDIR style condition when it required more effort (i.e., identifying persons vs. vehicles), and actual workload was higher in the NDIR condition than in the DIR. These findings indicate it was not so much the style of communication that caused these differences, but rather the amount of work to perform within a fixed amount of time. When workload was more manageable, there was no difference in performance. Overall, it appears task load had greater impact on task performance than communication style. However, the feedback those in the NDIR condition received on their assessments of robot perceptions may have influenced their perception of their workload.

These results are useful to interface designers, whose interface designs essentially script how the interaction between a human and an autonomous teammate will proceed when conducting their tasks. While the extra exchange in the NDIR

condition was not overly harmful to the team, it did require extra time. However, it appears to have offered some protection regarding their perceived workload, which could prove beneficial to the team when task timing is not crucial. This indicates that when it is deemed important for the agent to have the "final say" in an exchange, it can be implemented with little concern.

Portions of this work have been previously reported.[*]

[*] Wright JL, Lakhmani SG, Chen JYC. Bidirectional communications in human-agent teaming: the effects of communication style and feedback. Inter J of Human-Computer Interaction. 2022 May. doi:10.1080/10447318.2022.2068744.

6. References

Abich J, Reinerman-Jones L, Taylor G. Establishing workload manipulations utilizing a simulated environment. International Conference on Virtual, Augmented and Mixed Reality; 2013 July. Springer, Berlin, Heidelberg. p. 211–220.

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 modeling of networked human-automation performance using working memory capacity. Ergonomics. 2014;57(3):295–318.

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

Becker AB, Warm JS, Dember WN, Hancock PA. Effects of jet engine noise and performance feedback on perceived workload in a monitoring task. Int J Aviat Psych. 1995;5(1):49–62.

Bradshaw JM, Feltovich PJ, Johnson M. Human–agent interaction. In: Boy GA editor. The Handbook of Human-Machine Interaction: A Human-Centered Design Approach. CRC Press; 2011.

Breazeal C, Thomaz AL. Learning from human teachers with socially guided exploration. ICRA 2008: IEEE International Conference on Robotics and Automation; 2008 May. IEEE. p. 3539–3544.

Cakmak M, Thomaz AL. Designing robot learners that ask good questions. Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction; 2012 Mar. ACM. p. 17–24.

Calinon S, Billard A. Incremental learning of gestures by imitation in a humanoid robot. Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction; 2007 Mar. ACM. p. 255–262.

Chen JYC, Barnes MJ. Supervisory control of multiple robots in dynamic tasking environments. Ergonomics. 2012;55(9):1043–1058.

Chen JYC, Lakhmani SG, Stowers K, Selkowitz AR, Wright JL, Barnes MJ. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theor Issues Ergon Sci. 2018;19(3):259–282.

Chen JYC, Procci K, Boyce M, Wright JL, Garcia A, Barnes MJ. SA-based agent transparency. Army Research Laboratory (US); 2014 Apr. Report No.: ARL-TR-6905.

Chen JYC, Terrence PI. Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment. Ergon. 2009;52(8), 907–920.

Chien S, Lewis M, Sycara K, Kumru A, Liu J. Influence of culture, transparency, trust, and degree of automation on automation use. IEEE Trans Hum Mach Syst. 2020;50(3):205–214. doi: 10.1109/THMS.2019.2931755.

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

Ehmke C, Wilson S. Identifying web usability problems from eyetracking data. Paper presented at: British HCI Conference; 2007 Mar 9–July 9; University of Lancaster, UK.

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

Endsley M, Jones WM. Situation awareness information dominance and information warfare. Logicon Technical Services Inc; 1997.

Evans AW 3rd. Safe operations of unmanned systems for reconnaissance in complex environments–Army technology objective (SOURCE ATO) field experimentation observations and soldier feedback. Army Research Laboratory (US); 2012 July. Report No.: ARL-TN-0488.

Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39:175–191.

Fiore SM, Badler NL, Boloni L, Goodrich MA, Wu AS, Chen J. Human-robot teams collaborating socially, organizationally, and culturally. Proc Hum Fact Ergon Soc Ann Meet. 2011;55(1):465–469.

Fong T, Thorpe C, Baur C. Robot, asker of questions. Rob Auton Syst. 2003;42(3):235–243.

Graham EE, Barbato CA, Perse EM. The interpersonal communication motives model. Comm Quart. 1993;41(2):172–186.

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.

Hart S, Staveland L. Development of NASA TLX (task load index): results of empirical and theoretical research. In: Hancock P, Meshkati N, editors. Human Mental Workload. Elsevier; 1988. p. 139–183.

Hayes B, Scassellati B. Challenges in shared-environment human-robot collaboration. Learning. 2013;8(9).

Héder M. The machine's role in human's service automation and knowledge sharing. AI Soc. 2014;29(2):185–192.

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

Ishihara S. Test for colour-blindness. Kanehara Shuppan Co., Ltd; 1972.

Jian JY, 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 A, Hendrick H, Brookhuis K, Salas E, editors. Handbook of Human Factors and Ergonomics Methods. CRC Press; 2004. p. 419–427.

Kaupp T, Makarenko A, Durrant-Whyte H. Human–robot communication for collaborative decision making—a probabilistic approach. Rob Auton Syst. 2010;58(5):444–456.

Kilgore R, Voshell M. Increasing the transparency of unmanned systems: applications of ecological interface design. International Conference on Virtual, Augmented and Mixed Reality; 2014 June. p. 378–389. Springer, Cham.

Klein G, Feltovich PJ, Bradshaw JM, Woods DD. Common ground and coordination in joint activity. Org Sim. 2005;53:139–184.

Krausman A, Neubauer C, Forster D, Lakhmani SG, Baker A, Fitzhugh SM, Gremillion G, Wright JL, Metcalfe JS, Schaeffer KE. Team trust measurement in human-autonomy teaming. IEEE Transactions on Human-Machine Systems. 2022.

Labrou Y, Finin T, Peng Y. Agent communication languages: the current landscape. IEEE Intel Syst Appl. 1999;14(2):45–52.

Lakhmani SG, Wright JL, Schwartz M, Barber D. Exploring the effect of communication patterns and transparency on the attitudes towards robots. Springer, Cham. International Conference on Applied Human Factors and Ergonomics;2019a. p 27–36.

Lakhmani SG, Wright JL, Schwartz M, Barber D. Exploring the effect of communication patterns and transparency on performance in a human-robot team. Proc Hum Fact Ergon Soc Ann Meet. 2019b;(63).

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

Lueth TC, Laengle T, Herzog G, Stopp E, Rembold U. KANTRA-human-machine interaction for intelligent robots using natural language. RO-MAN'94. Proceedings of the 3rd IEEE International Workshop on Robot and Human Communication; 1994 July; Nagoya, Japan. p. 106–111.

Lyons JB. Being transparent about transparency: a model for human-robot interaction. AAAI Spring Symposium Series; 2013.

Matthews G, Lin J, Panganiban AR, Long MD. Individual differences in trust in autonomous robots: implications for transparency. IEEE Trans Hum-Mach Syst. 2019;50(3):234–244.

Mavridis N. A review of verbal and non-verbal human–robot interactive communication. Rob Auton Syst. 2015;63:22–35. doi: doi.org/10.1016/ j.robot.2014.09.031.

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

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 Fact. 2013;55(3):520–534. DOI:10.1177/0018720812465081.

Nakayama M, Takahashi K, Shimizu Y. The act of task difficulty and eye-movement frequency for the "Oculo-motor indices." Proceedings of the 2002 Symposium on Eye Tracking Research and Applications. 2002. p. 37–42.

Norton RW. Foundation of a communicator style construct. Hum Comm Res. 1978;4(2):99–112.

Ososky S, Sanders T, Jentsch F, Hancock P, Chen JY. Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. SPIE Defense+ Security; 2014 June. International Society for Optics and Photonics.

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

Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and 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 Dec Making. 2008;2(2):140–160.

Project Implicit. [accessed 2017 Sep]. https://www.projectimplicit.net/index.html.

Rau PP, Li Y, Li D. Effects of communication style and culture on ability to accept recommendations from robots. Comp Hum Behav. 2009;25(2):587–595.

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

Rubin RB. The role of context in information seeking and impression formation. Comm Mono. 1977;44(1):81–90.

Rubin RB, Perse EM, Barbato CA. Conceptualization and measurement of interpersonal communication motives. Hum Comm Res. 1988;14(4)602–628.

Ryan RM. Control and information in the intrapersonal sphere: an extension of cognitive evaluation theory. J Pers Soc Psych. 1982;43(3):450.

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.

Selkowitz AR, Lakhmani SG, Larios CN, Chen JY. Agent transparency and the autonomous squad member. Proc Hum Fact Ergon Soc Ann Meet. 2016;60(1):1319–1323.

Singh IL, Molloy R, Parasuraman R. Automation-induced “complacency”: development of the complacency-potential rating scale. The Inter J of Aviation Psychology. 1993;3(2):111–122.

Smith K, Hancock PA. Situation awareness is adaptive, externally directed consciousness. Hum Fact. 1995;37(1):137–148.

Stanton NA, Salmon PM, Rafferty LA, Walker GH, Baber C, Jenkins DP. Human factors methods: a practical guide for engineering and design. Ashgate Publishing, Ltd; 2012.

Stowers K, Kasdaglis N, Rupp MA, Newton OB, Chen JYC, Barnes MJ. The IMPACT of agent transparency on human performance. IEEE Trans Hum-Mach Syst. 2020;50(3):245–253.

Sycara K, Sukthankar G. Literature review of teamwork models. Robotics Institute, Carnegie Mellon University; 2006. Report No.: CMU-RI-TR-06-50.

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

Wickens CD. Multiple resources and performance prediction. Theor Issues Ergon Sci. 2002;3(2):159–177.

Wickens CD, Holland JG. Engineering psychology and human performance. 3rd ed. Upper Saddle River (NJ): Prentice Hall; 2000.

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

Wright JL, Chen JYC, Barnes MJ. Human-automation interaction for multiple robot control: the effect of varying automation assistance and individual differences on operator performance. Ergonomics. 2018;61(8):1033–1045.

Wright JL, Chen JYC, Lakhmani SG. Agent transparency and reliability in human–robot interaction: the influence on user confidence and perceived reliability. IEEE Transactions on Human-Machine Systems. 2020;50(3):254–263. doi: 10.1109/THMS.2019.2925717.

Yi D, Goodrich MA. Supporting task-oriented collaboration in human-robot teams using semantic-based path planning. Proc. SPIE. 2014;9084.

Zhou J. Feedback valence, feedback style, task autonomy, and achievement orientation: interactive effects on creative performance. J Appl Psych. 1998;83(2):261.

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):
- months 1 year 2 years 4 years or more

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

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

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

2) 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

1) 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

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

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

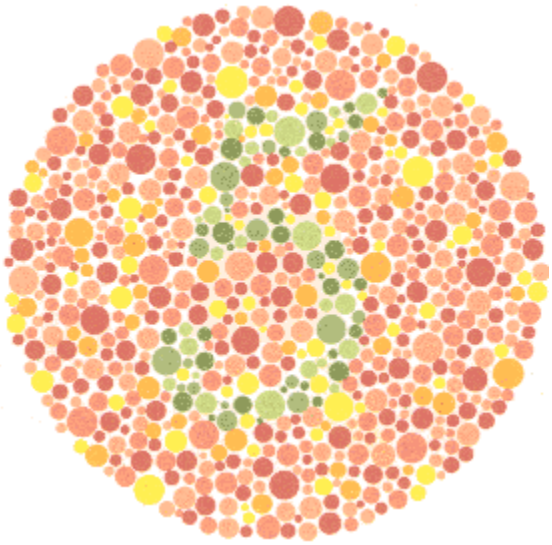
c. How often do you operate a radio-controlled vehicle (car, boat, or plane)?
Daily Weekly Monthly Once or twice a year Never

d. 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 the participant will see during the color vision test. A series of dots compose the number 5 among other dots that are of different colors.



Appendix C. Implicit Association Test

Implicit Association Test (IAT)

This implicit trust measure was adapted from the IAT of Merritt et al.[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, and happy; agony, terrible, horrible, nasty, evil, awful, failure, and 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 and the more difficulty he or she should have when "automation" and "bad" are paired.

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 Fact. 2013;55(3):520–534.

Appendix D. NASA Task Load Index (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).

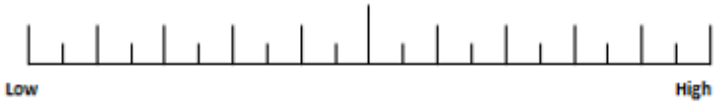
Participant ID: _____

TLX Workload Scale

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

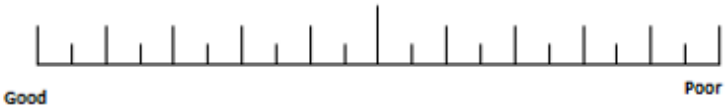
Mental Demand

Physical Demand



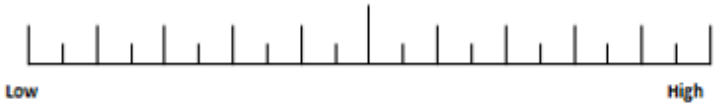
Temporal Demand

Performance



Effort

Frustration



Appendix E. Godspeed Measure

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

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 F. 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:

1) A: When the system is collecting and/or highlighting/filtering information.
2) B: When the system is integrating information, generating predictive displays, and/or presenting its analysis.
3) C: When the system is making decisions and/or selecting actions.
4) 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			*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

12) I am familiar with 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

Appendix G. Reading Span Task (RSPAN)

Participants will be administered a computerized version of the 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 broken down into two sections. First, participants receive practice and second, the participants perform the actual experiment. The practice sessions are further broken down into three sections.

The first practice is simple letter span. They 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 sense verification the participants are given feedback. The reading practice familiarizes participants with the sentence portion of the experiment as well as calculates 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 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 prevents participants from rehearsing the letters when they 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.

---

[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 Verb Learn Verb Beh. 1980; 19(4):450-466.

The experimental trials consist of three trials of each set size with the set sizes ranging from three to six. This totals 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 carefully watch 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 the study.

RSPAN Timing

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

Letter presentation: 800 ms.

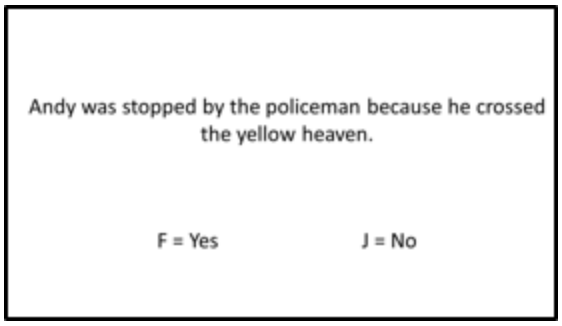
Recall screen: Min = none, Max = 2 min (there is a "Continue" button to move forward faster).

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

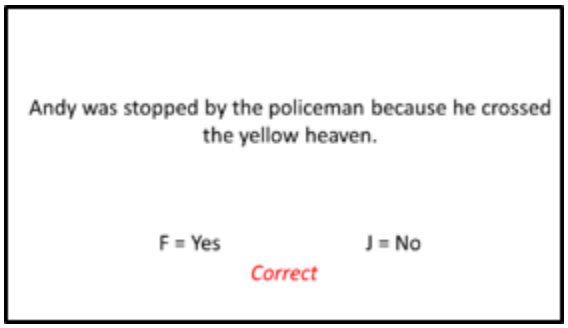
Slide Examples

Ready screen

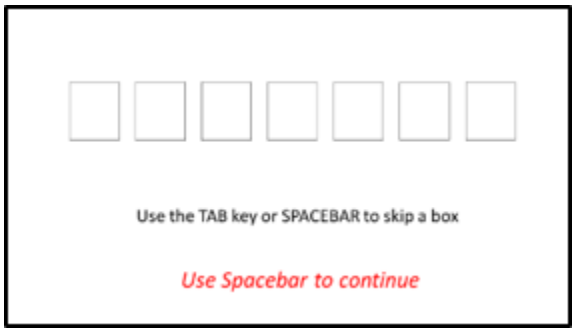
Letter screen



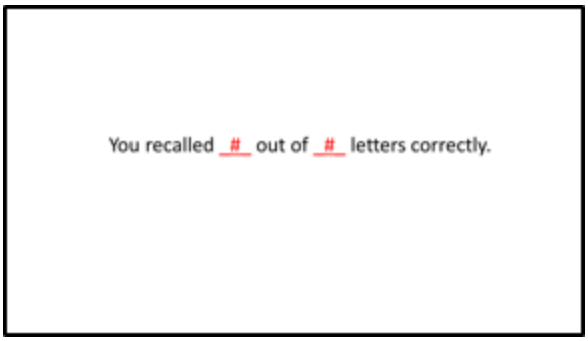
Sentence screen



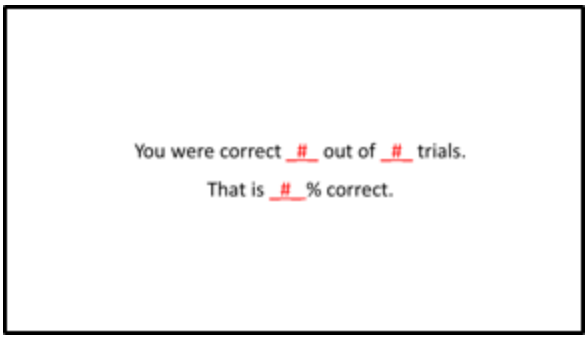
Sentence screen with feedback (for sentence practice only)



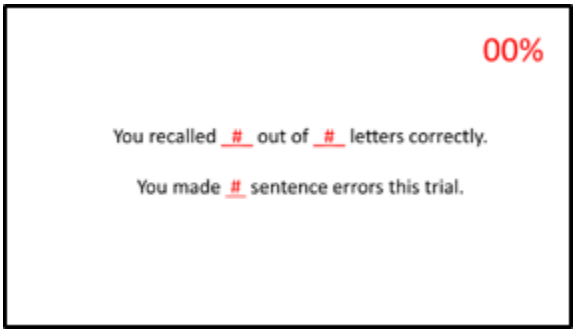
Recall screen; always 7 boxes shown



Feedback screen, letter practice



Feedback screen, sentence practice



Feedback screen, final practice and main experiment

Appendix H. 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 I. Situation Awareness Questions

After each event the simulation paused, the screens would blank, and the following questions were presented to the participants. Participants received 1 point for each correct response.

SCREEN 1

1) What did the robot encounter? (select one response)
    a. Money Cache
    b. Weapon Cache
    c. Information Cache
    d. IED
    e. Intruder
    f. Nothing
2) What is the robot doing? (select one response)
    a. Searching
    b. Documenting
    c. Dealing with Intruders
3) What did you see that could affect your task or the robot's task? (select one response)
    a. Obstacle
    b. Intruder
    c. Person (Distraction)
    d. Obstacle & Intruder
    e. Intruder & Distraction
    f. Obstacle & Distraction
    g. None of these

SCREEN 2

4) Did you encounter a dangerous event? (select one response)
    a. Dangerous Person
    b. Dangerous Vehicle
    c. A dangerous person and a dangerous vehicle
    d. No dangerous event
5) What is the robot's current priority? (select one response)
    a. Preserving Robot Safety

b. Maintaining Information Flow

6) What is your current priority? (select one response)
    c. Preserving Robot Safety
    d. Maintaining Information Flow

SCREEN 3

7) What is the likely outcome of the most recent event you observed in the cordon area? (select one response)
    - o The robot might be damaged
    - o You might be in danger
    - o There's nothing for you to say
    - o Your communication system will lose energy as you use it
    - o You'll be delayed before making a decision

8) Given the most recent event the robot encountered, what is the robot's most relevant projected outcome? (select one response)
    - o It will use energy
    - o It will be delayed
    - o It may be damaged
    - o It may suffer some signal interference

Appendix J. Task Performance Results' Tables

Table J-1 Communications task's descriptive statistics

Measure	Condition		*N*	*M*	*SD*	*SE*	*95% CI*	
							Lower	*Upper*
Correct responses (%)	**HTL**	**NDIR**	17	67.4	25.2	5.37	56.4	78.4
		DIR	15	63.9	18.1	5.72	52.2	75.6
	LTL	**NDIR**	17	68.4	19.9	4.72	58.8	78.1
		DIR	15	69.1	18.9	5.02	58.9	79.4
Response time—correct (s)	**HTL**	**NDIR**	16	4.0	0.7	0.16	3.6	4.5
		DIR	15	3.8	0.6	0.15	3.5	4.2
	LTL	**NDIR**	17	3.8	0.7	0.18	3.5	4.3
		DIR	15	3.8	0.7	0.19	3.4	4.2
Response time—incorrect (s)	**HTL**	**NDIR**	13	3.9	1.2	0.33	3.2	4.7
		DIR	14	3.2	1.0	0.27	2.6	3.8
	LTL	**NDIR**	16	3.1	1.0	0.25	2.6	3.6
		DIR	15	3.3	1.2	0.31	2.6	3.9
Response time—overall (s)	**HTL**	**NDIR**	17	4.0	0.7	0.16	3.6	4.3
		DIR	15	3.6	0.6	0.16	3.3	4.0
	LTL	**NDIR**	17	3.6	0.7	0.18	3.3	4.0
		DIR	15	3.6	0.7	0.18	3.2	4.0

N = number, M = mean, SD = standard deviation, SE = standard error of the mean, CI = confidence interval, HTL = high task level, LTL = low task level, NDIR = nondirective, DIR = directive.

Table J-2 Target identification (ID) task's descriptive statistics

Measure	Condition		*N*	*M*	*SD*	*SE*	*95% CI*	
							Lower	*Upper*
Person correct IDs (%)	**HTL**	**NDIR**	14	60.0	31.9	7.34	44.9	75.1
		DIR	13	77.3	21.7	7.61	61.6	93.0
	LTL	**NDIR**	17	67.9	24.5	5.54	56.6	79.3
		DIR	14	75.0	20.6	6.10	62.5	87.5
Person response time (s)	**HTL**	**NDIR**	14	4.9	2.3	0.53	3.8	6.0
		DIR	13	3.9	1.6	0.56	2.7	5.0
	LTL	**NDIR**	17	4.3	1.2	0.30	3.7	4.9
		DIR	14	3.5	1.3	0.33	2.8	4.2
Person false positives (%)	**HTL**	**NDIR**	17	36.8	35.7	8.06	20.3	53.2
		DIR	15	37.7	30.1	8.58	20.1	55.2
	LTL	**NDIR**	17	21.8	5.6	2.24	17.2	26.3
		DIR	15	16.0	12.1	2.38	11.1	20.9
Dangerous person correct IDs (%)	**HTL**	**NDIR**	17	77.0	22.5	5.01	66.7	87.2
		DIR	15	86.1	18.3	5.33	75.2	97.0
	LTL	**NDIR**	17	88.2	14.7	3.69	80.7	95.8
		DIR	15	92.8	15.7	3.93	84.8	100.8
Dangerous person response time (s)	**HTL**	**NDIR**	17	6.0	2.6	0.58	4.8	7.2
		DIR	15	4.0	2.1	0.61	2.7	5.2
	LTL	**NDIR**	17	4.0	1.9	0.44	3.1	4.9
		DIR	15	4.2	1.8	0.47	3.2	5.1
Dangerous person false positives (%)	**HTL**	**NDIR**	17	39.7	40.2	8.57	22.1	57.1
		DIR	15	38.2	28.7	9.12	19.6	56.8
	LTL	**NDIR**	17	28.2	14.0	3.87	20.3	36.1
		DIR	15	23.9	18.0	4.12	15.5	32.3

Table J-2 Target ID task's descriptive statistics (continued)

Measure	Condition		*N*	*M*	*SD*	*SE*	*95% CI*	
							Lower	*Upper*
Obstacle correct IDs (%)	**HTL**	**NDIR**	17	89.7	12.8	3.02	83.5	95.9
		DIR	15	88.8	12.0	3.22	82.2	95.4
	LTL	**NDIR**	17	91.1	10.5	2.47	86.0	96.1
		DIR	15	94.4	9.8	2.63	89.0	99.8
Obstacle response time (s)	**HTL**	**NDIR**	17	5.4	0.7	0.19	5.0	5.8
		DIR	15	5.2	0.8	0.20	4.8	5.6
	LTL	**NDIR**	17	5.2	0.9	0.20	4.8	5.6
		DIR	15	5.0	0.8	0.22	4.6	5.4
Obstacle false positives (%)	**HTL**	**NDIR**	17	40.2	15.1	3.42	33.2	47.2
		DIR	15	36.5	13.0	3.64	29.1	44.0
	LTL	**NDIR**	17	9.8	10.3	2.32	5.1	14.5
		DIR	15	7.8	8.6	2.47	2.7	12.8
Suspicious vehicle correct IDs (%)	**HTL**	**NDIR**	17	84.1	11.5	3.21	77.6	90.7
		DIR	15	82.3	15.0	3.42	75.4	89.3
	LTL	**NDIR**	17	88.2	11.3	3.01	82.1	94.4
		DIR	15	85.3	13.6	3.20	78.8	91.9
Suspicious vehicle response time (s)	**HTL**	**NDIR**	17	3.5	0.5	0.13	3.2	3.7
		DIR	15	3.3	0.5	0.14	3.0	3.5
	LTL	**NDIR**	17	3.1	0.8	0.17	2.8	3.5
		DIR	15	3.3	0.6	0.18	2.9	3.6
Suspicious vehicle false positives (%)	**HTL**	**NDIR**	17	61.2	15.9	4.51	52.0	70.4
		DIR	15	58.7	21.3	4.80	48.9	68.5
	LTL	**NDIR**	17	66.5	14.1	3.18	60.0	73.0
		DIR	15	58.7	11.9	3.39	51.7	65.6

Table J-3 Target ID task's t-test results, between communication styles, by task load

Measure	Overall			HTL			LTL		
	t(30)	*p*	d_s	*t(30)*	*p*	d_s	*t(30)*	*p*	d_s
Person correct IDs (%)	–0.96	0.345	0.58	1.39	0.174	0.63	0.22	0.825	0.31
Person response time (s)	1.20	0.242	0.43	1.40	0.173	0.54	1.78	0.086	0.64
Person false positives (%)	–0.11	0.211	0.14	0.29	0.774	0.11	0.96	0.345	0.62
Dangerous person correct IDs (%)	–1.23	0.229	0.43	–1.25	0.221	0.44	0.84	0.406	0.30
Dangerous person response time (s)	1.93	0.063	0.68	2.41	0.220	0.85	0.23	0.823	0.08
Dangerous person false positives (%)	0.35	0.730	0.12	0.12	0.909	0.04	0.76	0.453	0.08
Obstacle correct IDs (%)	0.03	0.973	0.01	0.21	0.839	0.07	0.93	0.362	0.33
Obstacle response time (s)	1.04	0.306	0.37	0.86	0.397	0.30	0.74	0.466	0.26
Obstacle false positives (%)	0.94	0.355	0.33	0.74	0.465	0.26	0.60	0.554	0.21
Suspicious vehicle correct IDs (%)	1.19	0.245	0.42	0.38	0.706	0.13	0.66	0.514	0.20
Suspicious vehicle response time (s)	0.19	0.849	0.07	1.17	0.251	0.41	0.58	0.568	0.23
Suspicious vehicle false positives (%)	1.07	0.292	0.38	0.38	0.706	0.14	1.68	0.103	0.59

Appendix K. Functional Trust Survey Results' Tables

Table K-1 Functional trust survey's descriptive statistics

Measure	Condition		*N*	*M*	*SD*	*SE*	*95% CI*	
							Lower	*Upper*
Overall	**HTL**	**NDIR**	17	5.45	0.89	0.22	4.77	5.91
		DIR	15	5.33	0.86	0.22	4.86	5.81
	LTL	**NDIR**	17	5.56	0.93	0.23	5.08	6.04
		DIR	15	5.31	0.67	0.17	4.93	5.68
A	**HTL**	**NDIR**	17	5.50	0.94	0.23	5.01	5.98
		DIR	15	5.68	0.79	0.20	5.25	6.12
	LTL	**NDIR**	17	5.74	0.93	0.23	5.26	6.21
		DIR	15	5.72	0.71	0.18	5.33	6.11
B	**HTL**	**NDIR**	17	5.36	1.06	0.26	4.81	5.90
		DIR	15	5.41	0.93	0.24	4.89	5.92
	LTL	**NDIR**	17	5.40	0.94	0.23	4.92	5.88
		DIR	15	5.47	0.86	0.22	5.00	5.95
C	**HTL**	**NDIR**	17	5.08	1.36	0.33	4.38	5.78
		DIR	15	5.02	0.96	0.25	4.49	5.56
	LTL	**NDIR**	17	5.28	1.11	0.27	4.71	5.85
		DIR	15	5.06	0.89	0.23	4.56	5.55
D	**HTL**	**NDIR**	17	5.52	1.09	0.26	4.63	6.10
		DIR	15	5.17	1.16	0.30	4.53	5.81
	LTL	**NDIR**	17	5.53	1.14	0.28	4.95	6.11
		DIR	15	5.27	0.77	0.20	4.85	5.70

N = number, M = mean, SD = standard deviation, SE = standard error of the mean, CI = confidence interval, HTL = high task level, LTL = low task level, NDIR = nondirective, DIR = directive.

Table K-2 Functional trust survey, between communication styles by task-load t-test results

	HTL			LTL		
Measure	***t(30)***	***p***	d_s	***t(30)***	***p***	d_s
Overall	0.37	0.714	0.13	0.89	0.382	0.31
A	–0.61	0.547	–0.22	0.06	0.950	0.02
B	–0.13	0.894	–0.05	–0.22	0.827	–0.08
C	0.15	0.886	0.05	0.62	0.538	0.22
D	0.89	0.379	0.32	0.74	0.465	0.26

Table K-3 Functional trust survey, within communication styles between task-load t-test results

	NDIR			DIR		
Measure	***t(16)***	***p***	d_s	***t(14)***	***p***	d_s
Overall	–1.07	0.299	–0.13	0.17	0.869	0.04
A	–2.31	0.034	–0.26	–0.18	0.858	–0.04
B	–0.33	0.743	–0.04	–0.33	0.745	–0.07
C	–1.04	0.315	–0.16	–0.16	0.872	–0.04
D	–0.05	0.959	–0.01	–0.45	0.658	–0.11

Appendix L. Workload Results' Tables

Table L-1 NASA task load index (TLX) scores' descriptive statistics

Measure	Condition		*N*	*M*	*SD*	*SE*	*95% CI*	
							Lower	Upper
Global–unweighted	**HTL**	**NDIR**	17	51.62	12.48	3.03	45.21	58.04
		DIR	15	52.77	15.83	4.09	44.01	61.54
	LTL	**NDIR**	17	52.86	11.62	2.82	46.89	58.84
		DIR	15	53.17	12.81	3.31	46.07	60.26
Mental demand	**HTL**	**NDIR**	17	75.88	15.13	3.67	68.10	83.66
		DIR	15	68.33	21.77	5.62	56.28	80.39
	LTL	**NDIR**	17	76.76	12.24	2.97	70.47	83.06
		DIR	15	73.00	12.93	3.34	65.84	80.16
Physical demand	**HTL**	**NDIR**	17	10.29	13.75	3.33	3.23	17.36
		DIR	15	19.33	18.70	4.83	8.98	29.69
	LTL	**NDIR**	17	7.35	9.70	2.35	2.36	12.34
		DIR	15	18.67	17.27	4.46	9.11	28.23
Temporal demand	**HTL**	**NDIR**	17	65.29	24.27	5.36	53.92	76.66
		DIR	15	64.33	20.86	5.39	52.78	75.89
	LTL	**NDIR**	17	64.71	20.76	5.89	52.23	77.18
		DIR	15	61.33	23.49	6.06	48.33	74.34
Effort	**HTL**	**NDIR**	17	63.53	11.56	2.80	57.59	69.47
		DIR	15	68.00	9.60	2.48	62.68	73.32
	LTL	**NDIR**	17	65.29	20.04	4.86	54.99	75.60
		DIR	15	59.67	16.20	4.18	50.70	68.64
Frustration	**HTL**	**NDIR**	17	46.47	26.21	6.36	33.00	59.94
		DIR	15	49.00	22.54	5.82	36.52	61.48
	LTL	**NDIR**	17	51.18	26.61	6.45	37.50	64.86
		DIR	15	51.33	24.09	6.22	37.99	64.67
Performance	**HTL**	**NDIR**	17	47.35	21.73	5.27	36.18	58.53
		DIR	15	49.00	24.36	6.29	35.51	62.49
	LTL	**NDIR**	17	43.24	26.92	6.53	29.39	57.08
		DIR	15	55.67	19.99	5.16	44.60	66.74

N = number, M = mean, SD = standard deviation, SE = standard error of the mean, CI = confidence interval, HTL = high task level, LTL = low task level, NDIR = nondirective, DIR = directive.

Table L-2 NASA-TLX scores, between communication styles by task-load t-test results

Measure	HTL			LTL		
	t(30)	*p*	d_s	*t(30)*	*p*	d_s
Global–unweighted	–0.23	0.820	–0.08	–0.07	0.945	–0.02
Mental demand	1.15	0.259	0.41	0.85	0.404	0.30
Physical demand	–1.57	0.127	–0.56	–2.32 [a]	0.027	0.82
Temporal demand	0.13	0.901	0.04	0.40	0.693	0.14
Effort	–1.18	0.247	–0.42	0.87	0.393	0.31
Frustration	–0.29	0.773	–0.10	–0.02	0.986	–0.01
Performance	–0.20	0.841	–0.07	–1.47	0.153	–0.52

[a] denotes <0.05

Table L-3 NASA-TLX scores, within communication styles between task-load t-test results

Measure	NDIR			DIR		
	t(16)	*p*	d_s	*t(14)*	*p*	d_s
Global–unweighted	–0.38	0.711	–0.10	–0.15	0.886	–0.03
Mental demand	–0.23	0.824	–0.06	–0.93	0.370	–0.26
Physical demand	1.98 [a]	0.066	0.25	0.26	0.796	0.04
Temporal demand	0.09	0.931	0.03	0.76	0.458	0.14
Effort	–0.39	0.702	–0.11	2.61 [a]	0.020	0.63
Frustration	–0.74	0.472	–0.18	–0.36	0.722	–0.10
Performance	0.50	0.621	0.17	–0.94	0.364	–0.30

[a] denotes <0.08

Table L-4 Eye-tracking measures' descriptive statistics

Measure	Condition		*N*	*M*	*SD*	*SE*	*95% CI*	
							Lower	Upper
Fixation count (average)	**HTL**	**NDIR**	17	866.12	101.76	24.68	813.80	918.44
		DIR	15	820.67	73.38	18.95	780.03	861.30
	LTL	**NDIR**	17	825.18	62.11	15.06	793.24	857.11
		DIR	15	834.00	108.11	27.91	774.13	893.87
Blink duration (s)	**HTL**	**NDIR**	17	0.47	0.27	0.07	0.33	0.61
		DIR	15	0.27	0.07	0.02	0.24	0.31
	LTL	**NDIR**	17	0.44	0.21	0.05	0.33	0.54
		DIR	15	0.29	0.06	0.02	0.25	0.32
Pupil diameter (mm)	**HTL**	**NDIR**	17	3.548	0.519	0.126	3.282	3.815
		DIR	15	3.434	0.490	0.127	3.163	3.706
	LTL	**NDIR**	17	3.474	0.520	0.126	3.207	3.742
		DIR	15	3.398	0.001	0.181	3.010	3.786

Table L-5 Eye-tracking measures, between communication styles by task-load t-test results

Measure	HTL			LTL		
	t(30)	*p*	d_s	*t(30)*	*p*	d_s
Pupil diameter (mm)	1.43	0.163	0.51	–0.29	0.776	–0.10
Blink duration (seconds)	2.69 [a]	0.011	0.95	2.73 [a]	0.011	0.97
Pupil diameter (mm)	0.64	0.530	0.23	0.35	0.728	0.12

[a] denotes <0.05

Table L-6 Eye-tracking measures, within communication styles between task-load t-test results

Measure	NDIR			DIR		
	t(16)	***p***	d_s	***t(14)***	***p***	d_s
Fixation count (average)	–1.71	0.107	–0.49	0.50	0.626	0.14
Blink duration (s)	–0.40	0.697	–0.14	0.69	0.499	0.19
Pupil diameter (mm)	–2.79 [a]	0.013	–0.14	–0.47	0.645	–0.06

[a] denotes <0.05

Appendix M. Situation Awareness (SA) Query Results' Tables

Table M-1 SA query scores' descriptive statistics

Measure	Condition		*N*	*M*	*SD*	*SE*	95% CI Lower	95% CI Upper
Overall	**HTL**	**NDIR**	18	23.61	6.40	1.54	20.48	26.74
		DIR	16	24.63	6.65	1.63	21.31	27.94
	LTL	**NDIR**	18	22.78	7.55	1.59	19.54	26.01
		DIR	16	22.94	5.67	1.68	19.51	26.37
Level 1	**HTL**	**NDIR**	18	8.50	2.04	0.53	7.43	9.57
		DIR	16	8.88	2.45	0.56	7.74	10.01
	LTL	**NDIR**	18	8.33	2.81	0.59	7.12	9.54
		DIR	16	8.25	2.14	0.63	6.97	9.53
Level 2	**HTL**	**NDIR**	18	8.50	2.04	0.53	7.43	9.57
		DIR	16	8.88	2.45	0.56	7.74	10.01
	LTL	**NDIR**	18	8.33	2.81	0.59	7.12	9.54
		DIR	16	8.25	2.14	0.63	6.97	9.53
Level 3	**HTL**	**NDIR**	18	6.39	1.88	0.44	5.48	7.30
		DIR	16	6.88	1.89	0.47	5.91	7.84
	LTL	**NDIR**	18	6.11	2.00	0.42	5.26	6.97
		DIR	16	6.44	1.50	0.45	5.53	7.35

N = number, M = mean, SD = standard deviation, SE = standard error of the mean, CI = confidence interval, HTL = high task level, LTL = low task level, NDIR = nondirective, DIR = directive.

Table M-2 T-test results for SA query score comparison between communication styles by task load level

	HTL			LTL		
Measure	***t(32)***	***p***	d_s	***t(32)***	***p***	d_s
Overall	–0.45	0.654	–0.16	–0.07	0.945	–0.02
L1	–0.49	0.629	–0.17	0.10	0.924	0.03
L2	–0.49	0.629	–0.17	0.10	0.924	0.03
L3	–0.75	0.459	–0.26	–0.53	0.598	–0.18

Table M-3 T-test results for SA query score comparison between task load levels, within communication styles

	NDIR			DIR		
Measure	***t(17)***	***p***	d_s	***t(15)***	***p***	d_s
Overall	–0.32	0.751	–0.12	–0.66	0.518	–0.27
L1	–0.19	0.851	–0.07	–0.67	0.516	–0.27
L2	–0.29	0.776	–0.11	–0.67	0.516	–0.27
L3	–0.40	0.694	–0.14	–0.62	0.545	–0.26

Appendix N. Godspeed Measures' Tables

Table N-1 Godspeed survey's descriptive statistics

Measure	Condition		*N*	*M*	*SD*	*SE*	95% CI	
							Lower	Upper
Anthropomorphism	**HTL**	**NDIR**	17	2.80	0.49	0.12	2.55	3.05
		DIR	15	2.91	0.85	0.22	2.44	3.38
	LTL	**NDIR**	17	3.06	0.60	0.15	2.75	3.37
		DIR	15	3.07	0.57	0.15	2.75	3.38
Animacy	**HTL**	**NDIR**	17	3.12	0.56	0.14	2.83	3.41
		DIR	15	3.25	0.48	0.13	2.98	3.52
	LTL	**NDIR**	17	3.49	0.49	0.12	3.23	3.74
		DIR	15	3.48	0.41	0.11	3.25	3.71
Likeability	**HTL**	**NDIR**	17	3.51	0.86	0.21	3.06	3.95
		DIR	15	3.75	0.89	0.23	3.25	4.24
	LTL	**NDIR**	17	3.74	0.62	0.15	3.42	4.06
		DIR	15	4.00	0.68	0.18	3.62	4.38
Perceived intelligence	**HTL**	**NDIR**	17	4.14	0.68	0.16	3.79	4.49
		DIR	15	4.15	0.52	0.13	3.86	4.43
	LTL	**NDIR**	17	4.20	0.66	0.16	3.86	4.54
		DIR	15	4.19	0.32	0.08	4.01	4.36
Perceived safety	**HTL**	**NDIR**	17	3.31	0.82	0.20	2.89	3.74
		DIR	15	3.38	0.58	0.15	3.06	3.70
	LTL	**NDIR**	17	3.39	0.60	0.15	3.08	3.70
		DIR	15	3.47	0.45	0.12	3.22	3.72

N = number, M = mean, SD = standard deviation, SE = standard error of the mean, CI = confidence interval, HTL = high task level, LTL = low task level, NDIR = nondirective, DIR = directive.

Table N-2 Godspeed survey, between communication styles by task-load t-test

Measure	HTL			LTL		
	t(30)	*p*	d_s	*t(30)*	*p*	d_s
Anthropomorphism	–0.44	0.663	–0.16	–0.04	0.970	–0.01
Animacy	–0.68	0.500	–0.24	0.07	0.948	0.02
Likeability	–0.78	0.444	–0.27	–1.12	0.270	–0.40
Perceived intelligence	–0.03	0.980	–0.01	0.07	0.943	0.03
Perceived safety	–0.25	0.803	–0.09	–0.39	0.698	-0.14

Table N-3 Godspeed survey, within communication styles between task-load paired t-test results

Measure	NDIR			DIR		
	t(16)	*p*	d_s	*t(14)*	*p*	d_s
Anthropomorphism	–1.95	0.069	–0.47	–1.46	0.166	–0.22
Animacy	–2.49	0.024	–0.70	–2.40	0.031	–0.51
Likeability	–1.78	0.094	–0.31	–2.24	0.042	–0.32
Perceived intelligence	–0.45	0.658	–0.09	–0.34	0.742	–0.09
Perceived safety	–0.47	0.642	–0.11	–0.84	0.413	–0.17

Appendix O. Individual-Difference Factors' Tables

Table O-1 Individual-difference factors correlations with communications task's performance measures

		IAT		WMC		PAC	
		r	*p*	*r*	*p*	*r*	*p*
Communications responses % correct	**NDIR**	0.166	0.525	0.371	0.143	–0.037	0.887
	DIR	–0.484	0.067	0.429	0.111	0.146	0.605
Communications correct responses RT	**NDIR**	–0.012	0.964	0.165	0.527	0.055	0.834
	DIR	0.230	0.409	–0.084	0.767	–0.314	0.254
Communications incorrect responses RT	**NDIR**	–0.256	0.321	–0.101	0.699	–0.273	0.289
	DIR	0.184	0.511	–0.204	0.466	–0.001	0.996
Communications overall RT		0.071	0.787	0.017	0.947	–0.046	0.860
	DIR	0.132	0.640	0.018	0.950	–0.287	0.300

IAT = Implicit Association Test, WMC = working memory capacity, PAC = perceived attentional control, NDIR = nondirective, DIR = directive, RT = response time.

Table O-2 Individual-difference factors' correlations with identification task's performance measures

		IAT		WMC		PAC	
		r	*p*	*r*	*p*	*r*	*p*
Person correctly identified %	**NDIR**	0.177	0.496	0.252	0.329	–0.009	0.973
	DIR	0.167	0.552	0.392	0.148	–0.069	0.807
Person correctly identified RT	**NDIR**	–0.327	0.200	0.170	0.514	0.051	0.845
	DIR	0.598 [a]	0.019	–0.002	0.994	–0.159	0.570
Person false positives %	**NDIR**	0.437	0.080	0.093	0.722	0.079	0.764
	DIR	0.550 [a]	0.034	0.029	0.918	–0.137	0.626
Dangerous person correctly identified %	**NDIR**	–0.031	0.907	0.079	0.763	–0.027	0.918
	DIR	0.176	0.530	0.372	0.172	–0.440	0.101
Dangerous person correctly identified RT	**NDIR**	0.006	0.982	–0.046	0.860	–0.280	0.277
	DIR	0.294	0.287	–0.424	0.115	–0.268	0.303
Dangerous person false positives %	**NDIR**	0.255	0.324	0.290	0.259	–0.365	0.149
	DIR	–0.082	0.770	0.184	0.512	–0.272	0.326
Obstacle correctly identified %	**NDIR**	0.007	0.979	–0.176	0.500	0.145	0.579
	DIR	–0.240	0.388	0.240	0.389	0.285	0.303
Obstacle correctly identified RT	**NDIR**	0.122	0.641	0.139	0.594	–0.103	0.695
	DIR	–0.137	0.627	–0.328	0.233	–0.323	0.241
Obstacle false positives %	**NDIR**	–0.048	0.854	0.039	0.881	0.117	0.655
	DIR	0.216	0.439	0.040	0.888	–0.218	0.434
Suspicious vehicle correctly identified %	**NDIR**	0.131	0.615	0.518 [a]	0.033	–0.489 [a]	0.046
	DIR	–0.334	0.223	0.631 [a]	0.012	0.237	0.396
Suspicious vehicle correctly identified RT	**NDIR**	–0.257	0.320	0.090	0.732	0.243	0.348
	DIR	0.489	0.064	–0.268	0.334	–0.119	0.673
Suspicious vehicle false positives %	**NDIR**	–0.072	0.784	0.450	0.070	–0.319	0.212
	DIR	–0.306	0.267	0.586 [a]	0.022	0.321	0.243

Note: p-value is two-tailed.
[a] denotes <0.05

Table O-3 Individual-difference factors' correlations with trust survey's scores

		IAT		WMC		PAC	
		r	***p***	***r***	***p***	***r***	***P***
Overall trust	**NDIR**	–0.225	0.386	0.395	0.117	–0.123	0.638
	DIR	0.341	0.213	0.088	0.754	–0.201	0.473
A	**NDIR**	–0.260	0.313	0.157	0.548	–0.044	0.866
	DIR	0.150	0.593	0.211	0.451	0.011	0.968
B	**NDIR**	–0.325	0.203	0.208	0.424	0.106	0.686
	DIR	0.289	0.296	–0.013	0.963	0.137	0.626
C	**NDIR**	–0.210	0.418	0.241	0.352	–0.062	0.814
	DIR	0.380	0.162	0.116	0.680	–0.254	0.360
D	**NDIR**	–0.258	0.318	0.289	0.261	–0.076	0.770
	DIR	0.255	0.360	0.227	0.415	–0.141	0.615

Table O-4 Individual-difference factors' correlations with cognitive-workload measures

		IAT		WMC		PAC	
		r	***p***	***r***	***p***	***R***	***p***
Global NASA-TLX score	**NDIR**	–0.160	0.539	–0.160	0.539	–0.447	0.072
	DIR	–0.197	0.483	0.085	0.765	0.025	0.929
Mental demand	**NDIR**	–0.226	0.382	–0.300	0.242	–0.095	0.716
	DIR	–0.224	0.422	–0.010	0.973	–0.008	0.976
Physical demand	**NDIR**	–0.006	0.983	0.071	0.788	–0.179	0.492
	DIR	0.150	0.594	–0.259	0.352	–0.276	0.319
Temporal demand	**NDIR**	–0.047	0.857	–0.284	0.269	–0.244	0.346
	DIR	–0.039	0.890	0.056	0.844	0.235	0.399
Effort	**NDIR**	0.068	0.794	–0.312	0.223	–0.115	0.660
	DIR	–0.201	0.472	0.246	0.377	0.119	0.673
Frustration	**NDIR**	–0.219	0.399	–0.272	0.292	–0.237	0.359
	DIR	–0.214	0.444	0.206	0.461	–0.139	0.622
Performance	**NDIR**	–0.244	0.345	–0.038	0.883	0.040	0.878
	DIR	–0.397	0.143	0.183	0.513	0.109	0.699
Fixation count	**NDIR**	0.152	0.561	0.302	0.238	–0.571[a]	0.017
	DIR	–0.133	0.635	–0.432	0.108	–0.021	0.940
Blink duration	**NDIR**	0.013	0.961	–0.078	0.767	0.250	0.334
	DIR	0.294	0.287	0.230	0.409	0.249	0.372
Pupil diameter	**NDIR**	0.353	0.164	0.219	0.399	–0.093	0.722
	DIR	–0.157	0.576	0.280	0.312	–0.355	0.194

Note: p-value is two-tailed.
[a] denotes <0.05

Table O-5 Individual-difference factor correlations with situation awareness (SA) scores, by SA level.

		IAT		WMC		PAC	
		r	*p*	*r*	*p*	*r*	*p*
SA Level 1	**NDIR**	–0.007	0.978	–0.036	0.891	0.303	0.238
	DIR	–0.019	0.947	0.333	0.225	0.224	0.422
SA Level 2	**NDIR**	–0.007	0.978	–0.036	0.891	0.303	0.238
	DIR	–0.019	0.947	0.333	0.225	0.224	0.422
SA Level 3	**NDIR**	–0.154	0.555	–0.066	0.802	0.226	0.383
	DIR	–0.119	0.672	0.184	0.510	0.130	0.644

Table O-6 Individual-difference factor correlations with Godspeed survey's scores

		IAT		WMC		PAC	
		r	*p*	*r*	*p*	*r*	*p*
Anthropomorphism	**NDIR**	0.153	0.558	0.301	0.240	–0.373	0.140
	DIR	0.116	0.682	–0.118	0.676	–0.493	0.062
Animacy	**NDIR**	0.019	0.942	0.372	0.142	–0.395	0.116
	DIR	0.217	0.437	–0.205	0.464	–0.392	0.148
Likeability	**NDIR**	–0.283	0.270	0.235	0.364	–0.222	0.392
	DIR	0.290	0.294	–0.306	0.268	–0.287	0.299
Perceived intelligence	**NDIR**	–0.174	0.504	0.253	0.326	–0.120	0.646
	DIR	–0.005	0.985	0.047	0.867	–0.368	0.177
Perceived safety	**NDIR**	–0.151	0.564	0.104	0.690	–0.223	0.391
	DIR	0.458	0.086	0.098	0.727	–0.241	0.386

List of Symbols, Abbreviations, and Acronyms

ANOVA	analysis of variance
ARL	Army Research Laboratory
d_s	Cohen's *d*, standardized (uses pooled variance)
DEVCOM	US Army Combat Capabilities Development Command
DIR	directive
GE	gaming experience
GQS	Godspeed Questionnaire Series
HAT	human–autonomy teaming
HTL	high task load
IAT	Implicit Association Test
ICM	Interpersonal Communication Motive
ID	individual difference
IR	infrared
LTL	low task load
M	mean
Mdn	median
NASA	National Aeronautics and Space Administration
NDIR	nondirective
PAC	perceived attentional control
POV	point of view
p_{single}	p-value, single-tailed
RSPAN	Reading Span Task
RT	response time
SA	situation awareness
SAGAT	Situation Awareness Global Assessment Technique
SAT	Situation-awareness-based Agent Transparency
SD	standard deviation

SE	standard error of the mean
TLX	task load index
UCF	University of Central Florida
WMC	working memory capacity

1 (PDF) DEFENSE TECHNICAL INFORMATION CTR DTIC OCA

1 (PDF) DEVCOM ARL FCDD RLD DCI TECH LIB

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

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

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

14 (PDF) DEVCOM ARL FCDD RLH J LANE Y-S CHEN P FRANASZCZUK K MCDOWELL FCDD RLH F J GASTON K OIE FCDD RLH FA AW EVANS G BOYKIN FCDD RLH FB J GARCIA (A) H ROY FCDD RLH FC J TOURYAN (A) T ROHALY FCDD RLH FD A MARATHE JL WRIGHT S LAKHMANI