Effects of Trust Calibration on Human-Machine Team Performance in Operational Environments

Beth M. Hartzler, Sandro Scielzo,
Alvin Abraham, Rachel Wong
CAE USA
Arlington, TX

{Beth.Hartzler, Sandro.Scielzo, Alvin.Abraham,
Rachel.Wong}@caemilusa.com

Spencer Kohn
Perceptronics Solutions, Inc.
Falls Church, VA

SpencerK@PercSolutions.com

ABSTRACT

Measuring mission-critical trust between human operators and collaborative synthetic teammates is a DoD priority for achieving third-offset goals, accelerating automation design and training for human-machine teams (HMTs), and supporting next generation multi-domain warfare. Achieving proper trust calibration has long been a primary mechanism by which HMT performance could be maximized by avoiding system distrust and over-trust, but this requires real-time trust assessment. The current study establishes the relationship between HMT trust, workload, and performance in a Search and Rescue (SAR) paradigm where human operators supervise intelligent Unmanned Air Vehicle (UAV) assets in a constructive synthetic environment. A novel trust measure was developed and piloted in this experiment to precisely measure subjective trust variations across time and in conjunction with target task elements. Twenty-eight participants, including UAV operators and novices, participated in a rigorously controlled, within-subjects experiment involving four SAR missions. Workload was manipulated by alternating the number of UAVs to supervise in each mission, and we sought to influence trust levels by varying the quality of flight path recommendations provided. Trust was assessed via our novel measure in addition to established metrics. Results demonstrated that the novel trust measure was effective in capturing participants' real-time confidence in the recommendations provided. We identified significant relationships between behavioral trust measures of compliance, verification, and rejection at different levels of trust, such that behaviors indicating high trust were more common when trust was reported to be high using the novel measure. Increased task workload was also associated with significantly poorer SAR outcomes. This experiment is unique as it provides a foundation for a real-time self-report measure of trust that can be directly compared to concurrent physiological measures. Study findings further discuss intervention techniques to maintain proper trust calibration in operational environments.

ABOUT THE AUTHORS

Beth M. Hartzler is a Senior Research Scientist in the Defense and Security division of CAE USA. Dr. Hartzler has 10 years of experience working in the defense industry, currently in conjunction with the Operational Learning Sciences Branch of the 711 Human Performance Wing/Air Force Research Laboratory.

Sandro Scielzo is a Human Systems Technical Authority and Learning Science Fellow at CAE USA. Dr. Scielzo has over 20 years of experience researching next generation training solutions for military and commercial applications. His current focus is on developing a virtual test range for validating human-machine team technologies.

Spencer Kohn is the Director of Human Factors Research at Perceptronics Solutions; his emphasis areas include calibrating trust in mixed human-automation teams. Dr. Kohn received his Ph.D. in Psychology, Human Factors and Applied Cognition from George Mason University.

Alvin Abraham is a Simulation & Modeling engineer in the Defense and Security division of CAE USA. Alvin has 3 years of industry experience, currently supporting the Human Systems Research & Development portfolio.

Rachel Wong is a Research Scientist with CAE USA, Defense and Security. Rachel received her master's degree in Experimental Psychology from Florida Atlantic University in 2023 and has 3 years of industry experience.

Effects of Trust Calibration on Human-Machine Team Performance in Operational Environments

Beth M. Hartzler, Sandro Scielzo,
Alvin Abraham, Rachel Wong
CAE USA
Arlington, TX

{Beth.Hartzler, Sandro.Scielzo, Alvin.Abraham,
Rachel.Wong}@caemilusa.com

Spencer Kohn
Perceptronics Solutions, Inc.
Falls Church, VA

SpencerK@PercSolutions.com

INTRODUCTION

Human-Machine Teaming

Human-machine teaming (HMT) has become increasingly vital as advanced technologies, such as artificial intelligence and autonomous systems, assume a growing role in aiding operational activities. By incorporating machines into human-centric tasks, we can reduce the cognitive burden on human operators and allow them to focus on higher-level decision-making tasks. However, balancing the need for relevant information with the potential for overwhelming distractions is crucial in maintaining operator attention and an appropriate level of workload. High workload may lead the operator to rely more heavily on a system that does not warrant an appropriate level of trust. Trust is fundamental in any type of team and it must be calibrated to maximize performance efficacy, even in the face of increasing stressors and distractions. Thus, one means of reinforcing warfare technologies is to aid operators in maintaining situational awareness (SA) while ensuring appropriate trust in the system.

The potential for cognitive workload mitigation through effective HMT is particularly promising in the control and tasking of drone swarms, when multiple Unmanned Aerial Vehicles (UAVs) are utilized to accomplish a common goal (Parnell et al., 2022). This technology has been increasingly investigated and adopted by military organizations due to the potential for intensifying warfighter lethality and limiting human endangerment in contested or unstable environments. The issue remains however that an operator can simultaneously interact with only a limited number of entities before experiencing cognitive overload. According to previous research, four UAVs is the largest number of drones an operator could actively control at once without creating excessive wait times between actions or losing SA (Hocraffer & Nam, 2017), thereby contributing to mission degradation. Furthermore, any decrease in the reliability of synthetic teammates contributes to a loss of the human's trust in the system, an effect compounded as the number of drones being operated simultaneously increases (Cummings et al., 2006).

Measuring Trust in HMTs

High-performing HMTs rely on appropriately tasking humans and machines based on their respective strengths: automation is consistent, fast, and accurate, while humans are flexible and creative. Humans often take a management role in HMTs, utilizing automation to the degree that it is helpful while monitoring it for errors. This oversight requires trust in the automated teammates, especially in DoD drone swarm tasks such as SAR and reconnaissance, where outcomes are uncertain and may affect the balance between life and death (Lee & See, 2004; Mayer et al., 1995). An appropriate level of trust is thus crucial to ensure the human does not disuse the system and take on too much workload themselves, or misuse the system and rely on imprecise automation (Parasuraman & Riley, 1997). Appropriate trust minimizes excess workload, reduces opportunities for errors, increases overall team efficiency, and generally enables a higher level of performance (see Scielzo & Kocak, 2021).

Given the evident importance of this construct, HMT developers must understand their user's degree of trust in automation so that they can help calibrate that trust via transparency (Lyons et al., 2017) and enhance predictions of performance. Trust in human-machine systems has been a focus of measurement for decades, with Lee & Moray's (1991; 1992) surveys of trust in supervisory control leading the shift into focusing on automation. In the intervening three decades, measures of trust in automation have expanded to include behavioral and psychological measures and more targeted self-report questionnaire batteries (see Kohn et al., 2021 for a review). Behavioral measures of trust capture the participants' patterns of behavior when teaming with automation, such as their compliance with recommendations (Meyer, 2004; de Visser et al., 2016) and reliance in decision-making (Muir, 1994; Rice, 2009; Wijnen et al., 2017), which are ideal for unobtrusively capturing the outcomes of trust on risk-taking behavior.

Similarly, physiological measures have been investigated as a means of capturing biological indicators of trust during interactions with automation. These methods tend to be more exploratory, but include the use of eye tracking (Hergeth et al., 2016; Tenhundfeld et al., 2019) and monitoring electrical activity in the brain to capture reactions to automation actions (de Visser et al., 2018; Desmet et al., 2014; Goodyear et al., 2017). Despite the substantial promise demonstrated, these methods are not frequently used due to equipment and training costs, as well as the obtrusiveness of the hardware. Much more common are self-report measures of trust, which prompt participants to report their own behaviors, beliefs, attitudes, or intentions in a structured manner. These surveys are substantially more flexible than other measures as they can be edited or manipulated to fit the required context.

The sheer abundance of self-report measures also facilitates their popularity. Lee & Moray's early work (1992; 1994) focused on a small set of two or four questions about trust and self-confidence, while more recent surveys have expanded and specified the questions to decompose trust into different components. Examples include Jian et al.'s (2000) 12-item checklist focused on trust and distrust in automation, Wojton et al.'s (2020) nine-item questionnaire regarding system understanding versus performance, and Malle and Ullman's (2021) 16-item scale distinguishing trust in performance and morals. While surveys differ in content and number of items, the vast majority are designed to be administered between experimental blocks or at the end of a session. Self-report measures have generally increased in quality and specificity, yet many suffer from at least one of the following problems: 1) they do not relate meaningfully to existing trust models; 2) they have not yielded convergent validity showing adequate construct operationalization (Kohn et al., 2021); and 3) they are difficult to deploy during experiments.

These problems often manifest when attempting to validate behavioral or physiological measures. Those metrics tend to capture different facets of trust than most existing self-report measures (see Kohn et al., 2021 for a model-based discussion). Viewed through the lens of popular trust models (Mayer et al., 1995; Lee & See, 2004), most self-report measures of trust such as the Checklist for Trust (Jian et al., 2000) predominately capture factors of perceived trustworthiness, information assimilation, and belief formation. Conversely, objective behavioral measures such as compliance (Lee & Moray, 1994) capture reliance actions that are influenced by trust intent and external factors that include the perception of risk. While both capture a construct that is generally referred to as "Trust", they are influenced by different factors. Additionally, these measures are captured at different points in time. Behavioral and physiological measures of trust are captured in real-time during the experiment, while the obtrusive application of most self-report measures mandates pausing the experiment or applying the measure between blocks: participants' level of trust is likely to change between these two capture times.

Current Study

A direct comparison between most objective real-time measures and self-report trust ratings is theoretically inappropriate (see Kohn et al., 2021) and unlikely to correlate well. We believe that this measurement gap has a resolution: a self-report measure applied concurrently with behavioral or physiological measures, capturing similar trust attitudes, with minimal disruption to the primary experiment. The basis and methodology of this concurrent online numerical trust measure is described in our methods section below. It was expected to: 1) provide a light-weight alterative to standard trust measures in scenarios that don't allow pausing the stimuli to administer a survey; 2) capture factors of trust that more closely align with existing behavioral measures, and thus enhance convergent validity; 3) aid in the validation of other novel real-time trust measures. The focus of the current effort is to assess the suitability of an online, explicit measure to capture participants' trust in flight path recommendations throughout a series of SAR mission simulations. Our overarching thesis was that operators would perform behaviors indicative of high trust states more often when they have reported high trust using the Continuous Online Numerical Score (CONS).

METHOD

Participants

Participants in this study represented a convenience sample of 31 adults in the local area, but data from three were excluded from analysis because the participants failed to comply with guidelines of the task. These participants were recruited without regard to age (M = 36.75, SD = 9.91) or gender (male = 27). The majority reported having a college degree, most commonly a bachelor's degree (n = 15), though 10 had earned an advanced degree (e.g., MA, MS). The remaining participants had an associate degree, technical certification, or a high school diploma (n = 1 each).

Within this sample, 39.3% (n = 11) reported a history of military service, either active duty or reserve. Most had served with either the Air Force or Navy (n = 5 each), while the last participant reported serving in the Army. Slightly less than half had aviation experience (n = 13), though only five reported experience as either a pilot or sensor operator for UAVs. Five participants also reported experience with either C2 (Command and Control) or C3 (Command, Control, and Communication) systems. Additionally, 71.4% reported spending at least 1 hour playing video games in an average week. Comparisons on objective measures of performance (e.g., proportion of survivors found and supplies delivered) revealed no significant differences between participants with regard to prior UAV experience or average time spent playing video games (all ps > .10), so these differences were not analyzed further.

Study Design

The study was designed as a 2x2 within-subjects experiment, manipulating workload (low | high) and recommendation quality (good | poor) to create four mission profiles, each lasting 12 minutes. All participants completed the same four missions but in counter-balanced order. Workload was manipulated by varying the number of drones and the size of the area of operation (AO) in a mission. Completion of the low workload missions involved operating four drones in an area of 1.85 mi2, whereas eight drones were used in a 3.60 mi2 area for the high workload missions. Both AOs included no-fly-zones (NFZs), representing sections of the map where smoke from building fires would impair performance of the drones. The smaller AO for the low workload missions included four NFZs, each roughly the size of four to six city blocks (~0.02 – 0.03 mi2), while the larger AO used in high workload mission had six NFZs approximately 10 city blocks in size (~0.05 mi2). The expectation for the workload factor was that operating a higher number of drones in a larger AO would be associated with greater cognitive stress.

The other factor, recommendation quality, referred to the utility of flight path and resource delivery suggestions provided by the system. Poor-quality recommendations showed participants a path that might take the drone beyond the AO or through an NFZ, whereas good recommendations demonstrated a clear path between a drone carrying supplies and survivors in need of that resource. Missions were divided into four phases, each 3-minutes long, during which participants would see blocks of either good or poor recommendations. Participants had not been told that the quality of recommendations would differ and there were no outward indicators when recommendation type changed. Recommendations were presented on the left side of the screen, with two to four presented at a time, and participants had to click on each to review the suggested flight path and from there could either accept or reject them individually. Alternatively, participants could forego verifying the recommendations and instead click "reject" to dismiss all listed recommendations, or "accept" and thereby causing one of the recommendations to be chosen at random. If participants did not respond within 1 minute, the recommendations were categorized as having been ignored.

As shown in Table 1, the combination of these two factors resulted in four possible mission profiles – Mission 1, low workload and starting with good recommendations ("G"); Mission 2, low workload and starting with poor recommendations ("P"); Mission 3, high workload and starting with good recommendations; and Mission 4, high workload starting with poor recommendations.

Table 1. Summary of Mission Orders and Types

	1st Scenario			2nd Scenario			3rd Scenario			4th Scenario		
Order	**Msn**	**Wrkld**	**Rec. Order**	**Msn**	**Wrkld**	**Rec. Order**	**Msn**	**Wrkld**	**Rec. Order**	**Msn**	**Wrkld**	**Rec. Order**
A	1	Low	G - P - G - P	3	High	G - P - G - P	2	Low	P - G - P - G	4	High	P - G - P - G
B	3	High	G - P - G - P	1	Low	G - P - G - P	4	High	P - G - P - G	2	Low	P - G - P - G
C	2	Low	P - G - P - G	4	High	P - G - P - G	1	Low	G - P - G - P	3	High	G - P - G - P
D	4	High	P - G - P - G	2	Low	P - G - P - G	3	High	G - P - G - P	1	Low	G - P - G - P

Note: Participants completed the missions ("Msn") in one of four orders, A-D. Missions differed by task workload ("Wrkld") and the order of recommendation types ("Rec. Order"), alternating good ("G") and poor ("P") quality.

Procedure

The premise of these missions was that a category F-4 tornado had struck Houston, TX, and participants were tasked with using drones to find survivors then deliver supplies to them. The locations of these survivors were randomized for each mission but held constant between participants. Participants were instructed to stay within the boundaries of the AO as no survivors were located outside this area, and to avoid flying through any NFZs as the smoke could damage the drones. All drones had identical capabilities, with a 100-meter field of view (FOV) radius and set speed of 25 meters/second. At the start of each mission, the drones were located at the southern end of the map (Figure 1).

Participants were instructed to define waypoints on the map as a means of directing the flight path of each drone. All drones were numbered and given a unique color to indicate both the drone icon and its path on the map. If any survivors are located within the drone's FOV along this path, they were indicated on the map with a small map vector icon (i.e., pin) showing their location. If the survivors were in critical need of resources, meaning death was imminent if supplies were not received within 2 minutes, their location would instead be indicated by a small star. The color of these pins or stars denoted what resources are required, such that blue indicated a need for water, yellow corresponded to food, and green corresponded to medical supplies. The hubs for these three resource types were represented as small rectangles, colored to match the three supply types, and were located near the southern boundary of the map. Once they started finding survivors, participants then used the same drones to retrieve supply kits from the hubs and deliver them to survivors with the corresponding need (i.e., survivors indicated with a green star needed supplies retrieved from the green hub). Each drone could only carry and deliver one supply kit at a time, requiring them to return to a hub between each delivery, but there were no restrictions on the type of supplies a drone could carry.

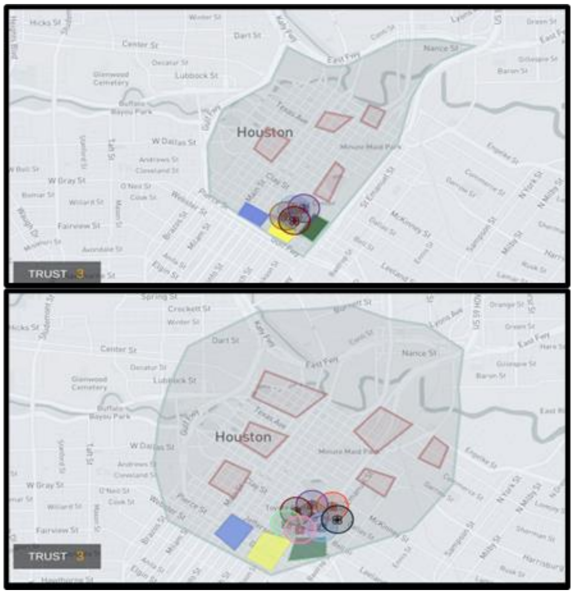
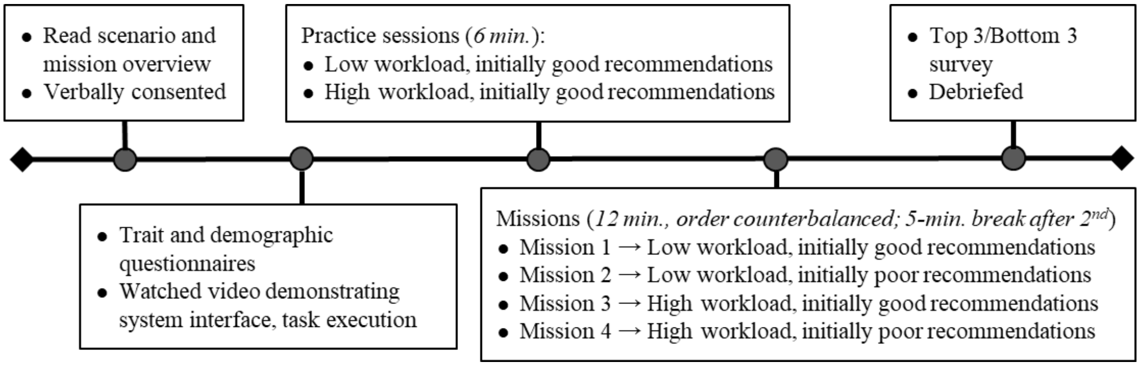


Figure 1. Sample AO for Low Workload (Top) and High Workload (Bottom) Missions

After consenting to participate in the study, participants completed a series of questionnaires, then watched a short video demonstrating the task, followed by two 6-minute familiarization sessions to practice using the interface. During this time, they experienced both the four- and eight-drone conditions and saw a combination of both recommendation types, starting with good recommendations. Following these sessions, participants completed the four missions in a pre-determined sequence, with a 5-minute break between the second and the third. After each mission, they were asked to complete two measures describing their experience and perception of the system, as well as a third measure summarizing their experience overall after completing the fourth mission. The experimental session took approximately 2 hours to complete (refer to Figure 2 for a detailed timeline). Performance, trust ratings, and recommendation response data from two participants were not properly saved for Mission 3.



- Read scenario and mission overview
- Verbally consented

- Trait and demographic questionnaires
- Watched video demonstrating system interface, task execution

Practice sessions (*6 min.*):
- Low workload, initially good recommendations
- High workload, initially good recommendations

Missions (*12 min., order counterbalanced; 5-min. break after 2nd*)
- Mission 1 → Low workload, initially good recommendations
- Mission 2 → Low workload, initially poor recommendations
- Mission 3 → High workload, initially good recommendations
- Mission 4 → High workload, initially poor recommendations

- Top 3/Bottom 3 survey
- Debriefed

Figure 2. Timeline of Experimental Procedures

Materials

Measures of Performance and Effectiveness

Participants' performance on the task was objectively measured as the number of survivors located, the number of supply kits delivered, and the number of survivors who died without receiving supplies in time. To better understand their engagement in the task, we also evaluated the number of times participants interacted with each drone. Effectiveness in the task was also assessed as the percentage of recommendations the participant selected against the number the system provided, which was used as a metric for participants' overall reliance on the system. Furthermore, this proportion was evaluated for the number of good versus poor recommendations selected.

A Novel Trust Measure

Our trust measure, the Continuous Online Numerical Score (CONS), was developed by examining the factors captured with behavioral and physiological measures, according to common trust models (Lee & See, 2004; Mayer et al, 1995; Hoff & Bashir, 2015; Kohn et al., 2021). The resulting mandate was that the trust measure must capture attitudes or intent to trust, and must capture it concurrently with behavioral or physiological measures. As those measures can be captured at any moment, and perception of the system may vary from one second to the next, participants should be allowed to update their self-reported trust at any time, thus determining the survey needed to be continuously available, or "online". Finally, for reasons of usability and statistical analysis, the output of this survey must be numerical. With this checklist in hand, we performed a literature review and found similar online measures previously deployed by Godfroy-Cooper (2022) and Desai (2012; 2013), exploring workload and trust, respectively. These measures validate the concept of asking participants to self-report the level of a given cognitive state during a concurrent task, and both approaches used a toggle to manipulate the level.

Inspired by these approaches, we used a conventional 5-point Likert scale modeled after Lee and Moray (1994) to create the CONS, where "1" corresponded to a lack of trust in the AI's recommendations and "5" corresponded total trust in the recommendations. Participants were trained to report their trust level when prompted or at any time when they felt their trust attitude had changed. A prompt reading "*Report trust level*" flashed in the bottom left of the user interface 30 seconds from the start of each mission and after the previous report. The most recent rating was always visible in the same location. To respond, participants were instructed to click on the "T" key on the keyboard to increase the trust level rating one point, "G" to maintain it at the current level, and "B" decrease the rating by one point. If participants failed to respond to the on-screen prompt within 60 seconds, a researcher verbally reminded them to respond.

This novel technique satisfied the intentions outlined above by being available concurrently with the focal task without substantially disrupting performance, allowing participants to self-report at any time their trust in the recommendations improved or degraded, and cultivating responses that could be analyzed using standard parametric procedures. Because participants were trained and encouraged to report their trust level at any time, the trust level that was previously reported can be assumed to be valid during any successive behavior or physiological sampling. Therefore, we can more accurately compare trust at time x to simultaneous behavioral or physiological measures. Similarly, capturing reported trust intent during the experiment more closely mirrors the sub-constructs of trust that behavioral and physiological measures proport to capture.

Behavioral Measure of Trust

Upon receiving a recommendation, participants could perform one of three behaviors which were indicative of trust states: Compliance, Verification, or Rejection. Blind compliance with the recommendation (i.e., accepting the recommendation without reading) is generally indicative of a high trust state, whereas verifying or reviewing the quality of the recommendation before making a decision suggests moderately low trust, and blind rejection of the recommendation suggests very low trust (see: Ezer et al., 2008; Lee & Moray, 1994; Moray et al., 2000). Participants could comply or reject the recommendation after verifying its content, but these actions are based on data interpretation, not trust. They influence future trust levels, but are functionally independent to trust attitudes when following verification. The cited works provide more detail on the trade-offs inherent in performing each of these behaviors.

Simulation Device and Environment

This study utilized the Vortex interface, developed, and customized through a partnership with Perceptronics Solutions, Inc. Vortex is a modular autonomy integration framework, previously demonstrated with the Army's Advanced Teaming Demonstration Program (A-Team), and designed to moderate the level of autonomy and thereby optimize human/machine team performance by assessing the operator's real-time workload and trust. Participants in the study used the Vortex interface to interact with the drones during the SAR tasks, and flight path recommendations were integrated into the display.

Surveys

Participants completed a series of questionnaires at the start of the experiment and throughout the session. All surveys were hosted online via Survey Monkey, and participants completed each by entering their responses on a Wi-Fi enabled tablet. The initial questionnaires included a demographic survey and three measures that asked participants to reflect on their experiences and feelings in working with automation and technology. These three questionnaires (Perfect Automation Schema, PAS (Tschopp & Ruef, 2020); Propensity to Trust Technology, PTT (Jessup et al., 2019); and Trust in Automation Inventory, TAI (Chien et al., 2014)) prompted participants to indicate their agreement to a series of statements using a 5-point Likert scale. Following each of the four missions, participants completed two surveys asking about their experience during that mission. The National Aeronautics and Space Administration Task Load Index (NASA-TLX) requests participants to report the difficulty experienced during the mission, and Shaefer's Trust Perception Scale – Human/Robot Interactions (HRI; Schaefer, 2016) measure requests them to estimate what percent of the time they expected the system to perform in a particular way. After the final mission and corresponding surveys, participants also completed a "Top 3, Bottom 3" measure asking them to describe the three best and worst parts of the study and whether they noticed any differences in the quality of recommendations.

Physiologic Sensors

The FX3 Remote Eye Tracker (Eye Tracking LLC) was used to unobtrusively track eye movements and measure pupillary response, and was used in conjunction with the EyeWorks Cognitive Workload Module to calculate the Index of Cognitive Activity (ICA), a real-time measure of workload. The camera was located directly below the primary display and the participant seated approximately 36" away. Throughout the experimental session, participants also wore an Empatica E4 (Empatica, Inc.) as a means of measuring participants' movement and physiologic states. In addition to tracking motion on the x-, y-, and z-axes, the E4 records heart rate, blood volume pulse (BVP), electrodermal activity (EDA), and skin temperature. To maintain brevity however and focus on the most critical elements of the study, analyses of these measures are not discussed here.

RESULTS

Task Performance

Participants performed as well as expected on the task, on average finding 87.0% (*SD* = 0.10) of the 20 survivors across each of the four mission profiles. Of the survivors located, 81.9% received the appropriate type of supply, representing 71.3% of all survivors in each mission, and only 2.3% died (*SD*s = 0.12 and 0.04, respectively). Participants also performed well in navigating around the NFZs rather than flying straight through them, averaging only 1.37 crossings per mission (*SD* = 1.54). When operating four drones simultaneously, participants interacted with each approximately 8.56 times during a 12-minute mission (*SD* = 2.52), which included an average of 3.67 supply deliveries made (*SD* = 1.29). The average number of interactions and deliveries per drone was somewhat lower when participants were responsible for operating eight drones, in those missions interacting with each 5.35 times (*SD* = 2.00) and making 1.98 deliveries (*SD* = 0.90). Finally, participants used each drone at least once across all missions, and completed at least one delivery with each when four drones were included in the mission, but on average used only seven drones to deliver supplies when eight were available.

Recommendations and Trust

The first step in assessing the suitability of an online numerical trust measure to capture participants' trust in the flight path recommendations was to determine whether participants' ratings corresponded to their responses to the suggested routes. Trust ratings were collected using the CONS, and compared to the three behaviors operators could perform that generally indicate trust levels (Lee & Moray, 1994): *Verification*, reading the recommendation before accepting or rejecting; *Compliance*, accepting the recommendation without reading; and *Rejection*, dismissing the

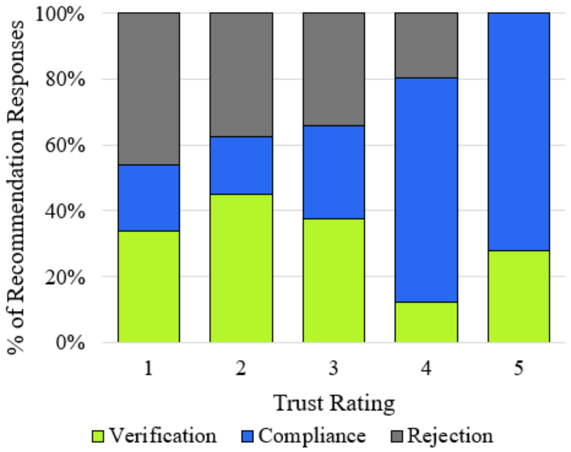


Figure 3. Recommendation Responses by Trust Rating

recommendation without reading. Of the 841 recommendation sets presented to participants across all four missions, 81.1% were read for verification, 13.8% were either accepted or rejected without reading, and 4.4% were ignored (note, participants' ignoring recommendations were not analyzed further because the data include no timestamp indicator for an event that did not occur). Among the unread recommendations, most were rejected (*n* = 96, 80.7%) rather complied with (*n* = 23). As demonstrated in Figure 3, when the distribution of participants' most recent trust ratings was compared against their responses to the recommendations, the results of a $\chi 2$ test for independence revealed a significant relationship between the two metrics ($\chi 2$(8, N = 749) = 42.71, *p* < .001, Cohen's ω = .262). This finding supports the hypothesis that response to the CONS measure aligned with participants' behavior, in that the rate of compliance was greater at higher levels of explicit trust.

Efficacy of the CONS tool to measure real-time trust was further demonstrated when analyses factored in the quality of recommendations to which participants were responding. As expected, when participants chose to verify the recommendation, they were significantly more likely to accept suggestions that were of good quality (*M* = 0.71) over those that were poor (*M* = 0.26; *t*(375.88) = 6.95, *p* < .001), and marginally more likely to reject poor-quality recommendations (*t*(446) = -1.90, *p* = . 058). Even when analyzing behavior separately based on whether they did (*n* = 15) or did not notice a difference in the quality of recommendations, acceptance following verification was significantly higher for good recommendations (*t*(191.02) = 6.07, *p* < .001 and *t*(179.18) = 3.82, *p* < .001, respectively), though the difference in response to poor recommendations was only significant for those who reported noticing variations in the quality (*t*(191.02) = -2.13, *p* < .05). These findings confirm that participants were at least somewhat sensitive to the differences in recommendation quality. There was however no effect on the mission performance measures (i.e., number of survivors located or number of supplies delivered) for the order of recommendation quality blocks, nor was there any effect on participants' responses to the NASA-TLX (all *p*s > .10).

Subsequent analyses though did point to a possible priming effect, such that the quality of the recommendation shown at the start of each mission influenced participants' perception and acceptance of recommendations throughout the scenario. In particular, participants were significantly more likely to reject good recommendations throughout a mission if the initial block of recommendations was poor (*M* = 2.68) rather than good (*M* = 2.05; *t*(107.69) = -2.11, *p* < .05). The converse of this was also demonstrated, as participants were more likely to accept good recommendations during a mission in which the first recommendation encountered was helpful (*M* = 0.68; *t*(95.74) = 3.73, *p* < .001) than if it was of poor-quality (*M* = 0.25). It is not surprising then that participants' online trust rating was significantly higher during the first phase of a mission when they received good recommendations (*M* = 2.98) over poor recommendations (*M* = 2.66; *t*(91.20) = 2.39, *p* < .05).

Workload

The second factor in this study was the manipulation of mission workload, having participants complete the same task either with four drones in a smaller AO (low workload) or eight drones in a larger AO (high workload). As hypothesized, participants found a significantly higher proportion of the 20 survivors during the low workload missions (*M* = 0.91, *SD* = 0.07) compared to the high workload (*M* = 0.83, *SD* = 0.18; *t*(93.03) = -4.79, *p* < .001), though the same effect was not evident for either the proportion of survivors who received supplies (*p* = .174) or the proportion of survivors who died due to not receiving critically needed (*p* = .812; Figure 4). In addition to this difference in objective task performance, responses to the NASA-TLX revealed that participants expended marginally greater effort for the high (*M* = 57.32, *SD* = 18.54) vs low (*M* = 50.63, *SD* = 18.29) workload missions (*t*(109.98) = 1.92, *p* = .057). Neither participants' trust in the system nor their responses to flight path recommendations differed in regards to the workload manipulation (all *p*s > .40).

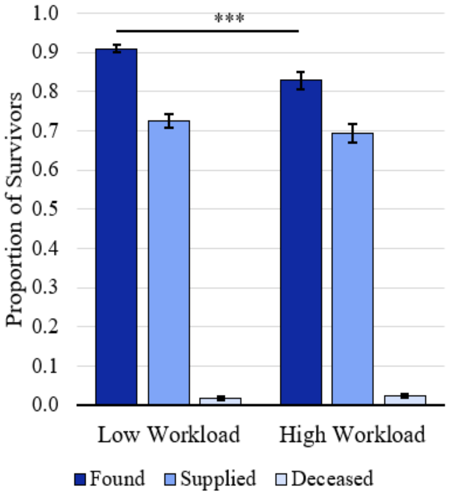


Figure 4. Performance by Workload Level

Differences by Mission Profile

Assessing the combined influence of manipulating participants' workload and the quality of flight recommendations they received revealed significant differences between the four missions regarding the number of tornado survivors who were found (F(3,56.94) = 8.07, *p* < .001; Figure 5). Post-hoc comparisons confirmed that the proportion of survivors found in both low-workload Missions 1 and 2 (*M* = 0.92 and *M* = 0.91, respectively) was significantly greater than the proportion found in either Mission 3 (*M* = 0.84) or Mission 4 (*M* = 0.82). Similar comparisons for the proportion of survivors who received supplies, died without receiving supplies, and the number of times drones were flown into a NFZ revealed no significant differences between the mission profiles. No effect was evident for the order of recommendation types, indicating this manipulation did not significantly influence participants' overall mission performance. Finally, there were no differences between the missions regarding participants' behavioral responses to the recommendations or their CONS trust ratings (all *p*s > .20).

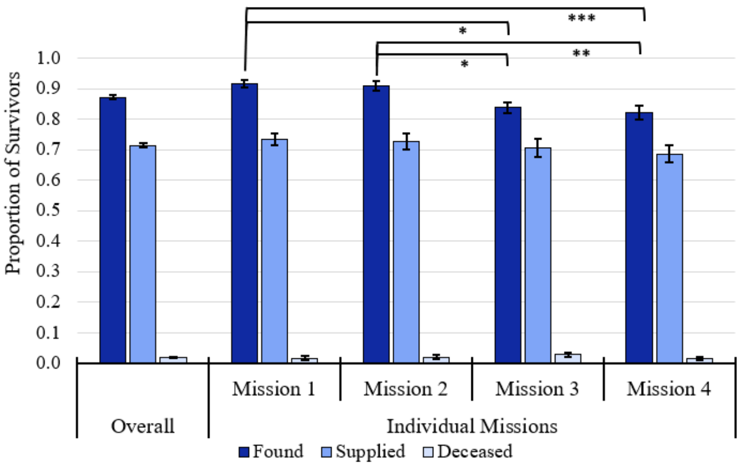


Figure 5. Performance Differences Across Missions

DISCUSSION

Overall participants performed well on the task, typically finding and delivering supplies to 71% of the tornado survivors, with few of those in critical need dying before they received help. As hypothesized, the quality of the flight path recommendations influenced participants' willingness to accept the suggestions provided, recognizing and rejecting inefficient or hazardous routes represented on a simple 2-dimensional display. Their responses to the CONS measure of trust further bore out this observation. In particular, we found a significant relationship between behavioral trust measures of compliance, verification, and rejection at different levels of trust, where behaviors that indicate high trust are more common when trust is reported to be high using CONS. This pattern supports our hypothesis that the novel trust score is a valid method of capturing trust attitudes and predicting trust behaviors. Additionally, increasing workload by doubling both the number of drones and the operational space had a detrimental effect on performance, decreasing the average number of survivors found by 8.8%, but no corresponding difference was evident for the number of supply kits delivered or the number of recovered individuals who subsequently died. These analyses however failed to find a meaningful correlation between participants trust ratings and task performance, suggesting fluctuations in trust did not influence participants performance on the mission as a whole.

While the novel trust measure was effective, we expect that several optimizations would improve the correlation between the self-report tool and existing trust measures. These optimizations include capturing perceived risk and self-confidence, and increasing workload. Existing models confirm the relationship between trust, risk, and self-confidence. Trust requires vulnerability (Lee & See, 2004), such that failure on the task by either the automation or human results in negative consequences. We can narrow our focus on trust by both increasing the consequences of failure on future iterations of this task, and capture participants' assessment of the risk level to control the effect of that moderator on trust ratings. Additionally, we can assess participants' self-confidence on the task, then control for these differences in analysis. Lee and Moray (1994) established that trust behaviors are influenced by the participant's perceived ability to perform the task: effectively, if they are more confident in themselves than they trust the automation, they will reject the automation's recommendation and perform the task unaided. We further observed that participants trended towards frequent verification of recommendations, which we believe was due to relatively low workload levels. Previous studies have demonstrated that trust measures are most sensitive when all behaviors require tradeoffs, such as when operators do not have enough available workload capacity to verify all recommendations, and must abandon other tasks to instead double-check the automation's work (Parasuraman et al., 2008; Salehi et al., 2021). Thus, workload must be strategically increased in future efforts, as excessive increases will force operators to perform compliance behaviors even when trust is low due to continuous overload. Possible modifications to increase the cognitive burden in future studies include requiring that participants respond to audio cues, monitor the energy or fuel levels on individual drones, or including an extraneous task such as the *n*-back (Ayaz et al., 2012).

Moreover, these findings served to replicate results from Lewis and colleagues (2010) who found that human performance on a foraging task was worse as workload increased proportionally with the addition of more drones to be operated. As predicted through scalability modeling (Humann & Pollard, 2019), most participants in the current study failed to utilize all available drones for the second phase of the high workload missions. However, the fact that the proportion of survivors who received supplies did not also differ between the workload levels indicates another factor may have contributed to the disparity on the proportion of survivors found. This was most likely due to participants' searching a larger area for the same number of survivors since lower density would increase the difficulty of the search. Future investigations into the effect of swarm size will either keep the size of the AO consistent across workload levels or increase the number of survivors in a proportional manner. Likewise, issues with the presentation and execution of recommended flight paths may have interfered with participants' perception of the suggestions and the system as a whole. When recommendations were generated, they were specific to a drone's location in that instant and the suggested flight path was not updated as it continued to move along the initial route. Consequently, if a participant chose to accept the recommendation, the corresponding drone would then have to move back to its location when the recommendation was generated before starting the suggested route, likely making the process more disruptive the longer the recommendation had been available.

We should also acknowledge the theoretical disadvantages. Participants were encouraged to supply trust ratings when their trust changed, but may not be motivated to update. On-screen prompts were included but participants too often ignored these until they were directed to respond by research personnel. Such reminders were necessary to ensure consistent ratings, but at the risk of inducing an unknown amount of additional workload. Subsequent studies will address this by emphasizing to participants the importance of this item and increasing the salience of the visual prompts. We are also interested in testing simplified versions of the prompt in the future, focused on a dichotomy of trust/no trust. This version may be useful for field applications where fast response times and minimal added workload are the utmost priority, rather than the current focus on stringent validation of other novel real-time measures. Though preliminary, this research provides important insight into the human factors of operating drone swarms as well as the critical role of trust when fielding decision-support tools. Chiefly, we successfully demonstrated that a subjective, explicit trust measure could be incorporated into an operational task with minimal disruption, thereby capturing trust attitudes in conjunction with demonstrated behaviors. By extension, the results support the possibility of implementing a lightweight, real-time measure into other HMT environments and allowing operators to describe their experience and provide feedback to the system without distraction from the focal task.

ACKNOWLEDGEMENTS

The authors would like to thank Colleen Mynaugh, Nicholas Crothers, and Diem-Thuy Le for their invaluable help and feedback throughout all steps of this research effort. Additionally, we are grateful for Athena Johnson's instrumental time supporting required Vortex interface changes needed to execute this study.

REFERENCES

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *Neuroimage, 59*(1), 36-47.

Chien, S. Y., Semnani-Azad, Z., Lewis, M., & Sycara, K. (2014, June). Towards the development of an inter-cultural scale to measure trust in automation. In *International Conference on Cross-cultural Design* (pp. 35-46). Springer, Cham.

Cummings, M. L., Nehme, C. E., Crandall, J., & Mitchell, P. (2007). Predicting operator capacity for supervisory control of multiple UAVs. *Innovations in Intelligent Machines*, *1*, 11-37.

de Visser, E. J., Beatty, P. J., Estepp, J. R., Kohn, S., Abubshait, A., Fedota, J. R., & McDonald, C. G. (2018). Learning from the slips of others: Neural correlates of trust in automated agents. *Frontiers in Human Neuroscience, 12*, 309.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied, 22*(3), 331.

Desai, M. (2012). *Modeling trust to improve human-robot interaction*. Doctoral dissertation. Lowell, MA: University of Massachusetts Lowell.

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). "Impact of robot failures and feedback on real-time trust." in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction* (HRI), March, 2013; 251–258.

Desmet, C., Deschrijver, E., & Brass, M. (2014). How social is error observation? The neural mechanisms underlying the observation of human and machine errors. *Social Cognitive and Affective Neuroscience, 9*(4), 427–435.

Ezer, N., Fisk, A. D., & Rogers, W. A. (2008). Age-related differences in reliance behavior attributable to costs within a human-decision aid system. *Human Factors 50*(6), 853–863. doi: 10.1518/001872008X375018

Godfroy-Cooper, M., Miller, J., Bachelder, E. N., & Szoboslay, Z. (2022, September) Operator state monitoring for workload prediction and management. *48th European Rotorcraft Forum*, September 6-8, Winterthur, Switzerland.

Goodyear, K., Parasuraman, R., Chernyak, S., de Visser, E., Madhavan, P., Deshpande, G., & Krueger, F. (2017). An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents. *Social Neuroscience, 12*(5), 570-581.

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human Factors, 58*, 509–519.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Human Factors, 57*, 407–434. doi:10.1177/0018720814547570

Humann, J., & Pollard, K. A. (2019, October). Human factors in the scalability of multirobot operation: A review and simulation. In *2019 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 700-707). IEEE.

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In *Virtual, Augmented and Mixed Reality. Applications and Case Studies: 11th International Conference, VAMR 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21* (pp. 476-489). Springer International Publishing.

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53-71.

Lee, J. D., & Moray, N. (1991). "Trust, self-confidence and supervisory control in a process control simulation." in *Conference Proceedings 1991 IEEE International Conference on Systems, Man, and Cybernetics*, October, 1991; 291–295.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243–1270. doi: 10.1080/ 00140139208967392

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*(1), 153-184. doi: 10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors 46*(1), 50–80. doi: 10.1518/hfes.46.1.50.30392

Lewis, M., Wang, H., Chien, S. Y., Velagapudi, P., Scerri, P., & Sycara, K. (2010). Choosing autonomy modes for multirobot search. *Human Factors, 52*(2), 225-233.

Lyons, J. B., Sadler, G. G., Koltai, K., Battiste, H., Ho, N. T., Hoffmann, L. C., Smith, D., Johnson, W., & Shively, R. (2017). Shaping trust through transparent design: Theoretical and experimental guidelines. In *Proceedings of the AHFE 2016 International Conference on Human Factors in Robots and Unmanned Systems, July 27-31, 2016, Walt Disney World®, Florida, USA* (pp. 127-136). Springer International Publishing.

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction* (pp. 3-25). Academic Press.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. doi: 10.5465/amr.1995.9508080335

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors, 46*(2), 196–204.

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology – Applied*, 6, 44–58.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics, 37*(11), 1905–1922. doi: 10.1080/00140139408964957

Parnell, K. J., Fischer, J. E., Clark, J. R., Bodenmann, A., Galvez Trigo, M. J., Brito, M. P., Soorati, M. D., Plant, K. L., & Ramchurn, S. D. (2022). Trustworthy UAV relationships: Applying the Schema Action World taxonomy to UAVs and UAV swarm operations. *International Journal of Human–Computer Interaction*, 1-17.

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

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making, 2*(2), 140-160.

Rice, S. (2009). Examining single-and multiple-process theories of trust in automation. *Journal of General Psychology, 136*(3), 303–322. doi: 10.3200/GENP.136.3.303-322

Salehi, P., Chiou, E. K., Mancenido, M., Mosallanezhad, A., Cohen, M. C., & Shah, A. (2021, September). Decision deferral in a human-AI joint face-matching task: Effects on human performance and trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 65, No. 1, pp. 638-642). Sage CA: Los Angeles, CA: SAGE Publications.

Schaefer, K. E. (2016). Measuring trust in human robot interactions: Development of the “Trust Perception Scale-HRI”. In *Robust Intelligence and Trust in Autonomous Systems* (pp. 191-218). Boston, MA: Springer US.

Scielzo, S., & Kocak, D (2021, December). A Multi-Domain Robotic Teammate Framework: Next Generation Human-Machine Interface Guidelines to Support Trust and Mission Outcomes. In *The Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), Orlando, FL*.

Tenhundfeld, N. L., de Visser, E. J., Haring, K. S., Ries, A. J., Finomore, V. S., & Tossell, C. C. (2019). Calibrating trust in automation through familiarity with the autoparking feature of a Tesla Model X. *Journal of Cognitive Engineering and Decision Making, 13*(4), 279–294. doi: 10.1177/1555343419869083

Tschopp, M., & Ruef, M. (2020, May 07). PAS – The Perfect Automation Schema: Influencing trust. *scipAG*. https://www.scip.ch/en/?labs.20200507

Wijnen, L., Coenen, J., & Grzyb, B. (2017, March). "It's not my fault!" Investigating the effects of the deceptive behaviour of a humanoid robot. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 321-322.

Wojton, H. M., Porter, D., Lane, S. T., Bieber, C., & Madhavan, P. (2020). Initial validation of the trust of automated systems test (TOAST). *Journal of Social Psychology, 160*(6), 735–750. doi: 10.1080/00224545.2020.1749020