5-2023

The Examination of Factors that Influence Trust in a Multi-Agent Team Context

Cherrise Ficke

The Examination of Factors that Influence Trust in a Multi-Agent Team Context

by

Cherrise Ficke

A thesis submitted to the College of Aeronautics of
Florida Institute of Technology
in partial fulfillment of the requirements
for the degree of

Master of Science
in
Human Factors in Aeronautics

Melbourne, Florida
May, 2023

We the undersigned committee hereby approve the attached thesis,
“The Examination of Factors that Influence Trust in a Multi-Agent Team Context
by
Cherrise Anne Ficke

____

Meredith Carroll, Ph.D.
Professor
College of Aeronautics
Major Advisor

____

Jessica L. Wildman, Ph.D.
Associate Professor
School of Psychology

____

Rian Mehta Ph.D.
Assistant Professor
College of Aeronautics

____

John E. Deaton, Ph.D.
Professor and Dean
College of Aeronautics

Abstract

Title: The Examination of Factors that Influence Trust in a Multi-Agent Team Context

Author: Cherrise Ficke

Advisor: Meredith Carroll, Ph.D.

Current dyadic teams in the human-agent teams literature demonstrates that Propensity to Trust in Technology (PTT), previous experience with an agent, levels of autonomy (LOA), workload, and mission performance affect trust and mission performance to some capacity. However, the purpose of this study was to expand this line of research by examining how these factors influence trust and mission performance in a multi unmanned aerial vehicle (UAV) context. To investigate the relationship between these factors and trust, an archival study was conducted using data from a previously-conducted, multi-UAV study. The previous study utilized a within-subjects repeated measures design in which participants conducted four separate, 5-minute, multi-UAV missions in four different LOA conditions utilizing four UAVs. The four LOAs included manual (agents did not assist in target selection), advice (agents provided suggestions for target selection), consent (agents pre-selected targets), and veto (agents completed all task independently). Measures of performance were collected through interactions with the drones, whereas measures of PTT, workload, and trust were collected via self-report surveys. Forty-seven participants experienced 4 trials each, resulting in a total of 188 trials to investigate. Utilizing this data, two multiple regression analyses were conducted. The first examined the relationship between the dependent variable of trust and independent variables of PTT, previous experience with agents, LOA, and workload ratings. The second examined the relationship between the dependent variable of mission performance and independent variables of reported trust, LOA, and workload. Findings revealed that PTT and mission performance positively and significantly influenced trust, whereas the advice LOA, consent LOA, and workload negatively and significantly influenced trust. Results from the second multiple regression found that the consent LOA, veto LOA, and trust positively and significantly influenced mission performance in a multi-HATs.

Table of Contents

List of Tables

Chapter 1
Introduction

Contemporary society has demonstrated technological breakthroughs in autonomous systems, which have shown increases in usage and complexity of autonomous agents. Agents are now able to carry out more complex tasks and work alongside human operators to complete missions such as Intelligence Surveillance and Reconnaissance (ISR) and cybersecurity defense operations, resulting in the proliferation of automated systems expanding beyond the use of agents being perceived as tools (Otto, 2016; Chen & Barnes, 2014). The United States Air Force (USAF) expects 60% of the USAF to be unmanned by 2035, exemplifying a technological shift to include more autonomous systems in their future endeavors (U.S. Department of Transportation, 2014; OSD, 2017). This trend will set the stage for an influx of human-agent teams (HAT) in which humans and agents will work together to accomplish a common goal. Due to recent advancements in autonomous capabilities, humans are becoming more reliant on agents to complete lower-level tasks so that the human operator can perform higher-level decision-making tasks (Cummings, 2015), allowing the team to achieve optimal levels of mission performance. To ensure overall team performance, appropriate trust dynamics between the human and the agent must be present to allow the human's trust accurately reflects the agent's limitations and capabilities (Yu et al., 2019; Kohn et al., 2020).

In the HAT literature, 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 & See, 2004, p. 54). Behaviors such as overtrusting or undertrusting an agent have been identified as inappropriate behaviors in HAT, due their adverse effects on mission performance and workload (Kox et al., 2021). For example, overtrusting behaviors can result in complacency, leading to overreliance on the agent. This is especially dangerous when a human trusts an agent to handle a situation that it is not equipped to handle (Parasuraman et al., 1993). For instance, when pilots engage autopilot in harsh weather conditions, there are times the weather conditions become too severe for autopilot to operate in, leading to the autopilot disengaging and offloading the workload to the pilot. Under these circumstances, the pilot may be ill-equipped to handle the situation as they may be out-of-the-loop, making it harder for them to manually fly the aircraft. On the other hand, if the human operator presents low levels of trust in an agent, this may lead to the operator rejecting useful assistance from the agent, leading to higher stress and workload, and the potential risk of impeding mission performance (Parasuraman et al., 1993). Therefore, it is important to ensure performers' trust accurately reflects the agent's capability for a given task (Bobko et al., 2022).

As the literature has identified trust as a pivotal component to HATs, most research has studied the effects of trust in HAT dyads in which one human and one agent are present. Studies have shown there are a range of different factors that influence trust in this context, including propensity to trust (PTT), previous experience with an agent, Level of Autonomy (LOA) and workload (Alacorn et al., 2016; Walliser et al., 2019; Johnson et al., 2011; Narayanan et al., 2014). However, few studies have examined the effects of trust or the factors that influence trust in multi-agent HATs. As more complex missions and tasks require numerous agents, the human operator must be able to trust a team of agents at the individual and team levels, demonstrating a need to explore trust dynamics in multi-agent teams. The purpose of this study is to examine the relationship between human trust in multi-agent HATs and a set of variables that have been shown in the literature to influence trust in HAT dyads.

Research Questions and Hypotheses

Research Questions (RQs):

The research questions for this study are:

RQ1: What is the influence of the following set of factors on an individual’s trust in multi-agent HATs?

1a. What is the influence of PTT on an individual’s trust in a multi-agent HAT?

1b. What is the influence of previous experience with the agent on an individual’s trust in a multi-agent HAT?

1c. What is the influence of level of autonomy (LOA) on an individual’s trust in a multi-agent HAT?

1d. What is the influence of workload on an individual’s trust in a multi-agent HAT?

1e. What is the influence of mission performance on an individual’s trust in a multi-agent HAT?

RQ2: What is the influence of the following set of factors on an individual’s mission performance in multi-agent HATs?

2a. What is the influence of LOA on an individual’s mission performance in a multi-agent HAT?

2b. What is the influence of workload on an individual’s mission performance in a multi-agent HAT?

2c. What is the influence of trust on an individual’s mission performance in a multi-agent HAT?

Hypotheses:

H_{1a}: PTT will significantly and positively influence an individual’s trust in multi-agent HATs when controlling for previous experience, current LOA, workload, and mission performance.

H_{1b}: Previous experience with an agent will significantly and positively influence an individual’s trust in multi-agent HATs when controlling for PTT, current LOA, workload, and mission performance.

H_{1c}: The agents current LOA will significantly influence an individual’s trust in multi-agent HATs when controlling for PTT, previous experience, workload, and mission performance.

H_{1d}: Workload will significantly and negatively influence an individual’s trust in multi-agent HATs when controlling for PTT, previous experience, current LOA and mission performance.

H_{1e}: Mission performance will significantly and positively influence an individual’s trust in multi-agent HATs when controlling for PTT, previous experience, current LOA and workload.

H_{2a}: The current LOA will significantly influence individual’s mission performance in multi-agent HATs when controlling for workload, and trust.

H_{2b}: Workload will significantly and negatively influence individual’s mission performance in multi-agent HATs when controlling for current LOA and trust.

H_{2c}: Trust will significantly and positively influence individual’s mission performance in multi-agent HATs when controlling for the current LOA and workload.

Significance of the Study

The practicality for employing multi-agent teams for different types of missions has become more widely used in the military domain (Otto, 2016). For instance, in a reconnaissance mission, agents will have different payloads, in which some agents will assist in target detection on the ground whereas other agents will provide aerial support. Due to the limited research available for trust in multi-agent teams, there is no clear understanding of what factors influence trust development in a multi-agent team compared to dyadic teams. Along with this, not as many multi-agent studies have executed tasks in dynamic environment where the human operator is interacting with the agent. Rather, many multi-agent HAT studies have implemented vignette-based studies in which interactions with the agent are not dynamic and do not reflect real-world scenarios. To address these gaps, the current study will contribute to the multi-agent literature by identifying factors that influence trust and mission performance in a multi-agent HAT context in which dynamic interactions occur.

Chapter 2
Literature Review

Introduction

Trust is integral to the success of HAT due to its effect on the human operator’s behaviors (Parasuraman et al., 1993). In the past, HATs have been typically comprised of one human and one agent, representing a dyadic relationship. As a result, there is an abundant amount of literature that has explored factors that influence trust in dyadic relationships in a HAT context e.g., Jung et al., 2017. However, as HATs operate in more elaborate and complex environments, the paradigm of HATs expands beyond a dyadic relationship, to one in which multi-agent teams have become more prevalent. As this shift continues, it is imperative for research to investigate whether factors that influence trust in a dyadic HAT context also impact multi-agent teams. The literature has found that many of these factors are closely linked together, and mission performance also plays an important role in the structure of factors that influence trust (Levinthal & Wickens, 2006). To illustrate these relationships, this section will review factors that have been identified in the dyadic HAT and multi-HAT literature which influence trust, along with factors that play an influential role in mission performance.

Propensity to Trust

PTT is defined as an individual difference that reflects one’s expectations of the trustworthiness of others and general willingness to trust others (Mayer et al., 1995; Rotter, 1967). Throughout the human-human teams (HHTs) and HAT literature, variations of the PTT questionnaire have been utilized to measure an individual’s general tendency to trust in automation or another human teammate before interactions occur. PTT is a well-established individual factor that has been shown to be a significant predictor of trust, as illustrated in a meta-analysis conducted by Colquitt et al. (2017). To provide further context on this measure, the following section illustrates the use and relevancy of PTT in HHT and HAT contexts.

In Alarcon et al. (2016) an empirical study was conducted to investigate whether PTT would act as a predictor for trustworthiness in unfamiliar dyads over time in an HHT context. Schoorman, Mayer & Davis' (1996) Propensity to Trust scale, containing eight items, each item on a 7-point Likert scale, was administered at the beginning of the study. In the experimental task, two participants were presented with the prisoner's dilemma task in which they discussed whether they would agree to cooperate (both reveal secrets, or both keep the secrets from the experimenter). After discussing what to do, participants met individually with the experimenter and informed them of the behavior they chose. Behaviors were coded as trusting if the participant performed the agreed behavior. Participants were then asked to complete a self-reported trustworthiness assessment. This process was repeated three times, each with different partners. The study manipulated the familiarity participants had with other participants, from unfamiliar (never met the other participant before) to familiar (have met the other participant before), to investigate how trust behaviors developed amongst different degrees of familiarity. Results of the study revealed that PTT was a significant predictor for the first rounds of trustworthiness for unfamiliar dyads p<0.05. In other words, the higher scores a participant had on the PTT scale, the greater likelihood they would report the agreed behavior to the experimenter in the first round.

As the previous study demonstrated the relevance of PTT in a HHT context, studies in the HAT domain have exhibited similar trends. To assess the effects of participants' trust, Zhang et al. (2021) conducted a study to investigate how trust fluctuated across different trust repair strategies in response to trust violations committed by an agent, and how PTT influenced trust throughout each condition. Independent Variables (IVs) in the study included the violation type and trust repair strategies. The three violation types included logic errors (errors causing machines to produce relevant but incorrect output), semantic errors (misunderstanding commands and exercising irrelevant action), and syntax error (failure to respond to instructions). The four types of repair strategies included no repair attempts, internal-attribution apology (e.g., “sorry, I was too timid to ask questions”), external attribution apology (e.g., “Sorry, the question was phrased weird”), and denial (e.g., “I didn’t do it”). This study utilized a mixed subject design under thirteen conditions (3 failure types x 4 repair attempts +1 control). In this vignette study, participants first completed a demographic questionnaire that included the Propensity to Trust Technology Scale by Jessup et al. (2019), which included six items on a Likert scale from 1 (Strongly agree) to 5 (disagree). Subsequently, participants watched a video of a robot committing one of the three types of performance failures (logic, semantic and syntax), followed by one of the four trust repair strategies (no repair attempts, internal-attribution, external-attribution apology, and denial). Participants were randomly assigned to three of the thirteen conditions. After watching each video, the Human Robot Interaction (HRI) trust perception scale (Schaefer et al., 2013) was administered to measure competency-based trust in agents. After conducting a Multivariate Analysis of Variance (MANOVA), results revealed the experimental conditions impacted post-interaction trust scores, in which PTT served as a significant covariate ($p < .001$). Based on these findings, the study demonstrated PTT scores significantly influenced post-interaction trust scores.

Additional studies have also shown PTT scores as a strong influencer of trust in dynamic HAT environments. Thomeson et al. (2022) investigated how PTT affected a participants trust development during a reconnaissance mission. In the experimental task, participants were instructed to collaborate with a drone to search through two abandoned houses in a virtual reality (VR) environment. The drone’s job was to scan the area and report to the participant whether the area was safe. Within each house, the participant and drone searched through 3 floors, in which a trust violation by the drone occurred on the second floor. IVs included agent anthropomorphism (machine-like drone vs. human-like drone) and an explanation of the error (present vs. absent). The study utilized a within-subjects design, in which each participant experienced two different types of explanation, and anthropomorphism was manipulated as a between-subjects factor, therefore half of the participants experienced a human-like drone, and half experienced a machine-like drone. The Propensity to Trust Technology scale (Jessup et al., 2019) was administered before participants interacted with the drone. To assess trust development during the task, a single-item trust scale was reported on each floor. After searching each house, a multidimensional trust survey containing 11 items on a 7-point Likert scale was employed. Results revealed that high PTT scores were associated with slightly higher forgiveness scores from the multidimensional trust survey (r=0.37, *p*=0.022). These results suggest there is a positive relationship between high PTT scores and forgiveness in an agent after a trust violation has been committed. Similar results have been found in the HHT literature where there is a positive correlation between PTT scores and the participant's agreed behavior with their teammate (Alarcon et al., 2016).

Along with self-reported trust scores, PTT scores have also shown correlations with trust related behaviors such as accepting or rejecting an agent's advice. Pynadath et al. (2019) aimed to predict subsequent trust behaviors in humans from pre-surveys including the PTT Scale (McShane, 2014). For the experimental task, the human teammate worked with different robots across eight reconnaissance missions in a virtual environment. The goal of the task was to search through 15 buildings within 10 minutes without facing fatal injuries from enemies. The agent served as a scout, where they relayed information to the human operator on whether the building was safe to enter. Within each mission, the robot had an accuracy rate of 80% in which they correctly reported the status of 12/15 buildings. Anthropomorphism of the robot was manipulated where half of the robots looked like robotic dogs and the other half were portrayed with a more robot appearance. Elaboration on the scouting reports also served as a manipulation. When relaying information back to the human, half of the robots would provide an assessment of the safety of the building as being "safe" or "dangerous" with no additional information. However, the other half of the robots included a confidence-level elaboration (e.g., “I am 80% confident the house is safe”). Trust was measured via behavioral trust indicators such as accepting or rejecting the robot's decision. Based on results from the experimental study, Pynadath et al. (2019) developed a predictive model to investigate whether factors like PTT could anticipate a human operator’s behaviors. Results from the model revealed participants who scored higher on the PTT scale were more likely to accept the robot's recommendations throughout the mission. Furthermore, when predicting trust behaviors throughout the experiment, the PTT scale exhibited the most influential impact on trust behaviors compared to other pre-survey questionnaires. Although the authors indicated PTT scores were predictive of the operator’s actions, no statistical significance was provided in the paper.

In summary, results from numerous studies reveal that PTT is a significant and positive predictor for trust in teammates in HHT and HAT contexts. Based on the studies identified in the literature, the hypotheses of the current study state that PTT will significantly and positively influence an individual’s trust in a multi-agent team context.

Previous Experience with an Agent

Studies have shown a person’s prior experience with an agent is another individual characteristic that influences trust in HATs. For example, in Walliser et al. (2019), results demonstrated that familiarity/and or experience with a teammate can affect initial trust and can continue affecting trust throughout the HAT mission. Although the current literature shows that more experience can help facilitate trust in teams, other studies suggest that only positive experiences with an agent can increase trust (Hafizogly et al., 2019). For further elucidation on this construct, the following section outlines studies that examine the effects of previous experiences with agents in relation to trust in HATs.

In Walliser et al. (2019), a study was conducted to determine whether team performance in HATs could be improved through team-building exercises before the actual task. In the task, participants collaborated with a teammate to defend a ship from incoming missiles. To achieve mission success, the player was required to correctly identify the missile type, heading, and time-to-impact, then deploy the appropriate countermeasure in a suitable location. The study utilized a between-subjects design, in which participants experienced one of two conditions: formal and informal team-building exercises. In the informal exercise, participants completed a non-task-related cooperative game with the confederate; whereas the formal team-building exercise was comprised of a task-related effort in which the participant and confederate engaged in a formal role clarification and goal-setting exercise. The participants also experienced one of two teammate types: a human or autonomous agent. To measure trust, the Jian et al. (2000) Trust in Automated Systems survey was administered at the end of the mission, along with behavioral trust measures such as the amount of communication exchanged between the participant and teammate. After conducting a 2 x 2 between-subject Analysis of Variance (ANOVA) to determine the effect of agent type and team building, results revealed participants in the formal team building condition significantly trusted their teammate more ($p<0.05$) regardless of agent type. Behavioral indicators of trust also indicated that participants who engaged in the formal team-building exercise exhibited more frequent chat communications between their teammates.

Along with team-building exercises directly before the beginning of a mission, other instances of prior experiences with an autonomous system or agent have also been studied in the HAT literature. Dikmen et al. (2017) conducted a study that examined how trust and reliance in automated driving systems are calibrated based on the driver's experience with the system. For this experimental task, participants performed a set of typical driving tasks in a Tesla such as switching lanes, turning on cruise control, accelerating/decelerating, and parking the vehicle. The study manipulated the autonomous systems within the Tesla: Autopilot and Summon. Autopilot is a combination of lane steering assistance and adaptive cruise control, whereas Summon is an automated parking system allowing vehicles to maneuver into and out of garages using a smartphone application. The study recruited a total of 99 Tesla users. Self-report trust surveys, which consisted of a 5-point Likert scale measuring trust and confidence in the autonomous system were administered before and after the experimental task to capture initial and current trust levels. Additionally, types of past experience (incident vs. no incident) and the frequency of system use with the autonomous systems was captured on a Likert scale. To compare the impact of trust amongst different driving experiences (incident vs. no incident), the study conducted a 2x2 ANOVA in which initial trust and current trust was the within-subjects factor, then Autopilot and Summon incidents served as the between-subject factor (incident, no incident). The results found that individuals who use the systems more often, also reported higher trust scores in Autopilot and Summons (p<0.001 eta squared=0.67). In other words, users reported higher levels of trust as they gained more experience with the autonomous system, regardless of whether had an incident with the system in the past.

As the previous study found higher frequency use increased operator trust regardless of the operator's negative experience in the system, Hafizogly et al. (2019) argues the importance of the type of interactions a teammate should have in order to grow and cultivate trust in agents. To examine this, they conducted an experiment in which participants played the Game of Trust (GoT), consisting of a two-player team game where both players were instructed to complete a total of five teammate-dependent interactions. The goal of this task was to finish as many tasks as possible. Each participant played two games, in which the purpose of the first game was for the participants to gain experience, and the purpose of the second game was to measure the effects of the participant's previous interactions. The study manipulated the experiences the participant encountered with the agent prior to partaking in the Game of Trust, which included positive experiences, negative experiences, and no experience with the agent. Positive past experiences refer to participant interactions with trustworthy agent teammates in previous teamwork instances. Whereas the negative past experiences consisted of the agent making unfair choices for the participants. The goal of the design for the negative past experiences was for the participant to believe their teammate was inclined to exploit them whenever there was a chance. Experience type served as the between-subjects factor. The dependent variables measured included the trust level (measured on a 5-point Likert scale) and the cumulative game results (total number of team goals and subtasks achieved), and the number of excess subtasks performed. These metrics were used to analyze the relationship between past experience and team performance. Results from the study found that participants who had positive past experience exhibited higher trust in agent teammates, whereas negative past experience hindered trust growth during the experimental task ($F(1,198)=3.29$, $p<0.1$). In summary, these findings provide evidence that positive prior experience with virtual teammates helps trust growth in HATs.

As presented in this section, there are conflicting findings in which different types of experiences yield contrasting trust results. For example, with tasks regarding automated driving systems, negative experiences such as incidents did not affect an individual's trust levels (Dikmen et al., 2017). Whereas Hafizogly et al. (2019) suggests negative experiences degraded an individual's trust levels. As there is no apparent trend on how previous experience affects an individual's trust levels, further research is needed to clarify conflicting results. For the current study, the agents demonstrated high levels of reliability (90%), in which the majority of agents assistance was useful. Based on the combination of findings from Hafizogly et al. (2019) in which positive previous experience increase trust and Dikmen et al. (2017) in which more frequency with a system increases trust, the hypotheses forthe current study states that previous experience will significantly and positively influence an individual's trust in a multi-agent team context.

LOA

LOA and Trust

Current technology has reached fully autonomous capabilities, in which input from the human operator is no longer needed to carry out low-level tasks. However, despite these capabilities, it is important to distinguish how different

LOAs may affect the operator, in relation to the task at hand. LOA is defined as "the range of design options implemented in a system to enhance self-sufficiency and self-directedness; ranging from manual operations which require humans to complete all functions, to fully autonomous operations, in which the system is able to perform the task in its entirety, requiring no assistance" (Johnson et al., 2011). Studies have shown that an agent's level of autonomous decision-making can affect an operator's trust, level of understanding of the task, and decision-making (Azhar & Sklar, 2017). Due to the importance of the task context, different LOAs provide optimal mission performance for certain tasks, which can cause repercussions or benefits to operator trust (Schneider et al., 2002). For further analysis of the effects of LOA on an operator's trust, the following section provides an overview of LOAs for trust in HATs.

In Azhar & Sklar (2017), a study was conducted to compare the effects of LOA and agent decision-making fidelity on operator trust, workload, and performance. Participants were instructed to collaborate with a robot to successfully find a treasure in a virtual maze. For the task, the participant and robot were required to make three decisions. At the first checkpoint, the team decided where to start the search, which was consequently followed by the second decision where the team determined the most efficient route the robot should take to reach its desired destination. Lastly, in the third decision, the robot and participant determined whether the images taken by the robot correctly identified the treasure. Each participant interacted with two robots that had two interaction modes: human-as-collaborator and human-as-supervisor. In the human-as-collaborator condition, the agent's actions were based on a shared-decision schema, this way the human and robot interacted as collaborating peers. In other words, the human and robot would work independently for the first and second decision points, then reconvene and discuss the final path the robot should follow. For the third decision point, the robot would send images to the human, in which the human would decide whether or not the images contained the treasure. In the human-as-supervisor condition, the human and robot did not share decisions, and the human only interacted with the robot in a supervisory capacity. More specifically, the participant was instructed to provide commands to the robot, without any questions or feedback from the robot. Another manipulated variable participants experienced was the type of robot: physical (the robot was in person with the participant) or simulated robot (robot only appeared on the screen of the virtual environment). The study utilized a between-subjects design, in which each participant interacted with one agent type in the experiment. The study used four performance metrics (deliberation time, execution time, the length of path traveled by the robot, and the total score in the game), workload via the NASA-TLX, and subject survey responses on a 7-point Likert scale to measure trust. Results from the study found that levels of trust were higher in the human-as-collaborator mode (F(1, 116)=21.36, p<0.05) compared to human-as-supervisor condition. Furthermore, participants in the human-as-collaborator condition performed statistically better than individuals in the human-as-supervisor condition (F(1,116)=27.32, p<0.05). This study presents the relationship between LOA and trust, but also LOA and performance, demonstrating the importance of shared-decision making in a target acquisition task context.

As the previous study investigated the effects of supervisory vs. collaborative LOAs, Ruff et al. (2002) compared the effects of lower-level LOAs in a more complex and dynamic HAT environment. Ruff et al. (2002) analyzed the impacts of different LOAs on the number of simulated remotely operated vehicles in a suppression enemy air defenses (SEAD) task. The experimental task simulated a SEAD mission scenario, in which participants were instructed to monitor a set of unmanned aerial vehicles (UAVs) whilst ensuring all UAVs correctly identified enemies and friendlies throughout the mission. The UAVs were programmed to follow a predetermined flight path, with the ability to be overridden by the participant. If an enemy was found on the UAVs flight path, the UAV would fire upon it, whereas when a target was identified as friendly, the UAVs would avoid it. Three LOAs were utilized in the study, including manual control (automation is dormant until initiated by operator), management by consent (automation proposes action, but cannot act without operator consent), and management by expectation (automation acts without consent, specific operator commands required to cancel automation). Another IV included in the study was the UAVs' decision-aid fidelity, which consisted of 100% and 95% accuracy, along with the number of supervised vehicles which fluctuated between one, two and four supervised vehicles. The decision-aid fidelity served as the between-subject factor whereas LOA and the number of supervised UAVs were within-subject factors. To assess the impacts of the different conditions, mission efficiency, mission performance, workload, and trust in automation were measured after mission completion. Trust in automation was based on subject trust ratings on a survey from Masalonis and Parasuraman (1999). The study conducted a three-way mixed design general linear model (GLM) between fidelity, LOA, and number of UAVs with respect to trust ratings. Results revealed a significant interaction effect of the three variables (F(4,40)=14.59, *p* <0.001). The study found that in the 95% fidelity group as UAVs increased, trust scores were significantly lower in the management-by-consent and the management-by-exception. On the other hand, in the 100% fidelity group as UAVs increased, trust scores were significantly higher in the manual control and management-by-consent LOA compared to the management-by-exception condition. In conclusion, the study found higher trust ratings were present when participants used lower-level LOAs, even as the number of UAVs increased.

There is an evident pattern between LOA and trust in the HAT literature, in which trust is typically higher during low-level LOAs and typically lower during high-level LOAs. However, this pattern is not consistent across all HAT studies, as the task's nature and other task-related variables can heavily influence the dynamic between LOA and operator trust. For example, in Khasawneh et al. (2019) different feedback delays and automation levels were analyzed to investigate the impact upon operator performance, trust and workload. The experimental task took approximately an hour to complete, in which participants collaborated with a UAV to complete two search and rescue missions in a virtual environment. Two different types of LOA were present in the study including manual control (participant had full control to navigate the robot through the environment), and semi-autonomous control (participant had full control of the robot for only advanced maneuvering such as deciding where to go during intersections). Furthermore, latency level was manipulated, which is defined as the amount of time the participant experienced between their input to the UAV and the UAVs response. Participants experienced one of two lag types during each mission, which included no lag or a 500 ms lag. Lastly, system complexity was also manipulated, by the participant controlling one UAV, or two UAVs simultaneously. This study utilized a mixed-subject experimental design in which automation level and latency level served as the between-subject factors, and the number of robots controlled by the operator was the within-subject factor. Dependent variables included operator performance, which was measured by the time to complete the task and operator error rate. Workload was also assessed using the NASA-TLX, which was completed after each mission and Jian et. al (2000)'s trust questionnaire was administered at the end of every mission. Furthermore, a single-item self-report measure was administered every two minutes to capture real-time trust in the agent. Results from the study did not find any significant differences in trust across the 2 automation levels (p=0.318). However, the study found that, when in control of two robots, participants' trust scores were significantly lower in the semi-autonomous condition compared to the manual condition (F(2,76)=8.62, *p*=0.004, eta squared=0.1). This may suggest that participants' controllability over the agent plays a significant role in their capacity to trust the automation, which explains why participants typically have higher trust ratings in manual LOA conditions. Additionally, real-time trust ratings demonstrated lower trust levels in the one-robot condition compared to the two-robot condition in both LOAs. As participants exhibited lower trust scores in low-workload scenarios (e.g., one robot condition vs. two robot condition) this may suggest that workload may have a greater effect than LOA in regard to trust scores in a search and rescue mission context. This finding may provide further insight into the effect of automation on workload which inherently affects trust scores as well.

As demonstrated in the previous study, there is an interaction between workload and LOAs regarding impact on trust. For example, low-level LOAs result in more operator control, inherently increasing workload. Whereas higher LOAs, in which the automation is alleviating task load, inherently decrease operator workload. In a study by Nam et al. (2018) trust and workload were assessed through various LOAs in a search task. The experimental task simulated a target search mission, in which the operator worked with a swarm of 32 homogenous robots to search through a virtual environment. The goal of the task was to successfully identify 100 hidden targets in the environment. Three LOAs were employed in the study including a manual condition (operator chose headings for the swarm), mixed-initiative condition (swarm control was switched from human to swarm or vice versa when performance declines), and the fully autonomous condition (swarm was redirected automatically). Participants could switch freely between different LOAs at any point in the mission. Performance was assessed by the number of targets found during the mission and workload was measured via the NASA-TLX. Trust was measured through a sliding trust scale from -10 (strongly distrust) to +10 (strongly trust) which was collected at 30-second intervals and was encouraged to be adjusted at any time the participant felt trust altered. After each mission, the study administered a self-developed trust survey to capture post-trust scores. Results revealed that participants exhibited higher trust when they had higher levels of control over the swarm, which is consistent with findings from Ruff et al. (2002). Furthermore, the study found a significant difference in post-trust scores across LOAs (F (1.37, 19.20) = 7.80, *p* =0.007) and average trust feedback values (F (2, 57) = 3.35, *p* =0.042). Specifically, the study found that participants provided higher trust ratings in the manual LOA compared to the autonomous LOA. Interestingly, when participants switched to the Mixed-Initiative LOA, workload and trust scores remained relatively constant. Additionally, task performance was also the highest in the manual LOA, however, workload ratings were reportedly higher in the manual LOA compared to the autonomous LOA. Results from the study suggest participants trust their autonomous agent more when they are in full control (manual LOA) and have higher task performance scores, even though they experience more workload.

Based on previous studies, there is an evident relationship between trust and mission performance, in which high trust is present as higher mission performance scores are reported. Kohn et al. (2020) suggests this is caused by differing impacts of trust on a human operator's behavior. For instance, if a human operator had low trust in an agent, this may lead to the human operator rejecting useful recommendations by the agent, leading to degraded mission performance. Azhar & Sklar (2017) found that higher trust scores were present when mission performance scores increased in the human-as-collaborator mode. Nam et al. (2018) also found similar results in which higher trust scores were present in the LOA in which task performance was significantly higher compared to the autonomous LOA. Based on the results from the previous studies, these findings provide support for the study hypothesis that states that trust will significantly and positively predict mission performance in a multi-agent teams context.

LOA and performance

The HAT literature has demonstrated that different LOAs not only affect trust, but also influence performance (Nam et al., 2018; Azhar & Sklar, 2017). More specifically, operators may perceive agents differently across LOAs, which can affect performance. In Narayanan et al. (2014), a study was conducted to examine the effects of two LOA types and how they impact human-agent collaboration, task performance, workload, and situational awareness. In the study, participants were given a limited amount of time to search through as many rooms as possible and successfully report the number of casualties present in a room in a virtual environment. To accomplish the task, participants interacted with a robot teammate who was inside the environment. To promote teammate collaboration, the simulation was designed so that certain regions would not be accessible for both teammates. For example, the participant would not have access to a door, whereas the robot had access to a neighboring room, to access the locked door. To prevent the human operator from micromanaging the robot, the participant was provided with a secondary visual task, which involved solving a three-dimensional visualization puzzle. Two types of automation were administered for this experiment: peer-to-peer and supervisory LOA, which acted as between-subject variables. The peer-to-peer condition instructed both teammates to plan the completion of the task separately, then inform each other of what they had planned. In the supervisory condition, the robot would request permission from the human to carry out actions. As the agent and human operator dynamically work together in this LOA, the peer-to-peer condition is a higher LOA compared to the supervisory LOA. Post-study questionnaires were administered at the end of the study to capture mental workload, situational awareness, complacency, automation effectiveness, likeability, and trust in robots (Parasuraman, 2000). Team performance was also measured through the number of correctly identified casualties found in each room, along with the number of rooms examined by the team. Results revealed that participants exhibited higher performance levels in the peer-to-peer LOA compared to the supervisory LOA in the primary task ($F(2, 19) = 19.56$, $p < 0.001$). This finding was also consistent with the participants' higher perceived likeability scores towards the peer-to-peer teaming condition. Furthermore, the study found that situational awareness scores were not impacted by the peer-to-peer teaming conditions. Overall, this study illustrated that certain LOAs that orchestrate shared-decision making led to higher performance levels along with higher scores of agent likeability.

Mid-level automation types in which the agent and teammate perform separate tasks synchronously have shown to be more flexible compared to lower level automation types in which the human operator is predominantly performing the task, or in higher level automation types where the agent is predominantly performing the task. In Valero-Gomez et al. (2011) different configuration types within LOAs (statistic adjustment vs. flexibility autonomy adjustment) were assessed to determine which condition worked better for the operator with the greater number of robots present. The experimental task was a search and rescue task, in which participants operated robots in a virtual environment. Within the task, participants could freely switch between four varying LOAs, ranging from manual LOA to a fully autonomous LOA. The LOAs included: teleoperation mode (participant controls robot path and sets the speed of the robot), safe teleoperation mode (participant controls robot path and set linear and angular speed control values and the robot used the parameters to maneuver its way around the area), shared-control mode (participant set a target point, robot attempted to reach it), and full autonomy mode (no operator input was needed). The two configuration types (static vs. flexible configuration) followed an adjustable autonomy paradigm, in which the operator was in full control of the autonomous system. The static adjustment configuration allowed the operator to choose the LOA and input commands according to the selected LOA. For example, if the operator selected a target point when working in teleoperation mode (which is a command that cannot be conducted in teleoperation mode), the system would change to the shared-control mode to carry out the human operator's command. In the flexible autonomy adjustment configuration, the operator was still in control of choosing the autonomy level, however, the operator could freely give commands to the agent regardless of whether the specific command could only be carried out in a specific operation mode. For example, if the operator selected the shared-control mode, it is the operator's job to set the robot's target point in which the robot would plan a path to follow. In the event the robot path's gets blocked and is unable to reach the target point on its own, the operator can briefly manually take control over the robot (can only be done in teleoperation and safe teleoperation mode). After the operator finishes assisting the robot, the robot would resume its task in the shared-control mode. In other words, the flexible configuration allowed operators to intervene at any commanding level independent of the selected LOA. The number of robots varied from one to four, which was utilized as a between-subjects variable. Performance metrics included the amount of area explored, the portion of time the robot was in each operation mode, and stoppage time which measured how long the robot was idle (insinuating the time it takes for the participant to make a decision). Results revealed that participants in the flexible autonomy adjustment group explored more area in the virtual environment compared to participants in the static adjustment group (F(3, 36) = 5.938, *p* = 0.002). As participants in the static adjustment group demonstrated higher levels of stoppage time, the study suggests the participant exhibited higher levels of workload which explains why less area was explored amongst the static adjustment group. Interestingly, the impact of the number of robots within the static adjustment model was not significant, however the impact of the number of robots in the flexible adjustment model significantly improved mission performance (F(2,42)=13.857, *p*<0.001). More specifically, when the number of robots increased two in the flexible adjustment model, more area was explored. In conclusion, this study revealed that the flexible autonomy adjustment condition promoted higher levels of performance, whilst reducing the operator's workload.

Workload

Workload and Trust

Current literature related to HATs typically reveals an inverse relationship between workload and self-reported trust (Hillsheim et al., 2017), however, this pattern is not consistent throughout all HAT tasks (Nam et al., 2018; Grimm et al., 2018). Due to the many different variables that influence trust in HATs, it is paramount that various constructs are studied to examine the degree of impact they may have on trust. To further investigate how workload plays an influential role in the development of trust, this section delineates different studies that have measured the relationship between workload and trust and the implications of their results.

Hillesheim et al. (2017) conducted a study to identify the relationship between individual characteristics (e.g., age, gender, technology experience and PTT) and how these variables influence a user's trust in an autonomous agent.

Participants took part in a space navigator game, in which they were instructed to accurately draw trajectories from spaceships to their corresponding planet (e.g., red spaceship to red planet) in a virtual environment. To complete the mission, the participant cooperated with an agent who assisted in drawing trajectories from spaceships to planets. Within the virtual environment, obstacles were placed for participants to actively avoid. In the event the participant disagreed with the agent's trajectory, the participant had the ability to re-draw the trajectory. The goal of this task was to accumulate as many points as possible, which were counted by the number of correct trajectories drawn in the environment. Each participant first familiarized themselves with the task goal by completing a training session, then played 12 four-minute missions. Within the 12 missions, participants experienced four different agent reliability levels which fluctuated between 95%, 90%, 80%, and 70% in randomized orders. Each reliability level was present in four different missions, whereas the remaining eight missions were provided with a 100% reliability level by the agent. Throughout this study, the participant was not notified of the changes in reliability. It is important to note that task load fluctuated in relation to these different conditions as the participant needed to correct the automation more often as reliability decreased. After each mission was completed, the NASA-TLX and self-reported trust surveys were completed to assess workload and trust. After conducting a multiple linear regression to predict trust, a significant regression equation found that total workload, reliability in the autonomous agent, gender, and education level influenced participants' trust ($F(6, 283)=26.504$, $p<.000$, $R^2=0.273$). For workload, the study found subjective workload inversely correlated to trust in agents (participants’ perceived reliability rate decreased by 0.5% for each point increase in total workload). More specifically, when the participant exhibited low levels of workload, the user’s trust in the agent was very high, whereas when the participant’s workload was high, the user’s trust was low. Additionally, results found that females tended to have higher levels of trust in agents compared to males, and that individuals who hold a college graduate education had higher levels of trust than participants that do not have a college education.

As demonstrated in the previous study, trust in HATs is constructed of a complex paradigm involving a multitude of individual and task-specific variables. However, the influences of these characteristics may fluctuate depending on the nature of the task at hand. In van der Waa (2021), numerous variables were assessed to investigate the impacts of HAT team composition in a dynamic and higher-stress environment. To assess task-specific and individual characteristics of trust, the study monitored several medical experts in a simulated hospital environment. Experts were instructed to assign medical care to incoming patients whilst accounting for the urgency exhibited by each patient and available resources. The domain experts could choose one of three options for incoming patients which included sending the patients home (receiving no care), assigning them to the general ward (receiving moderate care), or the intensive unit (receiving maximum care). For this task, medical experts used an application called MATRX, which provided relevant patient information such as age, profession, current health, symptom severity, and general fitness to simulate a realistic hospital environment. Within this application, a decision support agent in the form of a robot icon appeared on the participant's screen, recommending patient placement. Four LOAs were presented by the agent, which included (1) no involvement from the agent, (2) the agent providing advice, (3) the agent and medical expert placing patients synchronously, (4) the agent autonomously placing patients in accordance with the medical expert’s moral values. After the task, semi-structured interviews were conducted and surveys were completed to capture the expert’s perceived control over the task in relation to the LOAs, trust in the agent, workload, and level of team collaboration, along with a brief explanation for their answers. According to the qualitative comments from the interviews and questionnaires, the study found that experts felt more in collaboration with agents when they had more control over the agents (lower LOAs). However, this only occurred when the participant felt that sufficient time was present. In the event the medical expert experienced high time pressure in lower LOAs, lower trust scores were exhibited. The study also found higher trust scores were present when the agent reduced workload, whereas high workload and stress resulted in lower trust. These findings are consistent with the majority of the literature, which shows an inverse relationship between workload and trust in the context of fluctuating LOAs.

In the previous studies, the HATs were composed of one human and one agent, which is a common configuration in the HAT literature. However, there are also many instances in which HATs involve a human operator interacting with more than one agent, or two human operators interacting with one agent. For less common HAT compositions, there is a limited amount of research evaluating the effects of variables such as workload and trust, which is essential in understanding how variables may affect workload or trust differently in varying team configurations. In Grimm et al. (2018), fluctuations in workload were imposed by a combination of technological failures in a simulated Remotely Piloted Aircraft System (RPAS) testbed in which two participants collaborated with one agent. The objective of the experimental task was to reach as many critical waypoints in a virtual environment within a 40-minute period. Two participants played the role of the navigator and photographer, whereas the experimenter was assigned the pilot role and acted as the agent (i.e., Wizard of Oz method). The main manipulation in this study was the application of three types of agent failures which varied in degree of intensity, in which type 1 was the least intense, and type 3 was the most intense. A type 1 automation failure occurred when the pilot could not see the current or next waypoint information, which lasted for 300 seconds. A type 2 failure took place when the current and requested altitude and airspeed settings were not visible to the pilot, which lasted for 420 seconds. Lastly, a type 3 failure transpired when all information presented to the pilot was not visible, along with the bearing information to the next target waypoint area, which also lasted for 420 seconds. The team encountered failures at selected target waypoints, in which team members had limited time to overcome each failure. Performance measures including overall team performance (number of waypoints reached), target processing efficiency (time and accuracy of finding a target), and communication flow (message count and the flow of conversations) were assessed throughout the mission. After each session, questionnaires were administered to measure team coordination, team situational awareness, and trust. The NASA-TLX was also completed to measure workload. Results compared metrics between high-performing teams (groups that exhibited high-performance scores) and low-performing teams (groups that demonstrated low-performance scores). High-performing teams demonstrated effective team communication through high message counts, whereas the low-performing team failed to communicate in the event the agent exhibited a failure. Furthermore, as workload increased through the imposition of technological failures, low-performing teams provided relatively consistent trust scores on the agents over time. On the other hand, in high-performing teams, as workload increased through the imposition of technological failures, trust scores on the agents decreased over time. Due to the differences in results between a high-performing and low-performing team, this study demonstrated the delicate relationship between trust, workload, and performance in a dynamic HAT context. As workload and trust scores differed between the two groups, the study suggests future research should analyze different team characteristics that can correlate with HAT-related variables such as workload and trust under various performance conditions.

The previous studies from this section commonly found that workload was inversely related to trust. These studies support the hypothesis that workload will significantly and negatively influence an individuals trust in a multi-agent teams context.

Workload and Performance

Literature reviewed in the previous section illustrated a clear relationship between workload and trust, which varies depending on the task, the number of other teammates present, and the performance level exhibited by the team. To take a deeper look into the effects of workload and performance, the following section demonstrates the significance of this relationship from current HAT literature.

In Levinthal & Wickens (2006), a study was developed to assess how different levels of workload impact performance in HATs. Participants were instructed to perform two tasks in a UAV simulator: a UAV task and a tank detection task. The first task involved the participant navigating UAVs through a series of waypoints where the participant performed arithmetic operations between the UAVs current X and Y coordinates. The participant was told to select a Northside waypoint if the calculated value was greater than 50, or a Southside if the value was less than 50. Simultaneously, participants performed a tank detection task on the adjacent display. The goal of the second task was to find enemy tanks as quickly and accurately as possible. When participants found a tank, they responded "TANK" and pointed to the target. An automated target recognition aid was included in the tank detection task, which differed in automation reliability. The three levels of automation reliability included A90 (90% reliable, equally composed of false and missed alarms), FAP (60% reliable, with a 3:1 likelihood of committing false alarms over misses), and MP (60% reliable, 3:1 likelihood of committing misses over false alarms). Automation reliability was utilized as a within-subject variable. Workload was also manipulated by the number of UAVs required to be monitored, which was divided into two separate conditions: low workload (two UAVs) and high workload (four UAVs), serving as a between-subject variable. Performance metrics were measured by calculating the amount of time it took for participants to calculate the next waypoint, which was defined as "idle time". Results demonstrated that increasing workload was associated with deteriorating performance. For example, as workload increased from low to high, participants exhibited longer idle times ($F(1,72) = 248.3$, $p < .001$). More specifically, the time it took for participants to calculate the UAVs next waypoint tripled from approximately 600 seconds to 1800 seconds from low and high workload conditions. Additionally, results found there was no effect on the accuracy of the tank detection task among the different automation types. In conclusion, Leveinthal & Wickens (2006) found that performance decreased as participants reported higher workload scores.

Like the previous study, Zhang & Yang (2017) executed a similar task to investigate the effect of workload and automation aid on dual-task performance, trust in automation, and attention allocation in a simulated surveillance and detection task. For this experimental task, participants performed two tasks simultaneously in a desktop simulation, which were presented on two separate displays adjacent to each other. For the first task, participants monitored photos from an unmanned ground vehicle (UGV) to detect threats, in which automation aid was available to help identify enemies. In the second task, participants navigated the flight paths of two unmanned aerial vehicles (UAVs) through a series of waypoints. The flight between waypoints was automated, however, operators were required to select the next correct waypoint. To do so, participants were told to sum the x and y coordinates of the UAV and select the northmost waypoint if the sum was greater than 100 or select the southmost waypoint if the sum was less than 100. Workload and automation aid served as the independent variables for this study, in which two separate workload conditions and five varying automation aids were present. The researchers manipulated workload by fluctuating the time intervals the UAVs took to fly to each waypoint. This time interval fluctuated from 7.5 seconds in the high workload condition to 15 seconds in the low workload condition. For the second independent variable, automation aid had five levels: a non-automated baseline (BL), a 67% reliable aid with false alarms (67FA), a 67% reliable aid with misses (67M), a 67% aid with equal numbers of false alarms and misses (67MPFA), and a 100% reliable aid (100A). The experimental design utilized a mixed design, with workload as the within-subject factor and the automation aid as a between-subject factor. Task performance was measured by the number of correct identifications of threats and the amount of time participants took to identify an enemy in the detection task. In the waypoint task, performance was measured by the number of correct waypoint selections and the time it took to calculate the next waypoint. Furthermore, attention allocation was measured via eye-tracking devices to calculate the sum of all fixations between the two separate displays. Higher attention allocation for the detection task suggests the individual was more fixated on that specific display. Participants experienced one condition, which consisted of 18 trials, each lasting 30 seconds. After each trial, operators reported their trust in the automation aid and their confidence in performing the task. The specific questionnaire to collect trust scores was not specified in the publication. Results revealed that higher levels of workload led to longer response times (F(1,35) = 126.316, *p* < .001) and lower attention allocation (F(1,35) = 126.316, *p* < .001) in the detection task. However, higher levels of workload led to shorter response times (F(1,35) = 4.12, *p*= .05) and higher attention allocation in the waypoint task (F(1,35)=33.561, *p* < .001). Furthermore, workload had no effect on accuracy levels in the detection task, whereas higher workload led to lower accuracy levels in the waypoint task (F(1,35) = 8.066, *p*= .007). No differences were found in subjective trust scores across different workload or automation aid type (F(1,28) = .009, *p*= .923). In summary, this study revealed that higher workload led to accuracy decrements in the waypoint task, with no effect on accuracy to the detection task. These findings conflict with results from previous studies like Levinthal & Wickens (2006), suggesting further research is needed to explore the influential factors of workload in a multi-agent team.

McBride et al. (2021) studied how participants interacted with varying levels of imperfect automation to see how this affected workload and automation compliance. Participants acted as “warehouse managers” where they oversaw two tasks: (1) receiving packages into inventory, and (2) dispatching trucks once they were filled to capacity. For the task of receiving packages, participants were given a target barcode, which they had to match from a list of barcodes. If the participant failed to find the matching barcode after 7 seconds, or incorrectly entered the wrong barcode, points were deducted from their final score. If the participant correctly picked the barcode, points were added. In the second task, automation aid was provided to help the participant determine when trucks should be dispatched. For example, when the automation detected a truck was full, the participant was alerted and would approve of the automation's statement by pressing a dispatch key. If the participant failed to notice a truck was full after 10 seconds, points were deducted. The goal of the overall mission was to accumulate as many points as possible. Three different workload conditions were utilized between participants for the barcode task by altering the number of characters present in each barcode along with the list of possible matches included in the barcode list. The three workload conditions were low workload (3 characters in the barcode, with 3 barcodes in the list), moderate workload (4 characters in the bar code, 6 barcodes in the list), and high workload (6 characters in the barcode, 11 barcodes in the list). The NASA-TLX was administered after each mission to assess workload. Additional dependent variables included reliance (number of times the truck was not viewed when no alert was present, suggesting the participant trusted the automation was correctly performing the job), and compliance (number of times participants approved of the automation message without viewing the truck). Furthermore, performance measures were administered which included the number of correct barcode matches, incorrect barcode matches, time outs, trucks dispatched on time, dispatched trucks that weren't full, and dispatched trucks that were overloaded. Results revealed that the workload manipulation had a significant effect, with higher workload leading to a reduction in number of correctly matched barcodes (F(2, 39) = 58.01, $p < .01$, η2 = .74). More specifically, in the low workload conditions, participants presented a greater percentage of correctly matched trials. This also affected performance in the dispatch trucks task, in which the high workload group achieved a lower percentage of correctly dispatched trucks (t (39) = 2.63, $p = .01$). As a result of higher workload, higher levels of compliance was also observed (t (39)= -2.10, p=0.04), whereas reliance on automation was not significantly affected by the differing workload conditions ($p>0.16$). Findings from this study suggest that high workload levels lead to degradations in performance as well as higher levels of compliance in the automation.

In summary, studies included in this section generally reported that mission performance was degraded as workload increased. Findings from these studies provide support for they study hypotheses that workload will significantly and negatively influence an individual’s mission performance in a multi-agent team context.

Chapter 3
Methodology

Introduction

For the current study, archival data was collected from a previous experimental study (Rebensky et al., 2022) that examined the impacts of varying LOAs on mission performance, trust, and team effectiveness in multi-HAT missions. The experimental study was conducted previously in the Advancing Technology-Interaction & Learning in Aviation Systems (ATLAS) lab where I assisted in data collection and had access to the full dataset to conduct data analyses for the current study. The previous study collected an individual's current performance, stress, trust, and workload associated with each mission. Findings from the Rebensky et al. (2022) study revealed that participants had higher levels of performance, stress, and workload for the two higher LOAs (i.e., consent, veto)compared to the two lower LOAs (i.e., manual, advice. However, the study reported there were no significant differences in trust scores based on LOA. The current study utilized data from the Rebensky et al. (2022) study to examine if variables that influence trust dynamics in dyadic teams also impact trust in A multi-agent performance context.

Participants

A total of 49 participants completed the Rebensky et al. (2022) study, all of which were between 18-37 years old. Two participants were removed from the dataset due to low English proficiency, potentially impacting their subjective responses. From this, a total of 47 participants were utilized in the current dataset. Further details on the demographic information are presented in Table 4.1 and Table 4.2.

Ethical Considerations

An Institutional Review Board (IRB) protocol was submitted to the FIT IRB, which was later approved as an exempt study. The data set does not include any personal information regarding the participant. Furthermore, this dataset was not shared with anyone outside the research committee. Due to the absence of human participants or identifiable information in the archival study, there was very minimal risks to participants.

Study Design

The current study is an archival research design which utilized a correlational method to examine relationships between variables associated with one group of participants for which experimental data was collected for each of four trials with differing LOAs (Rebensky et al., 2022). This resulted in 188 trials for which data was available. The current study utilized this data set to examine two relationships. First, the relationship between the dependent variable of reported trust in the HAT and independent variables of PTT, previous experience with agents, LOA, and workload ratings was examined. Second, the relationship between the dependent variable of mission performance and independent variables of trust, LOA, and workload ratings was examined.

In the previous study (Rebensky et al., 2022), participants completed a military intelligence, surveillance, and reconnaissance (ISR) mission on a desktop simulator to identify the safest routes to send a convoy. During the task, participants interacted with four unmanned aerial vehicles (UAV) to identify and classify three types of targets: neutrals, friendlies, and enemies. A total of 26 targets were randomly distributed on each map for participants to find. Participants performed four scenarios, each with a different LOA: manual (M), advice (A), consent (C), and veto (V), each lasting 5 minutes, in one of the following counterbalanced orders: MACV, ACVM, CVMA, VMAC. In the manual condition, the agent only assisted in target detection, in which the participant was required to classify the target type and then click confirm. For the advice condition, the agent assisted in detecting the targets, and recommended the classification type to the participant. However, the participant was still required to classify the target type, then click confirm. For the consent condition, the agent detected and provided advice on the targets, but the target classification type was already pre-selected. From this, the participant was only required to click confirm if they agreed with the agent's classification. For the veto condition, the agent detected, classified, and confirmed all events, and participant input was not required for this condition unless they wanted to change the selection. Participants completed a range of individual difference measures prior to the experiment and workload, trust, and stress measures after each trial. A portion of these measures were utilized in the current study and are discussed in the following section.

Measures

The current study collected individual difference measures prior to interaction with the agents such as PTT. Further, three measures were collected during each trial of the experimental study, including mission performance, trust, and workload. The remaining two measures: previous experience and LOA were determined based on the current and previous conditions.

Propensity to Trust in Technology (PTT) Questionnaire

Participant’s Propensity to Trust (PTT) in agents was assessed using the Propensity to Trust in Technology (PTT) questionnaire, which was administered at the beginning of the study, and is a unidimensional questionnaire used to assess an individual’s general tendency to trust in automation. This questionnaire was developed by Schneider et al. (2017) and includes 6-items designed to measure the characteristics such as attitudes toward technology and the potential for collaboration with technology, where scores can range from 0 (strongly disagree) to 5 (strongly agree; Schneider et al., 2017). This questionnaire has an internal consistency with a Cronbach's α = .76 and convergent validity with perceived trustworthiness *r*=0.47 (Jessup, 2019).

Previous Experience with Agents

To assess the participant’s previous experience with the agents, the number of missions the participant completed prior to each trial was computed. A total of four categorical groups were identified from 0 (no experience) to 3 (three completed missions). For instance, after a participant completed their third mission in the study, their experience level was coded as “three LOA experienced”. On the other hand, if the participant was beginning their first mission, their experience level was coded as “zero LOA experience”.

LOA

To assess the participant’s current LOA, the specific type of LOA was determined for each mission the participant experienced. This resulted in four groups: Manual (1), Advice (2), Consent (3), and Veto (4). For example, if the participant was in the manual condition, this was counted as a “1” for the manual condition.

Workload

To assess the participant's mental workload, the NASA-Task Load Index (NASA-TLX) was administered after each trial. This measure has proven to be a reliable indicator of workload, as it has been used in over 500 studies as a subjective workload measure in varying contexts (Hart, 2006). A total of five items were included in the NASA-TLX, each of which represented five dimensions of workload, including mental demand, physical demand, temporal demand, and performance effort. The total workload score from the NASA-TLX was used as the workload rating for this archival study. For validity, the NASA-TLX has been shown to correlate with other workload measures and subjective ratings of mental workload (Longo, 2018).

Trust

To assess subjective perceptions of trust, trust ratings were completed for each agent at the end of each trial. Specifically, to calculate overall trust in agents, participants were asked to rate each agent on a sliding scale from 0 (no trust) to 100 (complete trust). To calculate the overall trust scores, trust scores from each of the four agents were averaged for each trial.

Mission Performance

To assess mission performance, the percentage of correctly identified targets out of the 26 targets present in each map and was calculated as the ratio of the number of targets correctly identified to the number of total targets. For example, if a participant missed two targets, their mission performance score would be 92% (24/26 *100).

Procedure

For the current study, the data was retrieved from the database of the experimental data that included all survey and performance data. To clean the data, relevant metrics were extracted, including trust, mission performance, PTT scores, previous experience with agents, LOA, and workload ratings. In all, the resulting database included data from 47 participants, in which each participant had four rows, representing each trial they experienced. Within each participant's four rows, the PTT scores remained the same as this was an individual difference measures collected only once at the beginning of the study. PTT scores in the database was the sum of the six subscales included in the survey. Workload scores in the database were also the sum of the NASA-TLX scores which consisted of five subscales. Mission performance scores were calculated by the number of correctly identified targets out of the 26 targets found in the map. Lastly, trust scores were averaged from individual trust ratings in the agent. Participant scores for previous experience with agents, LOA, and workload ratings differed throughout each of the four rows as they progressed through trials in the study and experienced different conditions. After this, further processing of the data was completed to transform the data into a more appropriate format for the multiple regression analysis. More specifically, one categorical variable, current LOA, was dummy coded into appropriate formatting for the multiple regression analyses as illustrated in table 1. For the current LOA, the manual LOA served as the reference group. The three remaining categories (advice, consent, and veto) were coded against the reference group. On the other hand, previous experience was categorized as an ordinal variable, which was coded based on the number of trials experienced by the participant.

Table 1

Dummy Coding Strategy for X_3 = Current LOA

Level of Automation	***Advice***	***Consent***	***Veto***
Advice	1	0	0
Consent	0	1	0
Veto	0	0	1
Manual	0	0	0

Data Analyses

The current study conducted two multiple regressions to investigate the relationship between individual variables and two different criterion variables: trust and mission performance. This is reported in Tables 3.2 and 3.3. With respect to the first regression analysis, trust was the dependent variable with X_1 = PTT, X_2 = previous experience, X_3 = current LOA, X_4 = mission performance as independent variables. With respect to the second regression analysis, mission performance was the dependent variable withs X_1 = Current LOA, X_2=workload, and X_3 = Trust as the independent variables.

Table 2

Regression Model 1

Dependent Variable	Definition
Y_1 = Trust	Y_1 is a continuous variable, overall trust scores in agents
Independent Variables	
X_1 = PTT	X_1 is a continuous variable, total scores from the PTT questionnaire
X_2 = Previous Experience	X_2 is a ordinal variable, total number of missions experienced by the participant
X_3 = Current LOA	X_3 is a categorical variable, current LOA experienced by the participant
X_4 = Workload	X_4 is a continuous variable, total scores from the NASA-TLX
X_5 = Mission Performance	X_5 is a continuous variable, percentage of targets correctly identified

Table 3

Regression Model 2

Dependent Variable	Definition
Y_2 = Mission Performance	Y_2 is a continuous variable, percentage of targets correctly identified
Independent Variables	
X_1 = Current LOA	X_1 is a categorical variable, total number of LOAs experienced by the participant
X_2 = Workload	X_2 is a continuous variable, total scores from the NASA-TLX
X_3 = Trust	X_3 is a continuous variable, overall trust scores in agents

Preliminary Analysis

A preliminary analysis was conducted for the archival dataset. A total of two outliers were removed due to the participant's low proficiency in English, potentially leading to skewed mission performance data from the previous experiment as mission goals may not have been understood correctly. The final sample size for each regression was 188. The predictors' variance inflation factor (VIF) values were calculated to assess multicollinearity. All predictors had a value less than 10, showing no multicollinearity amongst the independent variables (i.e., no independent variables showed any significant relationships with each other).

Regression Assumptions

Next, regression assumptions were checked. Both models were tested for the six-regression assumptions.

Multivariate Linearity. For assumption 1, two bivariate scatter plots were created in which the residuals and predicted values of the criterion variables (trust and mission performance) were plotted against each other. A smoother kernel line was layered onto each plot, revealing that lines from both plots hugged the zero line. This demonstrated that all independent and dependent variables were linear. From this, assumption 1 was satisfied for both linear regressions.

Correct specification of the IVs. For assumption 2, leverage plots were made for each IV to examine their relationship to the DV. P-values greater than 0.2 reveal no relationship between the IV and the DV and are recommended to be taken out of the linear regression. The leverage plots for both regressions revealed that several IVs did not meet this assumption. However, despite these results, IVs with a p-value greater than 0.2, and with significant theoretical and empirical support, were still included in both linear regressions.

Reliability. For assumption 3, all instruments used to collect data should present good reliability coefficients. For the PTT questionnaire an adapted version of the questionnaire was utilized to measure an individual's general tendency to trust in automation. By adapting the PTT to use the term "automated agent" increased reliability from $\alpha = .76$ to $\alpha = .84$ (Jessup, 2018). The adapted PTT accounted for a significant amount of variance in perceived trustworthiness (Jessup, 2018). Moreover, workload was assessed using the NASA-Task Load Index (NASA-TLX). The NASA-TLX has been shown to have high test-retest reliability with a Cronbach's alpha of .83 (Hart & Staveland, 1988) and .75 (Longo, 2018).

For validity, the NASA-TLX has been shown to correlate with other workload measures and subjective ratings of mental workload, as well as being sensitive enough to detect changes in workload (Longo, 2018).

Homoscedasticity of the Residuals. For assumption 4, the variance of the dependent variables (i.e., trust and mission performance) must be the same for all independent variables. However, because Assumption 1 was met, then Assumption 4 was also satisfied for both linear regressions.

Independence of the Residuals. For assumption 5, a bivariate plot of the residuals versus the case numbers was created for each linear regression. The kernel smoother line was imposed on the model, revealing the cases were randomized for both bivariate plots. From this, assumption 5 was satisfied for both linear regressions.

Normality of Residuals. For assumption 6, a histogram of residuals was plotted, in which a normal curve was imposed on the histogram for both regression models. Second, a normal q-q plot of the residuals was created at the 95% confidence interval. This assumption was satisfied for the first regression, where trust was the criterion variable. However, this assumption was not satisfied for the second regression, in which mission performance was the criterion variable. However, as regression is robust to non-normal data, we proceeded with the regression analysis.

Chapter 4 Results

This chapter presents the results of the current study, which used archival data from a previous study that collected data from 47 participants. Each participant experienced each of the four LOA conditions (manual, advice, consent, veto) resulting in four data points per participant. From this, each data point was counted separately resulting in 188 observations. The first section of the chapter provides an overview of descriptive statistics for demographic variables and independent and dependent variables including trust, PTT, previous experience, current LOA, workload, and mission performance. The second section presents the results of the descriptive statistics, inferential statistics, including the primary analysis for both regression models.

Descriptive Statistics

A total of 47 participants were included in the archival study, in which they were asked to report their age and gender. The average age of the participants was 24 years old, with a standard deviation of 5 years. Furthermore, 29 males were included in the current study accounting for 61.7% of the dataset, whereas 18 females were included in the current study accounting for 38.3% of the dataset. Table 4.1 and Table 4.2 present the descriptive statistics for age and gender.

Table 4

Descriptive Statistics for Age (N=47)

Variable	N	Mean	Median	SD	Range
Age	47	24	37	5	18-37

Table 5

Descriptive Statistics for Gender (N=47)

Variable	***N***	**Result (%)**
Gender		
Male	29	61.7%
Female	18	38.3%

Independent and Dependent Variables

As presented in Table 4.3 the descriptive statistics for the following variables are summarized: PTT *(M=20.4, SD=2.7)*, Workload *(M=58.2, SD=13.9)*, Mission Performance *(M=86.4%, SD=11.12%)*, and Trust *(M=75.3, SD=15.0).* There were also two categorical variables: LOA and previous experience. For previous experience, a total of four categorical groups were identified from 0 (zero completed missions) to 3 (three completed missions). Previous experience did not vary amongst participants, as each participant completed the same number of missions. Furthermore, LOA consisted of four categorical groups including Manual (1), Advice (2), Consent (3), and Veto (4), in which all participants completed each of the LOAs.

As these LOAs were counterbalanced in the original study, there is an equal number of datapoints for each LOA. Due to the nature of these two categorical variables, descriptive statistics were not included as there is an equal number of datapoints in each LOA and previous experience category.

Table 6

Descriptive Statistics for IVs and DVs (n=188)

Variable	*n*	M	SD	Range
PTT	188	20.36	2.668	13
Workload	188	58.16	13.936	81
Mission Performance	188	86.41%	11.12%	73.08%
Trust	188	75.25	15.02	84

Note. The PTT scale consisted of 6 items scored on a 5-point Likert scale (0=low to 30=very high). Workload was based on scores from the NASA-TLX which included 5 items on a 20-point scale (0=very low to 100=very high). Mission performance was scored based on the number of correctly identified targets out of the 26 targets present (0%=no targets found to 100%=All 26 targets found). Trust was averaged from individual trust ratings, which was rated on a 100 point sliding scale (0=no trust to 100=complete trust).

Inferential Statistics

Overview

The primary purpose of the current study was to explore the effects of PTT, previous experience, LOAs, workload, and mission performance on trust in a multi-agent context. The secondary purpose of the current study was to explore the effects of LOAs, workload, and trust on mission performance in a multi-agent context. Multiple linear regression was the research methodology best suited to address the research questions associated with the study purpose, as it can explain the relationship between trust scores and multiple factors such as PTT, previous experience, current LOA, workload, and mission performance.

Primary Analysis 1: Linear Regression Model 1

To investigate the influence of PTT, previous experience, current LOA, workload, and mission performance on an individual’s trust in a multi-agent HAT, a multiple linear regression was conducted. The criterion variable was trust and the predictor variables were PTT, previous experience, current LOA, workload, and mission performance. Table 4.4 contains a summary of the results of these analyses, along with a discussion of the unique contributions each of the predictors made in the regression model. As reported in Table 4.4 the variance explained by the predictors in Regression 1 was significant, R^2=.297, $R^2_{adjusted}$=0.27, F(7,180) = 10.88, p<0.001.

In model 1, it was found that X_1=PTT scores significantly predicted trust scores, $\boldsymbol{\beta}_{1\,=}$ 2.43, p < .001. This demonstrates a positive relationship between PTT scores and trust scores: for every one-point increase in PTT scores, on average, trust scores increased by 2.43 points. Based on this finding, individuals who score high in the PTT questionnaire are likely to have high trust scores.

X_2=Previous Experience did not significantly predict trust scores, $\boldsymbol{\beta}_{2\,=}$ 1.18, p = .183. From this, no significant relationship was found between previous experience and trust scores.

X_3=Advice LOA significantly predicted trust scores, $\boldsymbol{\beta}_{3\,=}$ -5.35, p = .046. This demonstrates a negative relationship between the advice LOA and trust scores: when participants are in the advice LOA, on average, trust decreased by 5.35 points compared to the manual condition. Based on this finding, when participants are in the advice LOA, they are likely to have lower trust scores compared to the manual condition.

X_4=Consent LOA significantly predicted trust scores, $\boldsymbol{\beta}_{4\,=}$ -6.44, p = .020. This demonstrated a negative relationship between the consent LOA and trust scores: when participants are in the consent condition, on average, trust decreased by 6.44 points compared to the manual condition. Based on this finding, when participants are in the consent LOA, they are likely to have lower trust scores compared to the manual condition.

X_5=Veto LOA did not significantly predict trust scores, $\boldsymbol{\beta}_{5\,=}$ -4.89, p = .077. From this, no significant relationship was found between the veto LOA and trust scores.

X_6=Workload significantly predicted trust scores, $\boldsymbol{\beta}_{6\,=}$ -.195, p = .006. This demonstrates a negative relationship between workload scores and trust scores: for every one-point increase in workload, on average, trust scores decreased by 0.2 points. Based on this finding, when participants experience a high workload, participants are likely to report low trust scores.

X_7=Mission performance significantly predicted trust scores, $\boldsymbol{\beta}_{7\,=}$ .34, $p <$ .001. This demonstrates a positive relationship between mission performance and trust scores: for every one-point increase in mission performance, on average, trust scores increased by 0.34 points. Based on this finding, when participants experience high levels of mission performance, participants are likely to report high trust scores.

Table 7

Summary of Multiple Linear Regression Analysis 2

Predictor	***B***	***SE***	***β***	***t***	***p***	***95% CI***	***R²***
Model 1							.297***
X_1 = *PTT*	2.43***	.35	.43	6.89	<.001	[1.73, 3.12]	
X_2=*Previous Experience*	-1.177	.88	-.09	-1.34	.183	[-2.91, 0.56]	
X_3=*Advice LOA*	-5.35*	2.66	-.16	-2.01	.046	[-10.60, -0.10]	
X_4=*Consent LOA*	-6.44*	2.73	-.19	-2.36	.020	[-11.83, -1.05]	
X_5=*Veto LOA*	-4.89	2.75	-.14	-1.78	.077	[-10.32, 0.54]	
X_6=*Workload*	-.20**	.07	-.18	-2.78	.006	[0.16, 0.52]	
X_7=*Mission Performance*	.34***	.09	.25	3.69	<.001	[-0.33, -0.06]	

*Note: *p<.05, **p<.01, ***p<.001*

Primary Analysis 2: Linear Regression Model 2

To investigate the influence of current LOA, workload, and trust on an individual's mission performance in a multi-agent HAT, a second multiple linear regression was conducted. The criterion variable was mission performance and the predictor variables were advice LOA, consent LOA, veto LOA, workload, and trust. Table 4.4 contains a summary of the results of these analyses, along with a discussion of the unique contributions each of the predictors made in the regression model.

As reported in Table 4.5 the variance explained by the predictors in Regression 2 was significant R^2=.107, $R^2_{adjusted}$=0.082, F(5,182)=4.342 at p<0.001.

In model 2, it was found that X_1=Advice LOA did not significantly predict mission performance, $\boldsymbol{\beta}_{1\,=}$ 3.51, p =.114. From this, no significant relationship was found between the advice LOA and mission performance.

X_2=Consent LOA significantly predicted mission performance, $\boldsymbol{\beta}_{2\,=}$ 6.82, p=.003. This demonstrates a positive relationship between the consent LOA and mission performance scores: when participants are in the consent condition, on average, mission performance scores increase by 6.82 points compared to the manual condition.

X_3=Veto LOA significantly predicted mission performance, $\boldsymbol{\beta}_{3\,=}$ 6.31, p=.006. This demonstrates a positive relationship between the veto LOA and mission performance scores: when participants are in the veto condition, on average, mission performance scores increase by 6.31 points compared to the manual condition.

X_4=Workload did not significantly predict mission performance, $\boldsymbol{\beta}_{4\,=}$ -.02, p=.790. From this, no significant relationship was found between workload and mission performance.

X_5=Trust scores significantly predicted mission performance, $\boldsymbol{\beta}_{5\,=}$ .16, p=.003. this demonstrates a positive relationship between trust scores and mission performance: for every one-point increase in trust, on average, mission performance increases by 0.16 points.

Table 8

Summary of Multiple Linear Regression Analysis 2

Predictor	*B*	*SE*	*β*	*t*	*p*	*95% CI*	*R²*
Model 2							.107
X_1 = *Advice LOA*	3.51	2.21	.14	1.59	.114	[-0.85, 7.87]	
X_2 =*Consent LOA*	6.82**	2.23	.27	3.05	.003	[2.41, 11.22]	
X_3 =*Veto LOA*	6.31**	2.25	.25	2.81	.006	[2.41, 11.22]	
X_4=*Workload*	-.016	.06	-.02	-.27	.790	[1.88, 10.74]	
X_5=*Trust*	.16**	.05	.22	3.04	.003	[0.06, 0.27]	

*Note: *p<.05,**p<.01,***p<.001*

Results of Hypotheses Testing

The research questions and corresponding hypotheses of the current study were stated in Chapter 1. The following section states whether the hypotheses were supported by the results of the respective primary analyses reported in this chapter. Table 4.6 also summarizes the results of hypothesis testing.

Table 9

Summary of Results of Hypothesis Testing

Hypotheses	**Result**
H_{1a}: PTT will significantly and positively influence an individual's trust in multi-agent HATs.	Fully Supported
H_{1b}: Previous experience with an agent will significantly and positively influence an individual's trust in multi-agent HATs.	Not Supported
H_{1c}: The agents current LOA will significantly influence an individual's trust in multi-agent HATs.	Partially Supported
H_{1d}: Workload will significantly and negatively influence an individual's trust in multi-agent HATs.	Fully Supported
H_{1e}: Mission performance will significantly and positively influence an individual's trust in multi-agent HATs.	Fully Supported
H_{2a}: The current LOA will significantly influence individual's mission performance in multi-agent HATs.	Not Supported
H_{2b}: Workload will significantly and negatively influence individual's mission performance in multi-agent HATs.	Fully Supported
H_{2c}: Trust will significantly and positively influence individual's mission performance in multi-agent HATs.	Partially Supported

***H_{1a}*: PTT will significantly and positively influence an individual's trust in a multi-agent HAT when controlling for previous experience, current LOA, workload, and mission performance.**

As reported in table 4.3 the variance explained by the PTT predictor was significant (t=6.89, p<0.001). As a result, H_{1A} was fully supported.

H_{1b}: Previous experience with an agent will significantly and positively influence an individual's trust in a multi-agent HAT when controlling for PTT, current LOA, workload, and mission performance.

As reported in table 4.3, the variance explained by the Previous Experience predictor was not significant (t= -1.34, p=0.183). As a result, H_{1B} was not supported.

H_{1c}: The agents current LOA will significantly influence an individual's trust in a multi-agent HAT when controlling for PTT, previous experience, workload, and mission performance.

As reported in table 4.3, the variance explained by the current LOA predictor was significant for the Advice LOA (t= -2.01, p=0.046) and the Consent LOA (t= -2.36, p=0.02) when these LOA were compared to the manual LOA. However, the variance explained by the current LOA predictor was not significant for the Veto LOA (t= -1.78, p=0.077) when compared to the manual LOA. As a result, H_{1C} was partially supported.

H_{1d}: Workload will significantly and negatively influence an individual's trust in a multi-agent HAT when controlling for PTT, previous experience, and mission performance.

As reported in table 4.3, the variance explained by the workload predictor was significant (t= -2.78, p=0.006). As a result, H_{1D} was fully supported.

H_{1e}: Mission performance will significantly and positively influence individual's trust in a multi-agent HAT when controlling for PTT, previous experience, current LOA, and workload.

As reported in table 4.4, the variance explained by the trust predictor was significant (t=3.69, p<.001). As a result, H_{1E} was fully supported.

H_{2a}: The current LOA will significantly influence individual's trust in a multi-agent HAT when controlling for workload and trust.

As reported in table 4.4, the variance explained by the current LOA predictor was significant for the Consent LOA (t=3.05, p=0.003) and the Veto LOA (t=2.81, p=0.006) when compared to the manual LOA. However, the variance explained by the current LOA predictor was not significant for the advice LOA (t=1.59, p=0.114) when compared to the manual LOA. As a result, H_{2A} was partially supported.

H_{2b}: Workload will significantly and negatively influence individual's trust in a multi-agent HAT when controlling for current LOA and trust.

As reported in table 4.4, the variance explained by the workload predictor was not significant (t= -0.27, p=0.79). As a result, H_{2B} was not supported.

H_{2c}: Trust will significantly influence individual's mission performance in a multi-agent HAT when controlling for current LOA and workload.

As reported in table 4.5, the variance explained by the mission performance predictor was significant (t=3.04, p=0.003). As a result, H_{2C} was fully supported.

Chapter 5 Conclusion

Overview

This study aimed to evaluate the influences of numerous predictors on trust and mission performance in a multi-UAV HAT context. Specifically, the study examined the effect that predictors including PTT, previous experience, current LOA, workload, and mission performance had on trust. Second, the study examined the influence of predictors including current LOA, workload, and trust on mission performance. Results revealed that PTT, advice LOA, consent LOA, workload, and mission performance significantly influenced trust. The results also revealed that the consent LOA, veto LOA, and trust scores significantly influenced mission performance. This section will provide further discussion of these results, practical implications, limitations of the study, and future research.

Theoretical implications

This section will evaluate the results from both linear regression models and discuss the theoretical implications of these findings.

Propensity to Trust

Results from the current study revealed that PTT positively influenced an individual's trust. Specifically, the higher an individual's PTT score, the higher the trust scores. This finding is in line with current research in the dyadic HAT and human-human team literature, as PTT is a well-established predictor of individual trust (Alarcon et al., 2016; Zhang et al., 2021). For instance, in Alarcon et al. (2016), results revealed that scores from the Mayer & Davis' PTT scale was a significant predictor for higher levels of trust behavior in unfamiliar dyadic teams. More specifically, when participants were presented with the prisoner's dilemma, participants who reported high PTT scores were more likely to exhibit trusting behaviors with partners if they were not familiar with them. Similar findings were also found in Thomson et al. (2022), in which individuals with high PTT scores also reported higher forgiveness scores on a trust scale after an agent committed a trust violation. In this study, higher trust scores were also reported on a single-item scale, demonstrating a positive relationship between PTT and an individual's trust scores. Findings from the current study suggest that the positive relationship between PTT and trust scores found in the dyadic HAT literature are consistent with relationships observed in multi-HATs.

Previous Experience

Results from the current study revealed that previous experience did not impact trust, demonstrating that trust did not increase as individuals gained experience with agents. These findings are most likely attributed to the limited interactions participants had with the agents. In the current study, each trial was 5 minutes, in which little experience was gained from participants who finished the experiment with a total of 20 minutes of interactions with the agents. From this, the differences in experience levels between the manipulations may not have been sufficient for trust impacts to occur, suggesting more time and interactions with an agent is needed to allow previous experience to impact trust. However, there has been research that showed limited experience can impact trust. For example, in Merrit and Ilegen (2008), findings revealed that as a participant interacted with an autonomous system over a 20-minute trial, trust levels increased as usage increased. To address the inconsistencies in the current literature, future research should examine the effects of previous experience and establish a time or number of interactions needed to allow previous experience to impact trust development.

Current LOA

The advice, consent, and veto LOAs were all compared to the manual LOA, which was a reference variable in the current study. From this, only advice, consent and veto LOA will be discussed in the following section.

Current LOA and Trust

The advice and consent LOA significantly and negatively influenced an individual's trust scores. In other words, when participants were in the advice and consent LOA, lower trust scores were exhibited compared to scores in the manual LOA. On the other hand, the veto LOA did not significantly predict trust scores. Therefore, no relationship was identified between the veto LOA and trust scores.

Current literature suggests that lower LOAs such as the advice LOA typically yield high trust scores (Ruff et al., 2002; Nam et al., 2018), which are not aligned with the findings in this study. Results from the current study may be attributed to the participant's responsibility of confirming the target categorization after discerning the agents recommendations. For example, in the advice LOA, the agents would recommend target type (e.g. "I think this target is an enemy"), however, it was the responsibility of the participant to categorize the target then confirm the target. In the consent LOA, the agent already categorized the targets for the participants, however it was the responsibility of the participant to confirm the target. Due to the participants responsibilities in confirming the target type in the advice and consent LOA, it may have been easier for participants to double check the agent's action and identify missed targets or mislabels from the agents as opposed to the veto LOA, in which no input from the participant was necessary because the agents categorized and confirmed targets themselves. Therefore, catching mislabels and missing targets may not have been as prominent in the veto condition. This could have resulted in lower trust scores being exhibited in the advice and consent LOA.

The highest LOA, the veto LOA, had no significant impact on trust, which is inconsistent with findings from the literature. Findings from the literature suggest that higher LOAs exhibit lower levels of trust, primarily due to the participant's decrease in control over agents (Nam et al., 2018). However, results from the current study may be attributed to the participants supervisory role in the Veto LOA. More specifically, in the veto LOA, no input was required from the participants because the agents completed the categorization and confirmation tasks. Therefore, mislabels and missed targets may have been detected less from participants, resulting in higher reported trust.

Similar behaviors were reported in Walliser (2011) in which lower LOAs (management by consent) were compared to higher LOAs (management by exception) in an intelligence surveillance and reconnaissance (ISR) mission in which participants labelled enemies and friendlies based on targets found in a video feed. In the management by consent (MBC) condition, the automation was required to have explicit consent from the human operator to carry out its actions, whereas in the management by exception (MBE) condition, the automation was allowed to freely perform tasks unless overruled by the human operator. Results revealed that participants exhibited significant differences in correct identification across the automation levels. Participants experienced difficulties in detecting errors in the MBE condition as participants were out-of-the-loop with the automation's actions. Whereas participants found more false alarms with the MBC automation, as participants were in the loop with the automation's actions allowing them to catch any errors committed by the agent. Although Walliser (2011) did not measure trust, the study found different behavioral impacts of varying LOAs. Similar behaviors were also reported in Olson and Sarter (1998) in an aircraft simulation context in which pilots found it harder to detect automation errors in the MBE condition compared to the MBC condition. Behavioral findings from these studies may explain trust results from the current study as lower LOAs allowed participants to identify erroneous suggestions by the automation, whereas participants encountered more difficulty with identifying erroneous suggestions in the highest LOA.

In summary, the advice and consent condition provided greater ability for participants to catch agent errors as these conditions required participants to confirm the target type. The likelihood of catching agent errors being higher in the advice and consent condition, may have led to lower trust scores reported. However, in the veto condition, the agents completed all tasks for the participants, therefore there was a lower likelihood of participants identifying agent errors, which may have led to the lack of influence on trust.

Current LOA and Mission Performance

The consent and veto LOA demonstrated a positive relationship with mission performance scores, whereas no relationship was identified between the advice LOA and mission performance. This may be accredited to the higher levels of workload experienced in the lower LOAs, such as the advice LOA, resulting in degraded mission performance, which is supported by the literature (Zhang & Yang, 2017; McBride et al., 2021). Although the automation assisted in recommending the participant in target categorization, the identification task was primarily the responsibility of the human. From this, the same actions were carried out for the manual and advice condition, revealing little difference in the task requirements between the manual and advice LOA. As the manual LOA was a reference variable, the lack of differences in task requirements between the manual and advice LOA may also explain why the advice LOA did not influence mission performance.

On the other hand, agents in the consent LOA provided assistance of pre-marking the target, most likely providing additional workload reductions. Due to the additional assistance provided by the agents in the two higher LOAs, consent and veto, the participant experienced a decrease in workload as reported in Rebensky et al. (2022). More specifically, the previous study found the advice condition resulted in significantly higher workload scores when compared to the consent and veto conditions (Rebensky et al., 2022). As the participants exhibited lower workload scores, this may have consequently led to better performance scores. This rationale is in line with previous research in which higher LOAs result in lower levels of workload (Chen & Barnes, 2014; Barnes et al., 2015). This in turn affects mission performance, as lower workload scores lead to higher mission performance scores (Grim et al., 2018). From this, findings from the current study are in line with the literature.

According to the results of the current study, LOAs did not demonstrate a clear relationship with trust and mission performance. Findings from the literature demonstrate similar results, in which differing LOAs do not present a distinct pattern on the impacts of trust and mission performance (Stewart, 2006). This may be attributed to the effects of LOAs amongst differing task contexts. For example, as certain LOAs influence trust and mission performance for lower-level tasks like calculating basic arithmetic, similar benefits may not be advantageous for higher-level tasks such as identifying and targeting enemies under a time-critical conditions (Schneider et al., 2002). Thus, it is important to consider the results of this study when comparing results from findings in the current literature, as some inconsistencies may be attributed to the differing task context.

Workload and trust

The results from the current study revealed that workload significantly and negatively predicted trust scores, in which participants who reported higher workload scores also reported lower trust scores. In the literature, an inverse relationship is commonly found between workload and trust (van der Waa, 2021). When participants experience reduced workload, more time is allotted for participants to play a more active role in checking agents' actions, while allowing participants to make more accurate decisions (van der Waa, 2021). Similar findings were found in Hillsheim et al. (2017), in which participants who reported a high workload in a space trajectory task consequently reported low trust scores in their agents. Results from this study inferred that participants most likely attributed higher workload to the untrustworthiness of the agent, which may also explain why higher workload scores were associated with lower trust scores. Based on these findings, the relationship between workload and trust directly aligns with trends in the literature.

Workload and mission performance

The results from the current study did not identify a relationship between workload and mission performance, which does not align with findings from the literature. In the literature, common trends reveal that workload and mission performance share an inverse relationship, in which lower workload leads to higher mission performance scores (Zhang & Yang, 2017). This study may have failed to identify this relationship as a ceiling effect appears to be present with mission performance scores. For example, the mean score for mission performance across all participants was 86.41%, revealing that participants found the majority of targets in each map. Therefore, there was a lack of variation in scores resulting in a potential ceiling effect, as the experimental task may not have been challenging enough for most participants. This lack of variability may have led to the inability to find a relationship between workload and mission performance.

Trust and Mission Performance

The current study revealed that trust and mission performance significantly and positively predicted each other, as illustrated in both multiple regressions. For example, when higher trust scores were present, higher mission performance scores were also reported, which directly aligns with findings from the literature (Wang et al., 2017). Trust plays a crucial factor in HATs as it can heavily influence a human operator’s behavior, in which human operators who have low trust in their agent may reject useful advice presented by the agent, which can inadvertently affect mission performance (Kox et al., 2021). As a result, performance in HATs can suffer without trust, demonstrating a need for trust to be present to provide high mission performance scores and ensure mission success (Joe et al., 2014). Based on these findings, the relationship between trust and mission performance directly aligns with trends in the literature.

Practical Implications

Based on the study’s results, variables including PTT, current LOA, workload, and mission performance significantly influence trust. The current study also found LOA and trust scores affect mission performance scores. There are several practical implications of these findings. First, for workforce hiring considerations, it is important to consider PTT scores, as individuals with higher PTT scores may be more suitable to work HAT teammates and this could be a potential consideration during workforce hiring decisions. For instance, individuals with low PTT scores may present a risk of not being able to work with agents effectively, as they will have a harder time developing trust in these systems. As trust and mission performance scores are closely related, human operators that report low PTT scores may present higher risk of adverse effects to mission performance compared to operators with higher PTT scores.

Furthermore, when designing the LOA for a multi-agent mission, it is important to consider the impacts of each LOA identified in the study, and how each LOA may affect trust and mission performance. For example, it should be anticipated that higher LOAs with high levels of reliability lead to higher levels of performance as demonstrated in the study. Based on these results, incorporating higher LOAs should be considered as it may benefit overall mission performance amongst HAT operations. However, practitioners should be aware there are issues with operators trusting automated systems, as found in the current study. To address this, training targeted to improve trust within higher LOAs may help combat this issue to assist human operators in appropriately calibrating their trust levels to the agents' capabilities and limitations.

Additionally, workload parameters should be considered when designing a multi-agent mission, as too much operator workload may degrade trust. From this, designing automated systems to alleviate operator workload will be useful in assisting human operators to experience optimal workload levels, inherently leading to high trust levels.

Limitations and Delimitations

The reader should interpret these results with caution, given the limitations of the study. First, the participants included in the archival study were limited to Florida Institute of Technology students, demonstrating convenience sampling. From this, the results presented in this study may not accurately represent the general population.

The current study obtained a small sample size of 47 participants, potentially leading to a type II error i.e., the hypothesis was not supported despite a relationship being present. However, as each participant contained 4 observations, there were over 10 observations per independent variable with the multiple regression model, satisfying the recommended number of observations per independent variable (Tripepi et al., 2008). Therefore, the likelihood of a type II error is low in the current study.

Selecting an archival study as opposed to an experimental study was a delimitation, as there was no control over the variables that were included in the archival data set, as opposed to an experimental study in which full control would have been present. Furthermore, the previous study was not set up to conduct a multiple regression analyses, resulting in a mix of categorical and continuous variables in the current study's dataset. To mitigate this issue, a dummy coding approach was conducted to incorporate categorical variables into two multiple regressions.

Additionally, the design of short experimental trials suggests there was not sufficient time for previous experience to be identified as a significant predictor. As participants interacted with agents during four separate 5-minute trials, little differences in experience occurred over the study. Furthermore, each 5-minute trial consisted of a different LOA, which may have led to participants associating each LOA with different agents. From this, experiences from each trial may not have been additive, which may be why previous experience was not identified as a significant predictor for trust.

To capture overall team trust scores, the participant's trust score ratings for each individual drone was averaged. Therefore, taking the average of these individual trust ratings may not accurately reflect the participants' total team trust scores.

Furthermore, mission performance scores from the search and rescue task in the archival study presented scores skewed towards higher performance scores. From this, the experimental task was not a difficult task for most participants, as the majority of participants scored above 80% across all conditions. This may have contributed to the results demonstrated in the current study.

Future Research

When designing upcoming HAT studies, future research should provide participants with a considerable amount of time to interact with agents to observe the effects of experience as the participants gains more time with the agents. Along with this, designing a challenging experimental task to provide variability in mission performance scores should be accomplished to assess the effects of mission performance amongst manipulated variables. For trust, it is important for future research to distinguish and measure the differences between trust levels associated with each agent individually and the overall team. Capturing trust development amongst dyadic agent relationships, as well as the overall team trust, can help us understand trust dynamics at an individual and overall team level.

In the current study, there was no clear relationship found between all LOAs in respect to trust and mission performance. Similar trends are also found in the literature, where there are conflicting impacts of LOAs across trust and mission performance (Stewart, 2006). This is primarily due to the impact of LOAs amongst different task contexts. For instance, certain LOAs provide optimal trust and mission performance for lower-level tasks, whereas the same benefits may cause repercussions for higher-level tasks under more time-sensitive circumstances (Schneider et al., 2002). From this, extensive research should be conducted on how different LOAs affect trust in agents, based on differing task contexts. Additionally, based on the misalignment of findings from the current study in relation to trends in the literature, these discrepancies identify gaps in our understanding of team dynamics for upcoming multi-HAT missions. As there is limited research in the domain, it is critical for operators to identify these gaps to provide further explanations for fluctuations in trust or mission performance in multi-UAV missions across different task contexts.

As the current study aimed to identify different variables that affect trust and mission performance in a multi-HAT context, future research should investigate the impact of these variables when multiple human operators and agents are present. HAT studies are shifting to more complex tasks, in which numerous human operators must be present to collaborate with agents. From this, investigating the effects variables in heterogeneous multi-HATs will be critical for the design of future HAT missions.

Conclusion

The proliferation of multi-HAT operations has exposed areas needing improvement for human operator training and agent design in a multi-agent HAT context. The aim of the current study was to examine the factors that influence trust and mission performance as participants reported PTT, workload, and gained experience across varying LOAs was obtained through the analysis of archival data. In the previous study, participants completed a military intelligence, surveillance, and reconnaissance (ISR) mission whilst identifying targets (civilian, enemy, and friendly) throughout the route. Surveys collected to assessed PTT, workload, and operator trust, LOAs (manual, advice, consent, and veto). Findings from the study revealed that PTT and mission performance positively and significantly influence trust, whereas the advice LOA, consent LOA, and workload negatively and significantly influence trust. Furthermore, the consent LOA, veto LOA, and trust positively and significantly influenced mission performance. Findings from the study revealed significant relationships between variables such as PTT, LOAs, workload, mission performance, and trust, across multi-HAT teams. Implications drawn from the study includes HAT design considerations when developing HAT missions or recruiting human operators to take part in HAT operations. More specifically, it is important to consider variables such as PTT, LOAs and workload when accounting for the effects of mission performance and trust in a multi-HAT context.

References

Alarcon, G. (2018). *The role of propensity to trust and the five factor model across the trust process*.

Alarcon, G. M., Lyons, J. B., & Christensen, J. C. (2016). The effect of propensity to trust and familiarity on perceptions of trustworthiness over time. *Personality and Individual Differences*, *94*, 309–315. https://doi.org/10.1016/j.paid.2016.01.031

Azhar, M., & Sklar, E. (2017). *A study measuring the impact of shared decision making in a human-robot team*. https://journals.sagepub.com/doi/full/10.1177/0278364917710540?casa_token=ynYgRB4D6isAAAAA%3Ai9uYLamA83yhg41wAzvRXoyM-jIeUrzTl05i8jb7h-OMlz80rl7NFTFKi7ea1WeiCBbqGxpgFRiw

Bobko, P., Hirshfield, L., Eloy, L., Spencer, C., Doherty, E., Driscoll, J., & Obolsky, H. (2022). Human-agent teaming and trust calibration: A theoretical framework, configurable testbed, empirical illustration, and implications for the development of adaptive systems. *Theoretical Issues in Ergonomics Science*, *0*(0), 1–25. https://doi.org/10.1080/1463922X.2022.2086644

Chen, J., & Barnes, M. (2014). *Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues | IEEE Journals & Magazine | IEEE Xplore*. https://ieeexplore.ieee.org/abstract/document/6697830?casa_token=eGeL1LlqrjwAAAAA:obt4m-eq2DyzDGsBdbM-0ootEPktecN5stuKPsZNEc_hx18OOy357RWHLImbTngABL6wgFpFCg

Chen, J. Y., & Barnes, M. J. (2013). *Human-Agent Teaming for Multi-Robot Control: A Literature Review*. ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD. https://apps.dtic.mil/sti/citations/ADA583900

Chen, J. Y. C. (2010). *UAV-guided navigation for ground robot tele-operation in a military reconnaissance environment*. https://www.tandfonline.com/doi/full/10.1080/00140139.2010.500404?casa_token=veLy_2_q7LUAAAAA%3AC13lCynSv0bnXfkTlAa9Odcm591b8v_FqepCYq88ZQ7czW0dLtnuntzDjusq8gEPX1sn8gRfFg2_

Chen, J. Y. C., & Barnes, M. J. (2014). Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues. *IEEE Transactions on Human-Machine Systems*, *44*(1), 13–29. https://doi.org/10.1109/THMS.2013.2293535

Chiou, E. K., & Lee, J. D. (2016). *Cooperation in Human-Agent Systems to Support Resilience: A Microworld Experiment—Erin K. Chiou, John D. Lee, 2016*. https://journals.sagepub.com/doi/full/10.1177/0018720816649094

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). *Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. - PsycNET*. https://psycnet.apa.org/doiLanding?doi=10.1037%2F0021-9010.92.4.909

Cooke, N., Demir, M., & McNeese, N. (2016). *Synthetic Teammates as Team Players: Coordination of Human and Synthetic Teammates*. https://apps.dtic.mil/sti/citations/AD1017169

Cummings, M. L. (2015). Automation Bias in Intelligent Time Critical Decision Support Systems. In *Decision Making in Aviation*. Routledge.

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics*, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

Dikmen, M., & Burns, C. (2017). Trust in autonomous vehicles: The case of Tesla Autopilot and Summon. *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 1093–1098. https://doi.org/10.1109/SMC.2017.8122757

Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert, L. P. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies*, *104*, 428–442. https://doi.org/10.1016/j.trc.2019.05.025

Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., & Endsley, M. R. (2008). The influence of agent reliability on trust in human-agent collaboration. *Proceedings of the 15th European Conference on Cognitive Ergonomics: The Ergonomics of Cool Interaction*, 1–8. https://doi.org/10.1145/1473018.1473028

Ferraro, J. C., Mouloua, M., Mangos, P. M., & Matthews, G. (2022). Gaming experience predicts UAS operator performance and workload in simulated search and rescue missions. *Ergonomics*, *0*(0), 1–13. https://doi.org/10.1080/00140139.2022.2048896

Gill, H., Boies, K., Finegan, J., & McNally, J. (2005). *Antecedents Of Trust: Establishing A Boundary Condition For The Relation Between Propensity To Trust And Intention To Trust | SpringerLink*. https://link.springer.com/article/10.1007/s10869-004-2229-8

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Greco, V., & Roger, D. (2001). Coping with uncertainty: The construction and validation of a new measure. *Personality and Individual Differences*, *31*(4), 519–534. https://doi.org/10.1016/S0191-8869(00)00156-2

Grimm, D., Demir, M., Gorman, J. C., & Cooke, N. J. (2018). The Complex Dynamics of Team Situation Awareness in Human-Autonomy Teaming. *2018 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, 103–109. https://doi.org/10.1109/COGSIMA.2018.8423990

Gugerty, L., & Johnell, B. (2004). *Reference-Frame Misalignment and Cardinal Direction Judgments: Group Differences and Strategies. - PsycNET*. https://psycnet.apa.org/record/2004-95231-001

Hafizoğlu, F. M., & Sen, S. (2019). Understanding the Influences of Past Experience on Trust in Human-agent Teamwork. *ACM Transactions on Internet Technology*, *19*(4), 1–22. https://doi.org/10.1145/3324300

Hancock, P. A., Billings, D. R., Oleson, K. E., Chen, J. Y., De Visser, E., & Parasuraman, R. (2011). *A Meta-Analysis of Factors Influencing the Development of Human-Robot Trust*. ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD HUMAN RESEARCH AND ENGINEERING DIRECTORATE. https://apps.dtic.mil/sti/citations/ADA556734

Hart, S. G. (2006). *Nasa-Task Load Index (NASA-TLX); 20 Years Later—Sandra G. Hart, 2006*. https://journals.sagepub.com/doi/abs/10.1177/154193120605000909?casa_token=73CPwrsfeRkAAAAA:YgKwrcKO6S_r-lJt1HOqgGTPnaFKPzFSVTqOLadaCVvwOff3vwq5xaMJdaKCE60jw9w8m8NUSOuN

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. A. Hancock, & N. Meshkati (Eds.), *Advances in Psychology, 52: Human Mental Workload* (pp. 139-183). Oxford, England: North-Holland.

Hegarty, M., Richardson, A. E., Montello, D. R., Lovelace, K., & Subbiah, I. (2002). Development of a self-report measure of environmental spatial ability. *Intelligence, 30*(5), 425–447. https://doi.org/10.1016/S0160-2896(02)00116-2

Hillesheim, A., Rusnock, C., Bindewald, J., & Miller, M. (2017). *Relationships between User Demographics and User Trust in an Autonomous Agent*. https://journals.sagepub.com/doi/abs/10.1177/1541931213601560?casa_token=Y5umBVcdLkQAAAAA:UB5MiRywfnb_gaoF16Hmxcq1U14CQHKZXJcpCu1MLTiu8TsDgVSl5H1V2atHSS14i6GXAIWAAf5D

Hou, M. (2020). IMPACT: A Trust Model for Human-Agent Teaming. *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, 1–4. https://doi.org/10.1109/ICHMS49158.2020.9209519

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The Measurement of the Propensity to Trust Automation. In J. Y. C. Chen & G. Fragomeni (Eds.), *Virtual, Augmented and Mixed Reality. Applications and Case Studies* (pp. 476–489). Springer International Publishing. https://doi.org/10.1007/978-3-030-21565-1_32

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: Vol 4, No 1*. https://www.tandfonline.com/doi/abs/10.1207/S15327566IJCE0401_04

Joe, J. C., O'Hara, J., Medema, H. D., & Oxstrand, J. H. (2014). *Identifying requirements for effective human-automation teamwork*. Idaho National Lab.(INL), Idaho Falls, ID (United States).

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Hoffman, R. R., Jonker, C., Riemsdijk, B. van, & Sierhuis, M. (2011). Beyond Cooperative Robotics: The Central Role of Interdependence in Coactive Design. *IEEE Intelligent Systems*, *26*(3), 81–88. https://doi.org/10.1109/MIS.2011.47

Jung, M. F., DiFranzo, D., Stoll, B., Shen, S., Lawrence, A., & Claure, H. (2018). *Robot Assisted Tower Construction—A Resource Distribution Task to Study Human-Robot Collaboration and Interaction with Groups of People* (arXiv:1812.09548). arXiv. https://doi.org/10.48550/arXiv.1812.09548

Khasawneh, A., Rogers, H., Bertrand, J., Madathil, K. C., & Gramopadhye, A. (2019). Human adaptation to latency in teleoperated multi-robot human-agent search and rescue teams. *Automation in Construction*, *99*, 265–277. https://doi.org/10.1016/j.autcon.2018.12.012

Kohn, S. C., Kluck, M., & Shaw, T. (2020). *A Brief Review of Frequently Used Self-Report Measures of Trust in Automation*. https://journals.sagepub.com/doi/abs/10.1177/1071181320641342?casa_token=7QtDxDnCkzkAAAAA:2RbGy9YtZuypw1dxI2b450Bj7s7jkyw7y_j0EY-c3ygDiQWoNEKhTwoI44hP6V6qJc3eFHGN3rbG

Kox, E., Kerstholt, J., Hueting, T., Barnhoorn, J., & Eikelboom, A. (2021). *AUTONOMOUS SYSTEMS AS INTELLIGENT TEAMMATES: SOCIAL PSYCHOLOGICAL IMPLICATIONS*. https://static1.squarespace.com/static/53bad224e4b013a11d687e40/t/5dc416da49d3b81baca3244b/1573131995316/24th_ICCRTS_paper_74.pdf

Lee, J. D., & See, K. A. (2004). *Trust in Automation: Designing for Appropriate Reliance*. https://journals.sagepub.com/doi/abs/10.1518/hfes.46.1.50_30392

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

Levinthal, B. R., & Wickens, C. D. (2006). *Management of Multiple Uavs with Imperfect Automation*. https://journals.sagepub.com/doi/abs/10.1177/154193120605001748?journalCode=proe

Longo, L. (2018). On the reliability, validity, and sensitivity of three mental workload assessment techniques for the evaluation of instructional designs: A case study in a third-level course. *Proceedings of the 10th International Conference on Computer Supported Education CSEDU, 166-178.*

Master, R., Gramopadhye, A. K., Melloy, B., Bingham, J., & Jiang, X. (2000). A questionnaire for measuring trust in hybrid inspection systems. *Proceedings of The Industrial Engineering Research Conference, Dallas, TX*.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

McBride, Sara. E., Rogers, W. A., & Fisk, A. D. (2011). *Understanding the Effect of Workload on Automation Use for Younger and Older Adults*.

McShane, S. L. (2014). *Propensity to Trust Scale*.

Mercado, J. E., Chen, J. Y. C., Rupp, M. A., Barnes, M. J., Barber, D., & Procci, K. (2016). *Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management*. https://journals.sagepub.com/doi/full/10.1177/0018720815621206

Merritt, S. M., & Ilgen, D. R. (2008). Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors*, *50*(2), 194–210. https://doi.org/10.1518/001872008X288574

Nam, C., Li, H., Li, S., Lewis, M., & Sycara, K. (2018). Trust of Humans in Supervisory Control of Swarm Robots with Varied Levels of Autonomy. *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 825–830. https://doi.org/10.1109/SMC.2018.00148

Nguyen, D. M. (2020). *1, 2, or 3 in a HAT? How a Human-Agent Team's Composition Affects Trust and Cooperation* [Thesis]. https://repository.lib.fit.edu/handle/11141/3192

Olson, W. A., & Sarter, N. B. (1998). "As long as I'm in control...": Pilot preferences for and experiences with different approaches to automation management. *Proceedings Fourth Annual Symposium on Human Interaction with Complex Systems*, 63–72. https://doi.org/10.1109/HUICS.1998.659955

OSD. (2017). *Unmannned systems integrated roadmap*.

Otto, R. P. (2016). *Small Unmanned Aircraft Systems (SUAS) Flight Plan: 2016-2036. Bridging the Gap Between Tactical and Strategic*. AIR FORCE DEPUTY CHIEF OF STAFF WASHINGTON DC WASHINGTON DC United States. https://apps.dtic.mil/sti/citations/AD1013675

Parasuraman, R., Molloy, R., & Singh, I. (1993). Performance Consequences of Automation Induced Complacency. *International Journal of Aviation Psychology*, *3*. https://doi.org/10.1207/s15327108ijap0301_1

Pynadath, D. V., Wang, N., & Kamireddy, S. (2019). A Markovian Method for Predicting Trust Behavior in Human-Agent Interaction. *Proceedings of the 7th International Conference on Human-Agent Interaction*, 171–178.

Ramchurn, S. D., Fischer, J. E., Ikuno, Y., Wu, F., Flann, J., & Waldock, A. (2015, June 23). A Study of Human-Agent Collaboration for Multi-UAV Task Allocation in Dynamic Environments. *Twenty-Fourth International Joint Conference on Artificial Intelligence*. Twenty-Fourth International Joint Conference on Artificial Intelligence. https://www.aaai.org/ocs/index.php/IJCAI/IJCAI15/paper/view/11293

Rebensky, S., Carmody, K., Ficke, C., Carroll, M., & Bennett, W. (2022). Teammates instead of tools: The impacts of level of autonomy on mission performance and human-agent teaming dynamics in multi-agent distributed teams. *Frontiers in Robotics and AI*, 102.

Rebensky, S., Carroll, M., Bennett, W., & Hu, X. (2021). *Full article: Impact of Heads-Up Displays on Small Unmanned Aircraft System Operator Situation Awareness and Performance: A Simulated Study*.

Ruff, H. A., Narayanan, S., & Draper, M. H. (2002). *Human Interaction with Levels of Automation and Decision-Aid Fidelity in the Supervisory Control of Multiple Simulated Unmanned Air Vehicles | MIT Press Journals & Magazine | IEEE Xplore*. https://ieeexplore-ieee-org.portal.lib.fit.edu/abstract/document/6790468

Schaefer, K. (2013). The Perception And Measurement Of Human-robot Trust. *Electronic Theses and Dissertations*. https://stars.library.ucf.edu/etd/2688

Schaefer, K. E., Straub, E. R., Chen, J. Y. C., Putney, J., & Evans, A. W. (2017). Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cognitive Systems Research*, *46*, 26–39. https://doi.org/10.1016/j.cogsys.2017.02.002

Schelble, B. G., Flathmann, C., & McNeese, N. (2020). Towards Meaningfully Integrating Human-Autonomy Teaming in Applied Settings. *Proceedings of the 8th International Conference on Human-Agent Interaction*, 149–156. https://doi.org/10.1145/3406499.3415077

Schneider, M., McGrogan, J., Colombi, J. M., Miller, M. E. M., & Long, D. S. (2014). *7.1.1 Modeling Pilot Workload for Multi-Aircraft Control of an Unmanned Aircraft System—Schneider—2011—INCOSE International Symposium—Wiley Online Library*. h

Schneider, T. R., Stokes, S. A., Rivers, S., Lohani, M., & McCoy, M. (2017). The influence of trust propensity on behavioral trust. *In Poster Session Presented at the Meeting of Association for Psychological Society, Boston.*

Spravka, J., Moisio, D., & Payton, M. G. (2003). *Unmanned Air Vehicles: A New Age in Human Factors Evaluations*. https://apps.dtic.mil/sti/citations/ADA432105

Stewart, G. L. (2006). A Meta-Analytic Review of Relationships Between Team Design Features and Team Performance. *Journal of Management*, *32*(1), 29–55. https://doi.org/10.1177/0149206305277792

Syrdal, D. S., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2009). *The Negative Attitudes Towards Robots Scale and reactions to robot behaviour in a live Human-Robot Interaction study*. http://uhra.herts.ac.uk/handle/2299/9641

Thomsen, A. M. (2022, January 28). *NoRobot's perfect: Trust repair in the face of agent error. How do individual factors influence trust development in human-agent teams?* [Info:eu-repo/semantics/bachelorThesis]. http://essay.utwente.nl/89379/

Tripepi, G., Jager, K. J., Dekker, F. W., & Zoccali, C. (2008). Linear and logistic regression analysis. *Kidney International*, *73*(7), 806–810. https://doi.org/10.1038/sj.ki.5002787

U.S. Department of Transportation. (2014). *Unmanned aircraft system (UAS) service demand*.

Valero-Gomez, A., Puente, P. de la, & Hernando, M. (2011). *Impact of Two Adjustable-Autonomy Models on the Scalability of Single-Human/Multiple-Robot Teams for Exploration Missions*. https://journals.sagepub.com/doi/full/10.1177/0018720811420427?casa_token=_0eygvZ7S64AAAAA%3AuupeXDaX6JlJ5MkCi9CkiQS4fwlUXYoOA3P888KPPm15-CwIzyAV5DBJLF718sBqANrCVnqAbPow

van der Waa, J., Verdult, S., van den Bosch, K., van Diggelen, J., Haije, T., van der Stigchel, B., & Cocu, I. (2021). Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations. *Frontiers in Robotics and AI*, *8*. https://www.frontiersin.org/articles/10.3389/frobt.2021.640647

Walliser, J. C. (2011). Trust in Automated Systems: The Effect of Automation Level on Trust Calibration. *Naval Postgraduate School Monterey CA*. https://apps.dtic.mil/sti/citations/ADA547808

Walliser, J. C., de Visser, E. J., Wiese, E., & Shaw, T. H. (2019). Team Structure and Team Building Improve Human–Machine Teaming With Autonomous Agents. *Journal of Cognitive Engineering and Decision Making*, *13*(4), 258–278. https://doi.org/10.1177/1555343419867563

Wang, N., Pynadath, D. V., Hill, S. G., & Merchant, C. (2017). *The Dynamics of Human-Agent Trust with POMDP-Generated Explanations In International Conference on Intelligent Virtual Agents (pp. 459-462). SpringerLink.* https://link.springer.com/chapter/10.1007/978-3-319-67401-8_58

Zhang, M., & Yang, J. (2017). Evaluating effects of workload on trust in automation, attention allocation and dual-task performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1799–1803. https://doi.org/10.1177/1541931213601932

Zhang, X. (2021). *"Sorry, it was my fault": Repairing Trust in Human-Robot Interactions*. University of Oklahoma.