


Mediating Agent Reliability with Human Trust, Situation Awareness, and Performance in Autonomously-Collaborative Human-Agent Teams

Sebastian S. Rodriguez , University of Illinois at Urbana-Champaign, USA, Erin Zaroukian, DEVCOM Army Research Laboratory, USA, Jefferson Hoye, VolunteerScience, USA, Derrik E. Asher, DEVCOM Army Research Laboratory, USA

When teaming with humans, the reliability of intelligent agents may sporadically change due to failure or environmental constraints. Alternatively, an agent may be more reliable than a human because their performance is less likely to degrade (e.g., due to fatigue). Research often investigates human-agent interactions under little to no time constraints, such as discrete decision-making tasks where the automation is relegated to the role of an assistant. This paper conducts a quantitative investigation towards varying reliability in human-agent teams in a time-pressured continuous pursuit task, and it interconnects individual differences, perceptual factors, and task performance through structural equation modeling. Results indicate that reducing reliability may generate a more effective agent imperceptibly different from a fully reliable agent, while contributing to overall team performance. The mediation analysis shows replication of factors studied in the trust and situation awareness literature while providing new insights: agents with an active stake in the task (i.e., success is dependent on team performance) offset loss of situation awareness, differing from the usual notion of overtrust. We conclude with generalizing implications from an abstract pursuit task, and we highlight challenges when conducting research in time-pressured continuous domains.

Keywords: human-agent teaming, human-autonomy teaming, trust, situation awareness, individual differences, predator-prey pursuit, group dynamics

Address correspondence to Sebastian S. Rodriguez, University of Illinois at Urbana-Champaign, Department of Computer Science, Urbana, Illinois, USA.
Email: srodri44@illinois.edu.

Journal of Cognitive Engineering and Decision Making
Vol. 0, No. 0, ■■ ■, pp. 1-23
DOI:10.1177/15553434221129166
Article reuse guidelines: sagepub.com/journals-rmissions
© 2022, Human Factors and Ergonomics Society.

Introduction

Teams are often defined as organizations that employ dynamic and adaptive behavior between individuals in order to achieve a common goal (Arrow et al., 2000). For many scenarios, good team performance can be the difference between achieving or failing the team's goal. With advances in computational technologies, the use of artificial intelligence (AI) and machine learning has allowed us to spring automated agents forth to autonomy: non-living entities which have the capability to be intelligent and make their own decisions. Because of this, the fundamental structure of teaming has changed—teams can now be comprised of a combination of human members and artificial agents. With automation and autonomy becoming increasingly ubiquitous in the 21st century, human-agent teams (HATs) already exist in a wide variety of domains (in research and practice), such as embodied agents for military operations (Kott, 2018), partially-autonomous driving (Lawson-Guidigbe et al., 2020), content recommendations (Schaffer et al., 2018), and algorithmic decision-making systems prominent in data analytics (Arnold et al., 2004, 2006), to mention a few. This results in a growing need to study the cooperative dynamic in HATs to allow effective and intuitive interaction for the growing usage of automation towards autonomy.

Thanks to advanced mathematical techniques in machine learning, complex and unpredictable operating environments are effectively actionable by agents. The capabilities of AI systems can often match or surpass human performance in specific tasks that utilize an agent's inherent advantages—such as speedy processing of the

operating environment's data for decision making. In other cases, the human's ability to visually perceive and adapt to the environment with high levels of judgment serve to complement the agent's processing speed (Korteling et al., 2021). Upon integration into HATs, general intuition established that the human and AI would collaborate effectively, resulting in increased performance across the board. Far from it, however, is that upon interaction with a capable AI, humans often end up overtrusting the system resulting in penalties to both individual and team performance (Bansal, Nushi, Kamar, Lasecki, et al., 2019; Rodriguez et al., 2019), a callback to similar issues in human factors research and replicated multiple times over the course of investigation (Parasuraman, 1997; Parasuraman et al., 1993; Parasuraman & Manzey, 2010). However, in these domains, the AI system is often treated as an assistant, providing decision aid to operators in order to complete a task, with the human retaining the role of the primary decision maker. However, if AI research is geared towards eventually reaching autonomy, these systems must be able to act independently to complete their task with limited to no intervention (Endsley, 2017). Human-agent teaming research often addresses situations where agents hold equal or higher responsibility than human operators, but few focus on the effect of varying agent reliability (e.g., Wright et al. (2020); McNeese et al. (2021)). The level of performance brought by these agents might significantly differ from or exceed human performance, resulting in potential overtrusting behavior, in part because the agent's performance is less likely to degrade over time (Langner & Eickhoff, 2013).

Work investigating different facets in human-AI interaction and HATs often makes use of experiments that operate in discrete domains: at a specific point in the task, the experiment prompts the user to make a choice to affect the outcome, often with no time pressure—which may significantly affect performance (Entin & Serfaty, 1999). However, the reality is that our world is continuous, and many advances in AI take advantage of continuity in order to provide more precise operation. For instance, drone technology for aerial domains (e.g., first

responders, military, and aviation) often acts beyond discrete decision-making. Determining whether they are operating correctly or not and what decision to make at the moment depends highly on a variety of different factors (often taking the form of continuous random variables, such as position, height, velocity, and goal), rather than a dichotomous summary of its actions. The agent's reliability can often shift from intended behavior to erratic movements, relying on internal system components to account for and correct the ill behavior. In these cases, we must ensure that established cognitive models of trust generalize towards operation in continuous spaces. As far as we are aware, investigating human collaboration and decision-making with fully autonomous agents is often not researched, thus we present the novelty of this work and our main contributions.

This paper presents an experiment that demonstrates the interplay of performance and perceptions of AI systems by varying agent reliability in a HAT operating in a continuous environment. Agents are assigned to complete a continuous pursuit task (dubbed the Predator-Prey game – PPG) by actively collaborating with a human. We measure performance and subjective perceptions of the agents to inform the discussion of trust and performance models in continuous tasks. This experiment explores a between-subjects design where we note changes in performance through perceptual factors and human predispositions. We aim to answer the following research questions:

1. How is human performance affected by varying reliability in collaborating agents in a continuous pursuit task?
2. How do human individual differences mediate perception of agents, situation awareness, and performance in continuous domains?

Background

Agents and Autonomy: Human-Agent teaming. Research often confounds the terms “agent”, “automation”, and “autonomy” to describe intelligent software that can process input and provide an output towards a goal, often without human intervention. Early automation allowed machinery to complete tasks that are

hazardous, unpleasant, or unattainable for humans (Lee et al., 2017). The development of AI and machine learning algorithms has resulted in adaptive and capable systems, with their deployment in practice forthcoming as we ensure their effectiveness within their operating environments. Their new capabilities have resulted in integration within operational teams with humans, culminating in the human-agent teaming paradigm. McNeese et al. defines HATs as a team comprised of at least one human working cooperatively with at least one autonomous agent (McNeese et al., 2018). The research additionally notes several overlaps with similar paradigms used in human-computer interaction, human factors, and robotics research, such as human-automation interaction, human-robot teams, human-AI interaction, human-AI collaboration, and human-autonomy teaming. For the purposes for this work, we use the term “agent” to refer to an AI-driven system that acts independently from human intervention, and we center the implications around human-agent teams, albeit many of the concepts discussed in this work will find applicability for paradigms with nomenclative similarities.

The agents that we intend to investigate can be defined according to the Levels of Automation (LOA) continuum established by Sheridan and Verplank (Sheridan & Verplank, 1978). The PPG implements agents at level 10, granting them complete independence from human intervention. According to the definition of autonomy, these agents are fully autonomous and effectively operational only in the continuous pursuit domain. Thus, this shifts away from supervisory control and moves towards equal collaboration, which some prior work has addressed (e.g., Hinds et al. (2004); Azhar and Sklar (2017)).

Game Theory and Continuous Pursuit

Game theoretic-scenarios are often employed to research the performance of algorithms that drive AI behavior. Heuristic algorithms and reinforcement learning policies are often benchmarked in strategy-based games (Bard et al., 2019; Campbell et al., 2002; Gibney, 2016; Silver et al., 2018; Vinyals et al., 2017),

with modern human-AI interaction and collaboration paradigms using game-based tasks as well (Ashktorab et al., 2021; Gero et al., 2020; Havrylov & Titov, 2017; Lazaridou et al., 2016; Liang et al., 2019). However, a large amount of work focuses on scenarios where participants encounter a discrete choice to make (but not to be confused with a discrete choice experiment). For instance, a turn-based game with or against an AI may allow infinite time to respond or interact with the agent, allowing user impressions not only to be formed given characteristics of the agent, but also through time variation. Interactions such as these can be generalized with real-life tasks associated with low-risk or with no time pressure, such as movie or product recommendations. However, previous research has yet to address cases with time critical constraints beyond supervisory control. Scenarios like these are of high interest for domains that require sophisticated embodied agents that cooperate to achieve a goal (e.g., robotics, elder care, drone control), often with little to no time to allow judgment or intervention from the human. Often, the technical advancement of autonomy supersedes human-centered uses and concerns when deploying these systems, and many have called to focus on human-centered issues when developing independent systems (Endsley, 2017; Hancock, 2017).

The proposed task in the PPG presents a specific case of interaction within an HAT: the agents are fully autonomous, have no line of communication with the human, and operate in a fully continuous space. Example real-world scenarios that this task would effectively model are interaction with fully-autonomous unmanned aerial vehicles (UAVs) (McNeese et al., 2018), human-swarm interaction (Kolling et al., 2016), or team-dynamics with drones as part of the Internet of Battlefield Things (IoBT) (DeCostanza et al., 2018); these are situations where supervisory control is either not desired nor feasible to attain. There has been longstanding interest in using continuous pursuit models to investigate emergent collaborative behavior between humans and agents (Asher et al., 2018; Barton & Asher, 2018). Although prior work has focused on establishing the

parameter space (Chung et al., 2011) and algorithmically identifying cooperative behavior from agents (Asher et al., 2019; Zaroukian et al., 2019), the question of the viability of these agents in operation with respect to models in human-AI interaction remains unaddressed. Ultimately, our work aims to shed some light on long-established models in human-AI interaction and HAT in a novel context, which will allow engineers and designers incorporating AI-based systems in these domains to be aware of their benefits and shortcomings.

Methodology

Overview

In this section, we describe the implementation of our experimental task. We had participants play the Predator-Prey game, where they are tasked with teaming up with two automated agents (predators) to “capture” (i.e., collide with) a third automated agent (prey), which is constantly evading in a continuous space. This serves to help our understanding of cooperative real-world pursuit tasks and strategy formation, as opposed to a simplified, discrete space. A screenshot of the PPG is shown in Figure 1.

The following description of the task environment maps 1 SI unit to 1 virtual unit (e.g., 1 m = 1 unit of distance in the environment). The task environment is comprised of a closed square arena of 2 m of width and 2 m of height. The players in the PPG move their circular avatar on a physics-based system by applying a force to their agent. In order to give predators and prey an equal chance to succeed, as well as to encourage the emergence of coordination among predators, the predators were made slower than the prey. The predators had a maximum speed of 1 m/s and accelerated at a maximum rate of 3 m/s². The prey had a maximum speed of 1.3 m/s and accelerated at a maximum rate of 4 m/s². The mass of all players was set at 1 kg. The diameters of the players were 0.15 m and 0.1 m for predators and prey, respectively. Upon capture, the capturing predator and prey would knock each other back at an impulse force of 1 m/kg*s until losing all momentum (by reducing the absolute velocity at

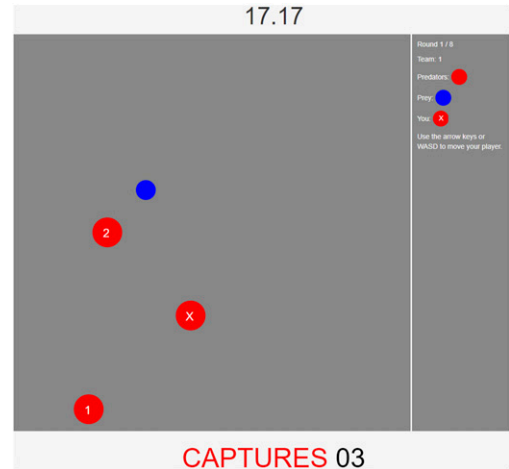


Figure 1. Screenshot of the Predator-Prey Game. Predators (red dots) are tasked to collectively work to capture the Prey (blue dot). The participant always assumes the role of the predator marked with an “X”. Predator teammates are labeled with either “1” or “2”. The round time remaining and number of team captures can be seen at the top and bottom, respectively. Instructional information is displayed in the sidebar.

the rate of 0.25 m/s²). Additionally, the prey is granted 0.5 seconds of invincibility after being captured, such that subsequent captures are not counted if multiple predators capture the prey at the same time. This task was designed according to a reinforcement learning testbed presented by OpenAI (Lowe et al., 2017), which has been of recent interest for investigating human-agent teams in embodied collaborative contexts (Barton & Asher, 2018).

The behavior of the predator and prey agents is outlined in the following section. The primary independent variable of interest is agent reliability. We varied the reliability of the agent teammates across conditions (i.e., in a given condition, both agent teammates had the same level of reliability). Dependent variables of interest include measuring performance, behavioral predispositions, attitudes towards the agents, and situation awareness. The proposed analysis entails modeling with inferential statistics, latent growth curve modeling, and structural equation modeling to answer our research questions.

Modeling Agent Behavior

A heuristic ruleset powered the agents' decision-making process in the PPG. Their behavior was governed by a procedure involving multiple geometric calculations per second to determine a target position toward which to move. For our experiment, agents calculated a new target position once every 0.25 seconds (4 Hz)—we refer each calculation as a step. Each predator was assigned a distinct strategy (either chaser or interceptor behavior), and the prey was assigned an evasive strategy. The predator chaser strategy attempts to close the distance to the prey at every step; that is, the target position is always the position of the prey. This is a rudimentary form of pursuit with an easy-to-establish a mental model for its behavior. The agent calculates a difference vector from the target position to their current position, and it moves in the direction of the vector. The predator interceptor strategy accounts for the prey's position and velocity in order to intercept it at some point in the near future, as it is found to be an optimal strategy for pursuers in differential games (Lin et al., 2011). The prey strategy calculates a set of candidate points on the edge of the play area to move towards based on the positions of the predators. The candidate points are generated by creating a triangle using the predator agents as vertices, followed by drawing a line that crosses the midpoint of each side of the triangle, that intersects with the bounds of the arena—this results in the farthest midpoints between every pair of predators, as the prey will attempt to maximize distance between itself and all predators. This results in six candidate points. The point that is selected is the one that is farthest from the predators based on the squared sum of the Euclidean distances from the predators.

In perfect scenarios, agents have the potential to respond perfectly to the environment, but it is often far from reality. For instance, robotics research establishes that frequent errors often plague robots, even after much investment and research to make them reliable (Honig & Oron-Gilad, 2018). Likewise for computational systems, limitations in information or processing power can lead to unexplained and sudden errors

(Schaffer et al., 2020). Because errors in an AI system can manifest in different magnitudes and frequencies (Honig & Oron-Gilad, 2018), we investigate a particular occurrence of error that hampers (but does not eliminate) the AI system's ability to complete their objectives.

For a continuous pursuit-evasion game, the optimal solutions are often formulated to require a pursuer to reach a certain target point in the action space under a certain time constraint (e.g., before a target escapes), after which the conditions have changed and require recalculation of said target point (Lin et al., 2011; Weintraub et al., 2020). By introducing error that completely hampers the functionality of the AI system (e.g., by making it non-functional or non-responsive), we trivialize the proposed research questions, as prior work has widely addressed the relationship between faulty systems and subjective attitudes (Parasuraman, 1997; Parasuraman et al., 1993; Wickens & Dixon, 2007). Instead, we opted to add error that introduces erratic behavior without preventing the agent from reaching their calculated target points. Such behavior is reminiscent of “juking” movement seen in gridiron football, albeit used from an offensive perspective.

We control this behavior by using a Markov Decision Process (MDP) to have agents behave according to two states: Clean and Noisy. Our independent variable, reliability, is the set of transition probabilities that govern which state the agent remains in most of the time. In the Clean state, the agent moves towards the calculated target position without any perturbation following its appointed strategy. In the Noisy state, the agent's target position is perturbed by adding a random vector with a magnitude of 0.8. The random vector is recalculated every time a new target position is generated, such that when the agent is in the Noisy state, it is constantly thrown off-course yet still following the trajectory of the target position. The magnitude of 0.8 was selected to contain the perturbed target position within the arena. The transition probabilities were selected by using Monte Carlo simulations to calculate the amount of time an agent would remain in a given state.

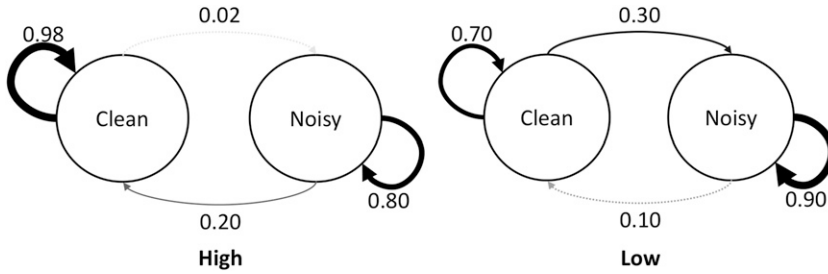


Figure 2. MDP transition probabilities for the High condition (left), and Low condition (right). The thickness and darkness of the arrow represents the probability of transition.

Measures

The independent variable of interest for these experiments is the reliability of the predator teammates. We define the agent's reliability as its potential capability to retrieve information from the environment and perform as designed. For instance, an agent with 80% reliability is akin to a real-world robot with sensors that function correctly 80% of the time. Prior research relating automation bias and complacency with human trust in automation indicates that reliability of the automation is a strong predictor of whether a human remains cognitively focused on the task at hand (Parasuraman, 1997; Wright et al., 2020). As discussed, the reliability is manipulated by changing the transition probability of the MDP such that we have distinct frequencies of transitioning and remaining in a specific state, resulting in distinct overall behavior. Thus, we define two different levels of reliability: High and Low, each with their own set of probabilities. The High condition remains in the Clean state for 90% of the time, whereas the Low condition remains for 25% of the time. A visualization of each condition's MDP is shown in Figure 2.

We measure performance, situation awareness, and subjective predispositions and attitudes towards automated agents through survey interventions. Our selected metrics to record from our task closely match prior research done in the human-agent teaming domain, as consolidated in a literature and meta-research review by O'Neill et al. (O'Neill et al., 2020).

Performance

Performance is measured by tallying the amount of captures by the participant and

captures by the team. As the prey is allowed an intangibility period after a capture to prevent repeat captures, the tally does not include captures while the prey is intangible. Performance is the main dependent variable analogous to success in real-life scenarios. The resulting performance of an agent will be a function of its reliability, the environment, and its interactions with other agents. All positional, input, and agent decision data is recorded for replay and re-simulation, resulting in multiple time series apt for analysis.

Situation Awareness

We employed a Situation Awareness Global Assessment Technique (SAGAT), often defined as the gold standard in measuring situation awareness (SA) (Endsley, 1988b,a). A SAGAT consists of interrupting the current task, disabling all interfaces and hiding all relevant artifacts that would provide knowledge about the scenario, and asking the participant to recreate the situation from memory. For the PPG, during a specified round at a random time, the game freezes and all players are removed from the play area. Then, the participant is tasked to recreate the position and direction of all players by dragging icons into the play area. This SAGAT measures perception of data and the scenario, often referred to as Level 1 SA (Endsley, 1988a, 1995). After participants complete the queries, they are allowed to resume the remainder of that round, with that data omitted when analyzing performance. An overview of the situation awareness probe can be seen in Figure 3.

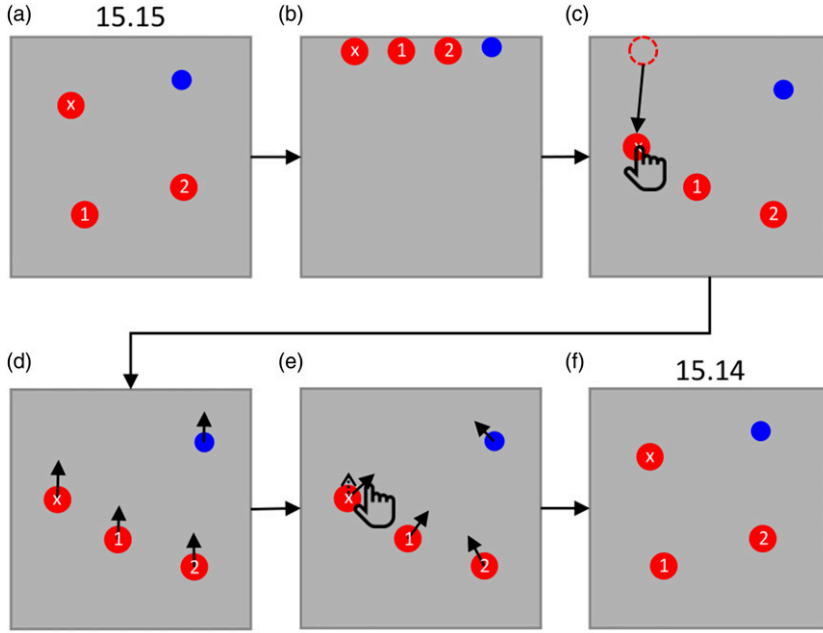


Figure 3. Situation awareness freeze probe done mid-round. (a) Participants play the game until a random time in the trial. (b) At the start of the probe, all players are moved at the top, and participants are prompted to (c) move the players to their last known position (as seen in (a)). (d) Arrows then appear in the position of all players, and participants are prompted to (e) drag the arrows in the directions toward which they were moving. (f) After completing the probe, participants are allowed to resume the round.

To measure SA, we establish two equations to quantify the amount of awareness from the participant's response (R) to the truth (T), that is, the state of the game when the task froze. Positional SA (SA_p) is defined as the Euclidean distance from the participant's positional response to the truth, divided by 2.83 (which is the maximum distance achievable in a 2 m by 2 m area), complemented with 1

$$SA_p(R, T) = 1 - \frac{\sqrt{(T_x - R_x)^2 + (T_y - R_y)^2}}{2.83} \quad (1)$$

Directional SA (SA_d) is the cosine similarity from the participant's vector response to the truth, normalized to a unit value

$$SA_d(R, T) = \frac{\left(\frac{R \cdot T}{\|R\| \|T\|} \right) + 1}{2} \quad (2)$$

Thus, as both measures approach 1, situation awareness is maximized. Both SA_p and SA_d are averaged to calculate a final value for situation awareness

$$SA = \frac{SA_p + SA_d}{2} \quad (3)$$

Survey Intervention

We used subsets of multiple validated surveys in order to measure relevant individual differences and human perception to model attitudes of trust on the agents. These surveys were abridged by selecting items with the highest factor loadings per survey factor to reduce the time needed to complete the experiment, as we had strict time limitations due to the nature of crowdsourcing payouts. We divide our surveys into two categories by utility: predisposition instruments and perception instruments. The selected questions and factor loadings are outlined in [Table 1](#).

Table 1. Survey interventions for the Predator-Prey Game. Items marked with (R) are reverse scored. The factor loadings are given by the related literature; NA indicates no factor loadings were found, thus multiple items were used.

Code	Item	Factor	Loading
Predisposition Instruments			
Automation-induced Complacency Potential (Merritt et al. (2019))			
aicp1	If life were busy, I would let an automated system handle some tasks for me	Alleviation	0.78
aicp2	It's not usually necessary to pay much attention to automation when it is running	Monitoring	0.67
Adapted Propensity to Trust Technology (Jessup et al. (2019))			
aptt1	Generally, I trust automated agents	Trust Propensity	NA
aptt2	Automated agents help me solve many problems		NA
aptt3	I don't trust the information I get from automated agents. (R)		NA
aptt4	Automated agents are reliable		NA
Perception Instruments			
Intrinsic Motivation Inventory (Ryan and Deci (2000) ; McAuley et al. (1989))			
imi1	I enjoyed playing with this team very much	Interest	0.8
imi2	I think I am pretty good at playing with this team	Competence	0.97
imi3	I tried very hard while playing with this team	Effort	0.85
imi4	I felt pressured while playing with this team	Tension	0.72
Trust in Automated Systems (Jian et al. (2000) ; Spain et al. (2008))			
tas1	My team is dependable	Trust	0.88
tas2	I am wary of my team. (R)	Distrust	0.87
NASA Task Load Index (Hart and Staveland (1988))			
tlx-me	How mentally demanding was playing with this team?	Mental Strain	NA
tlx-ph	How physically demanding was playing with this team?	Physical Strain	NA
tlx-te	How hurried or rushed was the pacing of this team?	Tension	NA
tlx-pe	How successful were you in capturing the prey with this team?	Competence	NA
tlx-ef	How hard did you have to work to capture the prey with this team?	Effort	NA
tlx-fr	How discouraged, stressed, and annoyed were you while playing with this team?	Tension	NA

Predisposition Instruments

The predisposition instruments are comprised of the Automation-induced Complacency Potential (AICP) scale and the adapted Propensity to Trust Technology (aPTT) scale.

The AICP scale measures a participant's tendency towards sub-optimal monitoring

patterns through two factors: workload alleviation and frequency of monitoring (Merritt et al., 2019). The highest factor loadings for the AICP was included in its manuscript, resulting in two questions.

The aPTT scale measures a participant's tendency to trust technology, modified to

include language explicitly referring to “automated agents” rather than technology (Jessup et al., 2019). According to Jessup et al., using the specific language of “automated agents” allows the measure to predict behavioral trust apt for the PPG. The aPTT did not include factor loadings, thus we selected items that were unique and non-congruent in nature (e.g., “Automated agents are reliable” and “I rely on automated agents” are conceptually similar, so only one of them was included). Both of these scales measure a participant’s bias to trust or distrust technology, which serves useful for predicting potential issues with trust calibration. We expect both scales to positively correlate the participants’ propensity to trust automation and their perceived trust of the automated teammates.

Perception Instruments

The perception instruments are comprised of the Intrinsic Motivation Inventory (IMI), Trust in Automated Systems scale (TAS), and the NASA Task Load Index (NASA-TLX).

The IMI focuses on task evaluation, measuring interest and enjoyment, perceived competence, effort and importance, and pressure and tension (Ryan & Deci, 2000). Designed to be reworded to contextualize the inventory for the task, questions were modified to make them relevant to the PPG (e.g., “I felt pretty skilled when cooperating with this team” instead of “I felt pretty skilled at this task”). The IMI was given to measure changes in motivation, as has been known to accurately predict performance and engagement in game-based tasks (Ryan et al., 2006). The IMI factor loadings were given by a confirmatory factor analysis (McAuley et al., 1989), resulting in four questions.

The TAS scale measures how trustworthy the participants perceived the system that they just interacted with—in our case, the automated teammates (Jian et al., 2000)—to be. The TAS factor loadings were given by a confirmatory factor analysis (Spain et al., 2008), resulting in two questions. The questions in this scale were rephrased with the automated teammates as the object of reference. We expect the observed trust measured by this scale to be related to the previous aPTT.

The NASA-TLX is a prevalent and strongly validated tool to measure perceived workload throughout a task (Hart & Staveland, 1988). Since complacency might relate to the amount of workload a participant perceives, we expect to find a correlation between reliability and perceived workload. Additionally, any variance demonstrated from predicting complacent behavior using reliability could be clarified using workload.

Additionally, a demographic survey was employed to collect participant information, including their age, gender, race/ethnicity, education, and experience with video games (on hours per week). Game experience was collected to model any performance variance due to familiarity with game-based tasks.

Latent Growth Modeling and Structural Equation Modeling

In order to answer our research questions, we turn to latent variable modeling as a method to investigate hidden relationships between measures and extract valuable information from the variance often ignored in traditional inferential statistics approaches such as ANOVAs (Duncan & Duncan, 2009). In addition to finding any effects given by varying reliability, we aim to discover any changes in participants’ perception towards the agents and situation awareness, as valuable information is found when noting changes as the participants interact with the agents (Desai et al., 2013).

Latent growth modeling establishes two factors that capture differences between groups and over time, often referred to as the Intercept factor and the Slope factor. For clarity, we refer to group differences as the Base factor and growth differences as the Change factor. The Base factor determines whether there is a difference between observed measures in groups, whereas the Change factor determines if there is a difference in the trajectory of the observations over time between groups. Therefore, two groups may begin at the same quantification of a given observation (e.g., trust), and then diverge over time. This allows us to determine not only if but how much attitudes and perceptions of the agents change through the trials.

Furthermore, we use structural equation modeling (SEM) to determine mediations through variables. Most effects found in HAT models cannot be observed in a vacuum, as there is a wide variety of interdependence between human characteristics and individual differences that interplay and affect performance and SA in any given task. Guided by literature, we hypothesize an initial SEM and iterate the model by modifying pathways and gauging the strength of the relationships between variables, aiming to find a pathway from reliability to performance. The initial SEM consisted of the following associations:

- Reliability drives the performance of the automated agents. Thus, it will directly affect team performance (2 automated agents and 1 human operator) and will correlate with individual performance.
- Reliability affects the perceived trust by the operator, as trust and reliability are known to be correlated in prior research (Lee et al., 2017). As reliability is the independent variable that can be controlled, it directs towards trust.
- Propensity to trust technology (measured by aPTT) predicts the estimated trust by the human to an agent, thus driving the demonstrated trust measured by the TAS (Jessup et al., 2019).
- Complacency potential (measured by AICP) also drives the perceived trust by the human, as overtrusting automation usually leads to the presence of complacent behavior (Parasuraman, 1997; Parasuraman et al., 1993; Parasuraman & Manzey, 2010). Thus, if there is varying complacency potential, we expect it to drive perceived trust.
- Propensity to trust technology and complacency potential are correlated through trust (i.e., overtrust) (Lee et al., 2017; Parasuraman & Manzey, 2010).
- Perceived trust affects motivation, under the rationale that trusting automation may lead to less incentive from the operator to perform (i.e., the value of their contribution is lessened by the capability of the automation, “The agent is doing well, I think I’ll sit back and allow it to do it’s job”) (Shepperd, 1993).
- The operator’s motivation and complacency potential drive their individual performance in the PPG, as combinations of these factors lead operators to retain certain mental states during the task (e.g., absentminded vs. disengaged).

Procedure

Participants were recruited through Amazon Mechanical Turk (AMT)¹, an online crowdsourcing platform. Participants selected the associated Human Intelligence Task (HIT) from AMT and signed an initial consent form. Participants were then redirected to VolunteerScience² to complete the PPG. Participants first filled the demographic survey, read instructions on how to play the PPG, and then completed two practice rounds (no recorded data) followed by 20 recorded rounds of 30 seconds each. Both the pre-disposition and perception instruments were administered upon completion of rounds 2, 10, and 18. The situation awareness probe freeze was administered at a random interval between 10 and 20 seconds into rounds 4, 10, and 16. Participants were allowed four rounds (+2 practice rounds) to interact with the predator agents before the first situation awareness probe freeze, as any changes in situation awareness would not be reflected immediately upon interaction, but requires a certain amount of time for complacency to impact the participant’s cognitive state. Reliability was treated between-subjects: participants completed the entire PPG with one level of reliability, where both predator teammates were set to the same level. The experiment flow is visualized in Figure 4.

Results

Demographics

104 participants completed the PPG. Demographics are reported in Table 2. In summary, 60% of participants were male, 46% of participants were in the age range of 25–34, 65% finished a 4-year college education, and 35% played videogames for 2–4 hours a week. All participants were compensated 10 USD for completing the experiment. The collected data

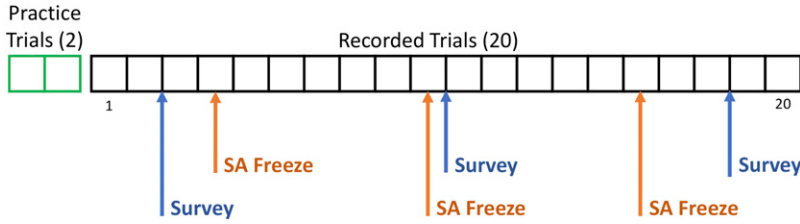


Figure 4. Experiment flow of the PPG experiment.

Table 2. Resulting demographics for the PPG experiment. Categories with 0 participants in all conditions were not included.

	High	Low
Sample size (n)	52	52
Age		
18–24	2	4
25–34	25	24
35–44	16	13
45–54	6	6
55–64	3	4
65+	0	1
Gender		
Male	32	30
Female	20	22
Race/Ethnicity		
White	47	45
African American	2	2
Native American	2	2
Asian	0	2
Hispanic/Latino	1	1
Highest education completed		
High school	8	6
2-year college	3	3
4-year college	34	33
Graduate/Professional	10	7
Hours/Week playing videogames		
<1	0	4
1–2	14	15
2–4	20	16
4–7	11	11
>7	7	6

was checked for satisficing and completion, resulting in no records dropped.

Inferential Statistics, Behavior, and Growth Modeling

Performance. We compare the performance of both the individual participant and the

whole team across the two reliability conditions. Since the number of captures is count data (i.e., non-negative integers with a low tendency of centrality), we conduct non-parametric statistical inferences from participants playing the PPG. The score of every trial is treated as a single data point. Plotted scores across trials and condition are visualized in Figure 5.

We conducted a Mann–Whitney U test to determine significant differences between conditions in individual performance and team performance. For individual performance, the test revealed no difference between the High and Low conditions (μ : 1.45 vs. 1.38, M : 1 vs. 1, U = 555210, n = 2080, p = 0.27). For team performance, the test revealed that there was a statistically significant difference between the High and Low conditions (μ : 3.5 vs. 3.87, M : 3 vs. 4, U = 485150.5, n = 2080, p < 0.001).

We compare the performance of each predator (including the participant) within the team per condition. In the High reliability condition, a Kruskal–Wallis test and follow-up Dunn’s test indicate that each predator’s performance was significantly distinct (χ^2 (2, 3119) = 82.93, p < 0.001; all pairwise comparisons: p < 0.05). The highest performing member was the participant, followed by the Interceptor, with the Chaser trailing last. In the Low reliability condition, a Kruskal–Wallis test found no significant differences in the performance of each member of the team (χ^2 (2, 3119) = 0.15, p = 0.92).

To ascertain the contribution of each agent predator, we conduct Mann–Whitney U tests across reliability conditions for each predator. The Chaser had a significant increase in its performance in the Low condition (μ : 0.85 vs. 1.26, M : 1 vs. 1, U = 436695, n = 2080, p <

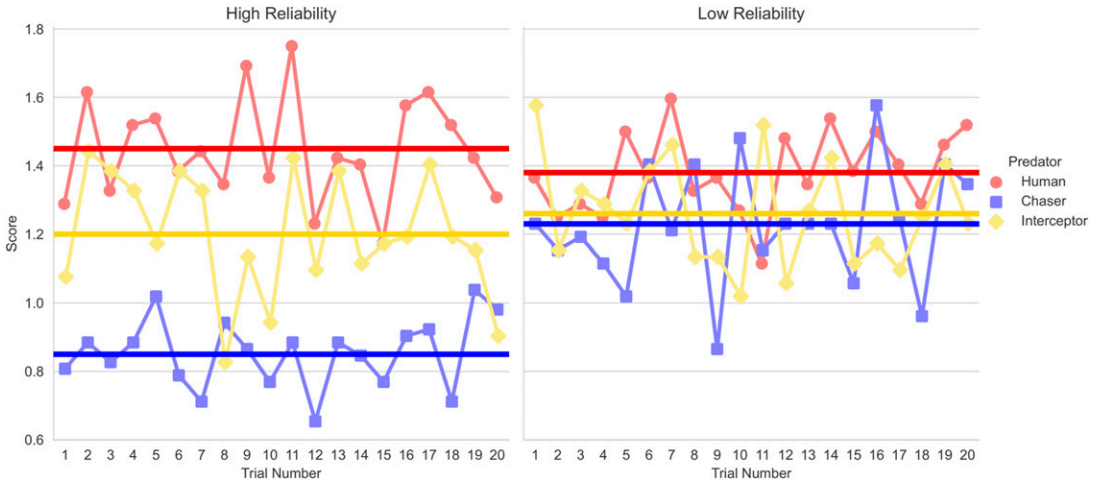


Figure 5. Plotted scores for every member of the predator team across trials. Each point in the x-axis is the mean of all scores for a given trial. The horizontal lines denote the average for the predator according to the legend. Comparing the High to Low conditions, the Chaser predator substantially increased its performance (0.85 → 1.23 captures; blue). Non-significant effects were found in the Human predator's performance (1.45 → 1.38 captures; red) and the Interceptor predator's performance (1.2 → 1.26 captures; yellow).

0.001), increasing its contribution by 0.41 captures per trial. In contrast, the Interceptor had a close but non-significant increase in its performance in the Low condition (μ : 1.20 vs. 1.26, M : 1 vs. 1, $U = 515826.5$, $n = 2080$, $p < 0.001$), increasing its contribution by 0.06 captures per trial.

Movement behavior. We visualize the positional data of all players through heatmaps in Figure 6. This denotes how frequently a player was located in a given area of the arena. From here, we draw several important comparisons. First, we see the static strategy the High reliability Chaser exhibits: the prey's best strategy to escape was to circle around the area away from the predators, and the Chaser followed through by using the same circling as defined by its strategy. It is only when noise is added to the Chaser where it chases all across the arena, similar to an Interceptor (both in High and Low reliability). Interestingly, the human has a high incidence of remaining in the corners of the area, potentially demonstrating a “wait and attack” strategy, where it allows the other agents to draw the prey to the corner of the arena where the human awaits for a quick capture. Second, one argument to be brought

forth is that the more coverage an agent has, the higher number of captures it can potentially achieve, as exemplified by the Chaser and Interceptor. The number of Chaser and Interceptor captures increased in the Low condition, which aligns with the increased coverage shown in their respective heatmaps. Similarly, coverage subtly decreases for the human in the Low condition (areas beyond radius 0.5) along with its performance (although this was found to be non-significant). Finally, the distinct coverages from the High and Low reliabilities led to the prey to adopt a distinct strategy, as the Prey could mostly avoid the predators by staying adjacent to the edges of the arena, but once coverage was increased, the prey adjusted by considering new routes closer to the center of the arena.

Survey Interventions and Situation Awareness Probe

As discussed, survey responses and the SA probe were analyzed using latent growth curve modeling (Duncan & Duncan, 2009; Raykov, 1994) in order to shed light not only on subjective differences between the High and Low conditions but on how these differences change

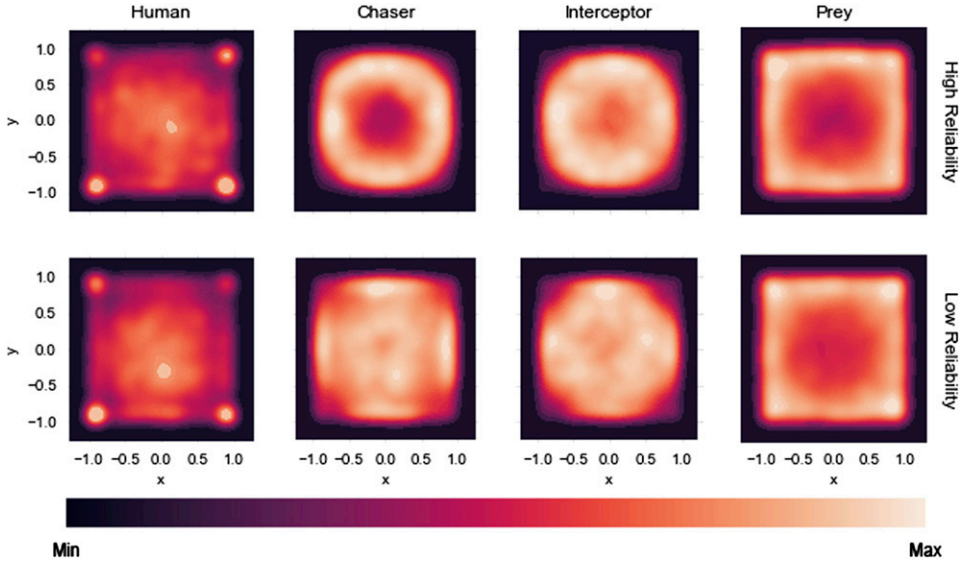


Figure 6. Position heatmaps for all players ($n = 104$) and conditions in the PPG across all trials. Positions were recorded at 4 Hz, resulting in nearly 1,000,000 data points. The top row covers the High condition, the bottom row covers the Low condition. The gradient represents the frequency of positions.

over time as participants interact with the automated agents. Therefore, we analyze each survey question with a growth curve model, along with aggregating the questions with their respective instrument. A summary of the growth curve model results can be found in Table 3.

For the predisposition instruments, we had expected and found no change between conditions or through time (i.e., non-significant Base and Change factors) since complacency potential and propensity to trust technology is a predisposed attitude rather than being affected by the intervention. For the perception instruments, there was no significant change per condition or through time, demonstrated by the non-significant p -values in both the Base and Change factors. Responses over time are plotted in Figure 7 and Figure 8.

For all positional and directional situation awareness in the growth curve model, there were no significant differences between conditions or through time for all PPG players, except for prey directionality in the Base factor ($p < 0.05$). Summary statistics of the probe results along with the p -values of the growth curve model are found in Table 4, and responses over time are plotted in Figure 9.

Structural Equation Model Fit

Much prior research shows complex relationships between system reliability, human performance, and individual differences that prove difficult to investigate with inferential techniques. Developing an SEM to describe these relationships allows us to paint a clearer picture of what factors alter performance and situation awareness in continuous pursuit tasks. Figure 10 outlines a visual representation of the final SEM, defining relationships between exogenous and endogenous variables, along with fit and regression coefficients. The numbers reported on every arrow are coefficients (B), which predict a change on the regressand when the generating regressor is increased. The SEM model was built using R 3.5.2 with lavaan 0.6-7 (Rosseel, 2012).

Both individual and team performance were aggregated for fit in the SEM by taking the median of a participant's trials. Survey items were aggregated for analysis via parceling, accounting for reverse scoring and common factors. Parceling allows us to improve estimates and model fit by reducing measuring error through aggregation (Matsunaga, 2008).

Multiplicity control using the False Discovery Rate (FDR) is often recommended for exploratory SEM analysis (Cribbie, 2007) to prevent the issue of multiple comparisons (as every model fit is a new hypothesis). This is because Standard Family-wise Error Rate

controlling procedures (e.g., Bonferroni) are rather conservative for exploratory purposes. We controlled the FDR through the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) with Q set at 0.15, which allows us to determine which factors in the final SEM model are expected to be false positives. The total number of models tested for the PPG and effects is 46.

Table 3. Survey results and growth curve modeling summary. The bolded represents the parceled aggregation of the scale.

	High M (MAD)	Low M (MAD)	Base $p (> z)$	Change $p (> z)$
aicp1	6 (0.99)	5 (1.03)	0.34	0.58
aicp2	5 (1.38)	5 (1.39)	0.61	0.73
aicp	5.5 (1.05)	5 (1.09)	0.4	0.55
aptt1	5 (0.87)	5 (0.97)	0.5	0.67
aptt2	5 (0.96)	5 (1.14)	0.83	0.48
aptt3	5 (1.44)	5 (1.33)	0.9	0.87
aptt4	5.5 (0.85)	5 (1)	0.87	0.22
aptt	5 (0.66)	5 (0.7)	0.97	0.43
imi1	6 (0.91)	6 (0.88)	0.37	0.77
imi2	5 (1.04)	5 (1.13)	0.85	0.6
imi3	6 (0.71)	6 (0.99)	0.17	0.06
imi4	5 (1.2)	5 (1.28)	0.82	0.85
imi	5.5 (0.62)	5.5 (0.66)	0.85	0.73
tas1	5 (0.91)	5 (1.04)	0.27	0.4
tas2	5 (1.32)	5 (1.18)	0.23	0.33
tas	4.5 (0.72)	4 (0.76)	0.09	0.21
tlx-me	70 (13.71)	75 (12.56)	0.50	0.09
tlx-ph	65 (18.25)	70 (22.7)	0.73	0.54
tlx-te	70 (12.8)	75 (12.95)	0.67	0.73
tlx-pe	70 (15.64)	70 (16.7)	0.85	0.99
tlx-ef	75 (11.15)	75 (10.7)	0.65	0.47
tlx-fr	65 (22.52)	65 (22.31)	0.88	0.24

Discussion

We situate our findings by first discussing the results from inferential statistics and growth curve modeling, and we use those to guide our exploratory SEM analysis. Next, we discuss the implications of these effects in the context of continuous cooperative multi-agent systems, and finally, we answer our research questions and highlight future research challenges.

Statistics and Growth Curve Modeling

We begin by noting the performance of the predator team during the PPG. We find that individual human performance is slightly higher when the other team members are not impacted by noise ($\mu_H = 1.45$, $\mu_L = 1.38$). Such behavior could indicate the human's ability to predict future positions of their teammates and optimize their possibilities to capture the prey. Albeit the mean difference between the individual performance of the two conditions is 0.07, the calculated effect size is 0.91—that is, if we generate all ordered pairs between participants in the High condition and participants in the Low condition, in 91% of the pairs, the High

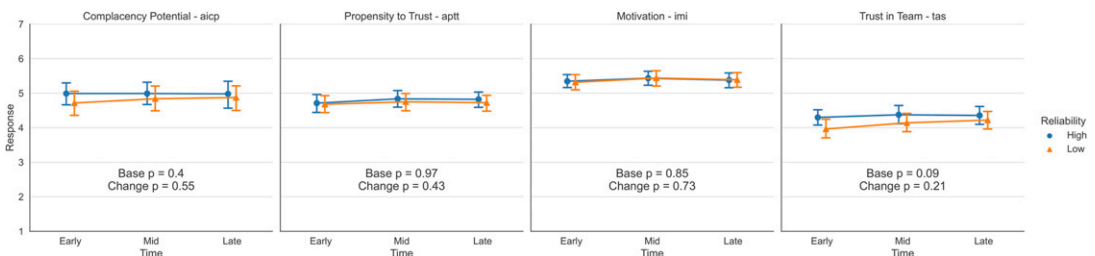


Figure 7. Plotted Likert response average with error bars over time for parceled survey interventions (except NASA-TLX responses). The p -value of the growth curve model factors (Base and Change) are annotated below. 1 = Lowest, 7 = Highest.

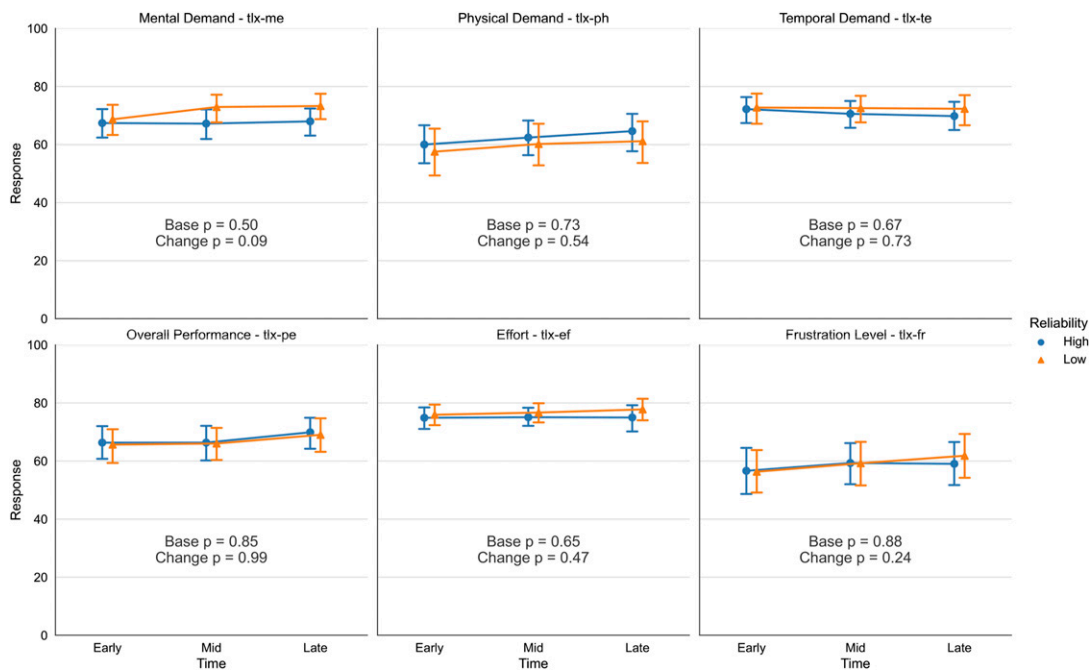


Figure 8. Plotted response average with error bars over time for NASA-TLX items. The p -value of the growth curve model factors (Base and Change) are annotated.

Table 4. Situation Awareness probe results and growth curve modeling summary. Values were rounded to two significant figures. H = High condition, L = Low condition.

Probe	Cond	N	Mean	StDev	Min	Max	Skewness	Kurtosis	Base $p (> z)$	Change $p (> z)$
Position										
Self	H	156	0.68	0.18	0.17	0.98	-0.52	-0.55	0.81	0.41
	L	156	0.71	0.18	0.20	0.98	-0.59	-0.34		
Chaser	H	156	0.68	0.16	0.27	0.98	-0.26	-0.55	0.84	0.89
	L	156	0.68	0.16	0.24	0.98	-0.26	-0.63		
Interceptor	H	156	0.70	0.14	0.32	0.97	-0.24	-0.64	0.33	0.23
	L	156	0.70	0.16	0.33	1.00	-0.37	-0.54		
Prey	H	156	0.69	0.19	0.17	0.99	-0.60	-0.42	0.54	0.2
	L	156	0.70	0.19	0.13	0.99	-0.64	-0.26		
Direction										
Self	H	156	0.59	0.33	0.00	1.00	-0.38	-1.13	0.3	0.73
	L	156	0.54	0.34	0.00	1.00	-0.13	-1.37		
Chaser	H	156	0.58	0.36	0.00	1.00	-0.39	-1.43	0.47	0.29
	L	156	0.58	0.35	0.00	1.00	-0.37	-1.31		
Interceptor	H	156	0.56	0.36	0.00	1.00	-0.23	-1.52	0.69	0.35
	L	156	0.54	0.33	0.00	1.00	-0.14	-1.32		
Prey	H	156	0.56	0.35	0.00	1.00	-0.25	-1.45	0.049*	0.13
	L	156	0.51	0.34	0.00	1.00	-0.05	-1.43		

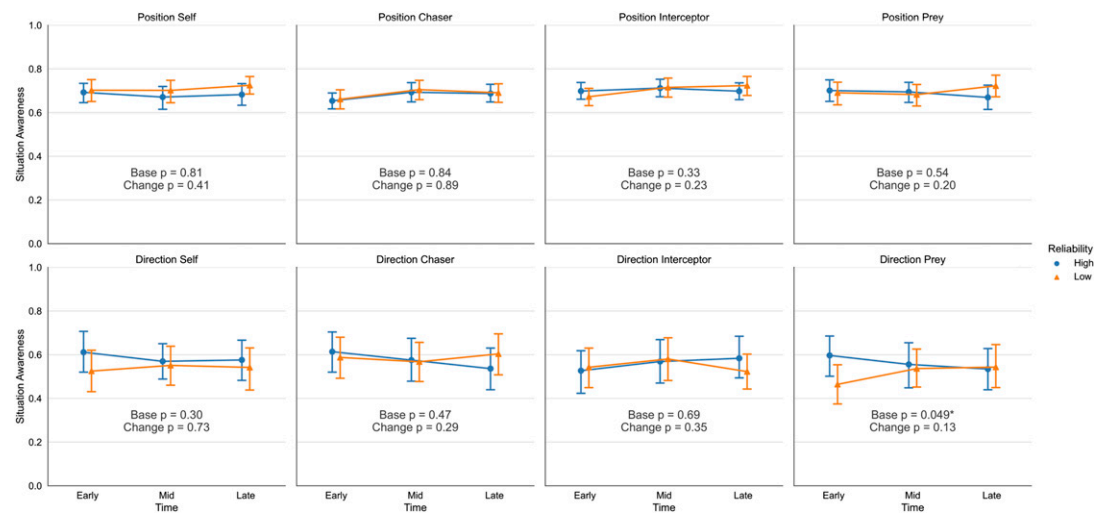


Figure 9. Plotted situation awareness response average with error bars over time for the situation awareness probe. The p -value of the growth curve model factors (Base and Change) are annotated.

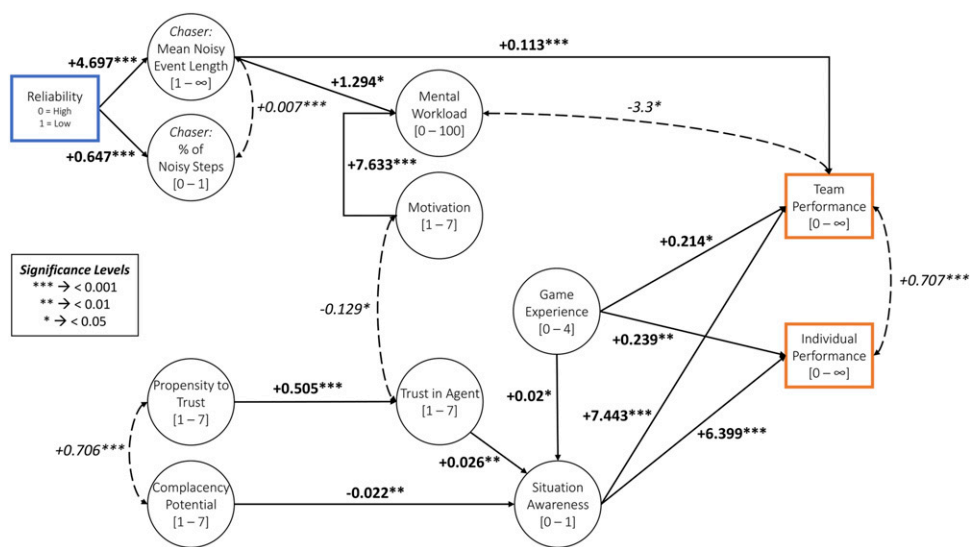


Figure 10. Fitted SEM for the PPG experiment. Model fit: $N = 104$ with 36 free parameters. RMSEA = 0.037 (CI = [0, 0.074]), TLI = 0.991, CFI = 0.994, over null baseline model, $\chi^2(78) = 1363.238$. Solid lines indicate significant regressions, and dashed lines indicate correlations. Numbers represent the fitted coefficient along with its level of significance.

condition participant outperforms the Low condition participant. More than individual performance, however, we observe that the Chaser agent had far superior performance in the Low condition compared to the High condition, in contrast to the reduced effects from the human

and the Interceptor agent. We hypothesize that by adding a random perturbation vector to the target point of the Chaser agent during its Noisy state, its actions were subtly altered to match the Interceptor behavior (as the Interceptor strategy is theoretically equal to the Chaser, with

a consistent target position offset accounting for the prey's position and velocity). Additionally, the perturbation could have led the predators to adopt unpredictable behavior that became difficult for the prey to account for. By having two interceptors in a team—with interceptor behavior being a very viable strategy in a continuous pursuit game (Fernandez et al., 2021; Foley & Schmitendorf, 1974; Ho et al., 1965)—we note the difference in team performance across conditions. This hypothesis aligns with the position heatmaps visualized in Figure 6; to reiterate, a Noisy Chaser had more coverage across the arena, and had a similar coverage to an Interceptor agent, overall forcing the prey to cover new routes. Further analysis can be conducted through histogram analysis, as previous work has shown that the distribution of the distance between predators and prey is the most characteristic feature of agent strategy (according to a Principal Component Analysis), and is a strong indicator of Chaser versus Interceptor behavior within teams (Howell et al., 2021). It is interesting to note how unintended noise can bring benefit to the agent, as it begins exploring alternative strategies that may help it achieve its goal at a more unorthodox pace, much akin to research that suggests that a perfect, transparent agent may not be the perfect teammate (Bansal, Nushi, Kamar, Weld, et al., 2019, 2020; Stumpf, 2016). Alternatively, however, we should be cognizant of potential overtrusting that can occur if agents indeed receive unintended benefits from a lowered reliability.

We opted to take measurements for the survey instruments at the early, middle, and late interactions with the agents to determine any changes through time, as interactions and impressions with automated agents are often conditioned by the worst recalled interaction (Lee & See, 2004; Muir & Moray, 1996). Thus, if any negative interactions occurred with the Low condition agents, it would rapidly reflect on the items. For the predisposition instruments, we found no significant change through time, as we expected that beliefs already held before interaction with the agents would not change with a short amount of interaction (it took participants around 15–20 minutes to complete the experiment, with 10 minutes directly interacting with the agents). In

a similar vein, the perception instruments show that participants did not perceive a change in motivation or trust across conditions or through time. We take note of the most significant factors to inform the construction of our exploratory SEM.

The SA growth curve model demonstrates consistent SA across all probes, conditions, and time, except for the awareness of prey direction. This indicates that participants had adjusted to an information processing strategy that was consistent across 20 trials, regardless of agent reliability. This may suggest that participants did not perceive the Low reliability agents to be any different from the High reliability agents or require a different strategy to keep track of their state (such as vigilantly monitoring a vehemently unreliable agent). The exception present in the prey direction probe where the Base factor was significant indicates a distinct level of initial situation awareness by completion of the first probe (High SA_d : 0.594 vs. Low SA_d : 0.469, $p < 0.05$). This may be explained by an initial adjustment of expectations with respect to behavior from the predators in the Low condition: the perceived erraticness of the participants' teammates could have led to higher vigilance towards them while reducing awareness of the prey's direction. At future probes, the situation awareness of Low condition participants increased linearly, indicating that they had adjusted to such behavior, whereas the High condition participants saw a linear reduction, possibly indicating overtrust and, consequently, loss of situation awareness. However, the change factor indicates non-significance, even though the change is opposite per condition (most likely due to the small magnitude of the change slope: -0.031 vs. 0.042). Future research may address specific points of interests and objects of focus that lead to varying situation awareness.

Structural Equation Model Effects

We use our SEM to outline relationships between factors and answer our research questions. As this exploratory model aims to inform future research, we encourage readers to not take the described model as a definitive description of the human cognitive process with respect to

reliability and performance, but to consider the mediated effects when formulating future research.

An initial point of discussion is how well each MDP executed to produce distinct reliability behavior between the two agents. After all, if at every step the agent would have alternated between the Clean and Noisy states, behaviorally, the agent would have continued its original trajectory after correcting for the incorrect state with non-perceptible error. Hence, we hypothesized that a longer stay in the Noisy state would produce erratic behavior perceived by a human. Thus, we considered both the percent of Noisy steps (i.e., how much time the agent was in the Noisy state), and the average length of a Noisy event (i.e., how long the agent remained in the Noisy state before transitioning out to the Clean state). We found that for both the Chaser and Interceptor agents, the High condition had 10% of total Noisy steps, with a mean Noisy length of 4 steps (1 second), whereas the Low condition had 75% of total Noisy steps, with a mean Noisy length of 8 steps (2 seconds). As the surveys and probes indicate, the slight difference in the length of the Noisy event might be imperceptible to humans but results in distinct behavior from the agents—behavior which leads to altered performance. The SEM shows that the MDP was able to manipulate the agents' behavior ($p < 0.001$), with the caveat that since the Interceptor behavior did not influence any further variables downstream, it was removed from the model. As expected, the percent of Noisy steps and the average Noisy length are correlated.

As to control a portion of the performance variance given participants' experience with game-based entertainment (since our task is a game-style task), we categorized participants based on their video gaming experience. This allows us to use their experience with such entertainment as a proxy to engagement, as we expect participants to remain interested throughout the task due to familiarity. Engagement with the PPG led to higher individual and team performance ($p < 0.05$). This strengthens other factors in the model to explain the remaining variance.

The length of the Noisy state and the participant's current motivation affected the mental

workload they perceived ($p < 0.05$). The erratic behavior directly influenced the team's performance positively ($p < 0.001$), as a direct effect from the increase in Chaser performance as discussed earlier. However, a negative correlation existed between a participant's mental workload and the performance of the team ($p < 0.05$). Mental workload may have been increased by the nature of the task, as the time pressure to capture the prey along with erratic behavior of the predator teammates may have affected coordination (Entin & Serfaty, 1999; Fan et al., 2010).

The predisposition instruments accurately predicted participants' attitudes on the reliability of the AI, validating the instruments' utility in continuous tasks. A participant's propensity to trust accurately predicted their amount of trust in the agents after the PPG ($p < 0.001$). Complacency potential predicted their level of situation awareness ($p < 0.01$), validating overtrust and automation complacency paradigms discussed in the literature (Bahner et al., 2008; Parasuraman & Manzey, 2010). Both of the predisposition instruments strongly correlated with each other, of almost one point on each scale ($p < 0.001$).

Trust and situation awareness were strong mediators between individual characteristics and observed performance. Complacency potential may describe an initial state of situation awareness (lowering SA more as the predisposition to complacency is higher) but is simultaneously increased by trust ($p < 0.01$) and engagement ($p < 0.05$). Interestingly, this is an opposite effect than what is usually found in SA-based studies and metrics, where often an increase in trust hampers awareness (Endsley, 2017; Endsley & Kiris, 1995) due to improper trust calibration (Lee & See, 2004). We are unable to determine whether the increase in situation awareness due to trust is driven by trust calibration since the model did not find a pathway from the agents' reliability to perceived trust. Another explanation is given by the nature of the task: the participant's active role in the PPG may have influenced them to exercise higher awareness due to their dependency on the other agents to capture the prey. That is, if the participant was alone with the prey, it would

become incredibly challenging for them to capture a prey due to the asymmetry of their capabilities. Much prior research often places participants hierarchically superior to automated or decision support systems, where little time pressure is found for participants to make a decision, often deferring their judgment to an AI. Further investigation about the benefits of the nature of the task is warranted, as trust, engagement, and complacency potential only describe a small amount of the effect ($R^2 = 0.142$), albeit very significantly. A possible explanation is that video gaming experience allows for increased multitasking performance, which allows for more effective situation awareness (Chen et al., 2011). On the other hand, we found that motivation and trust in the agents are negatively correlated ($p < 0.05$), echoing an effect similar to social loafing in working groups (i.e., the less one is motivated, the more they will trust the AI out of inactivity) (Karau & Williams, 1993; Latané et al., 2006).

Ultimately, performance was largely mediated by situation awareness ($p < 0.001$) and engagement, explaining almost half of the variance in both individual and team performance ($R^2 = 0.428, 0.411$). Breaking down the variance contributors, we find that for individual performance, the variance was explained by game experience (26%) and situation awareness (74%). Similarly, for team performance, the variance was explained by game experience (18%), how long agents were in a Noisy state (18%), and situation awareness (64%). The behavior of the agents influenced participants' perceived workload, but performance in the task was predicted by situation awareness. A correlative bridge exists between motivation and trust, yet further research is required to solidify these relationships with respect to continuous contexts. In light of this model, we return to our research questions:

1. How is human performance affected by varying reliability in collaborating agents in a continuous pursuit task?

According to the resulting SEM, human performance was not directly affected by agent reliability, but reliability was mediated by pre-dispositional and perceptual factors that

ultimately influenced resulting performance. A caveat does exist with team performance, as slight noise added to the Chaser agent resulted in increased performance due to its hypothesized similarity to an Interceptor, largely due to the nature of the task facilitating the use of optimal strategies. This resulted in a more capable teammate and a higher share of captures for the agents, all while being minimally perceptible to the human operator.

2. How do human individual differences interplay with performance, situation awareness, and perception of agents in continuous domains?

Constructs such as complacency, engagement, motivation, and social loafing were observed in Human-Agent Teams in game-theoretic continuous tasks. We observed that multiple individual characteristics (such as propensity to trust and complacency potential) affected performance, mediated by trust and situation awareness. In continuous tasks where the reliability of a system may not be defined through a number or probability, the system's behavior can provide a strong signal for the operator to make a judgment on whether the agent is behaving correctly. Depending on the situation or scenario, agents may be subject to a varied amount of noise (either internal or external), which may result in inaccurate perceptions of the agents' efficacy. This work has shown that a system may be highly erroneous due to system failure or environmental constraints yet can be perceived as effective as an agent operating correctly.

Limitations and Future Work

This study is not without its limitations. Determining trust during a continuous task is a challenging prospect that research should focus on addressing, as we develop embodied AI agents that are active operators of a task holding a certain amount of responsibility and, consequently, risk. We were limited in measuring trust at discrete points, thus losing granular information on what events can cause trust to change during the task. Additionally, this particular task (continuous pursuit) ends up being abstract and specific for a certain kind of utility,

and thus factors discussed here may change depending on the domain, task, or situation. As this work is meant to be exploratory, research can focus on solidifying and replicating certain effects in the context of continuous tasks, as the growth of our AI-based systems now often operate in these domains. For instance, the perceptible difference of AI performance is a common effect found when controlling for reliability (Rodriguez et al., 2019; Schaffer et al., 2018), yet this threshold is not well defined.

Conclusion

Our world is continuous, and we are slowly beginning to integrate AI-based systems into continuous spaces that will allow us to achieve our goals more efficiently with active agents rather than supervisory control. We performed a study demonstrating how controlling reliability of an AI can affect how people perceive, trust, and perform in human-agent teams. We also emphasize using mediation models to connect cognitive factors often studied in a vacuum and present a holistic view on how agent performance can ultimately affect individual performance. Albeit this work uses an abstract task to demonstrate the studied effects, many operative environments may have similar structures or goals as those presented in the PPG. In the future, replacing the task with well-defined goals will allow us to verify the effectiveness of modeling cognition in HATs through mediation analysis. If we ever aim to have agents to walk alongside us, research should continue to uncover the factors that drive proper human-automation interaction in continuous domains.

Acknowledgments

We would like to thank James Schaffer for his advice in developing the mediation analysis in this work. We would like to thank the reviewers for their thoughtful insights on earlier drafts of our work.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was sponsored by the U.S. Army Combat Capabilities Development Command Army Research Laboratory. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

ORCID iD

Sebastian S. Rodriguez  <https://orcid.org/0000-0002-7003-1764>

Notes

1. <https://www.mturk.com/>
2. <https://volunteerscience.com/>

References

- Arnold, C., Collier, L., & Sutton, L. (2006) The Differential Use and Effect of Knowledge-Based System Explanations in Novice and Expert Judgment Decisions. *MIS Quarterly* 30(1): 79. <https://doi.org/10.2307/25148718>. URL <https://www.jstor.org/stable/10.2307/25148718>
- Arnold, V., Clark, N., Collier, P. A., Leech, S. A., & Sutton, S. G. (2004) Explanation provision and use in an intelligent decision aid. *Intelligent Systems in Accounting, Finance & Management* 12(1): 5-27. <https://doi.org/10.1002/isaf.222>. URL <http://doi.wiley.com/10.1002/isaf.222>
- Arrow, H., McGrath, J., & Berdahl, J. (2000) *Small Groups as Complex Systems: Formation, Coordination, Development, and Adaptation*. 2455 Teller Road, : SAGE Publications, Inc. ISBN 9780803972308. <https://doi.org/10.4135/9781452204666>. URL <http://sk.sagepub.com/books/small-groups-as-complex-systems>
- Asher, D., Garber-Barron, M., Rodriguez, S., Zaroukian, E., & Waytowich, N. (2019) Multi-Agent Coordination Profiles through State Space Perturbations. In: 2019 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE. ISBN 978-1-7281-5584-5, pp. 249-252. [10.1109/CSCI49370.2019.00051](https://doi.org/10.1109/CSCI49370.2019.00051). URL <https://ieeexplore.ieee.org/document/9070901/>
- Asher, D., Zaroukian, E., & Barton, S. (2018) *Adapting the Predator-Prey Game Theoretic Environment to Army Tactical Edge Scenarios with Computational Multiagent Systems* URL <http://arxiv.org/abs/1807.05806>
- Ashktorab, Z., Dugan, C., Johnson, J., Pan, Q., Zhang, W., Kumaravel, S., & Campbell, M. (2021) Effects of Communication Directionality and AI Agent Differences in Human-AI Interaction. CHI 2021 <https://doi.org/10.1145/3411764.3445256>
- Azhar, M. Q., & Sklar, E. I. (2017) A study measuring the impact of shared decision making in a human-robot team. *International Journal of Robotics Research* 36(5-7): 461-482. <https://doi.org/10.1177/0278364917710540>

- Bahner, J. E., Hüper, A. D., & Manzey, D. (2008) Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies* 66(9): 688-699. <https://doi.org/10.1016/j.ijhcs.2008.06.001>. URL <https://linkinghub.elsevier.com/retrieve/pii/S1071581908000724>
- Bansal, G., Nushi, B., Kamar, E., Horvitz, E., & Weld, D. S. (2020) Optimizing AI for Teamwork URL <http://arxiv.org/abs/2004.13102>
- Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019) Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7(1): 19. URL www.aaai.org
- Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019) Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 33, 2019 : 2429-2437*. <https://doi.org/10.1609/aaai.v33i01.33012429>
- Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., Dunning, I., Mourad, S., Larochelle, H., Bellemare, M. G., & Bowling, M. (2019) *The Hanabi Challenge: A New Frontier for AI Research* <https://doi.org/10.1016/j.artint.2019.103216>. URL <http://arxiv.org/abs/1902.00506>
- Barton, S., & Asher, D. (2018) Reinforcement learning framework for collaborative agents interacting with soldiers in dynamic military contexts. *SPIE 1065303*(2018): 2. <https://doi.org/10.1117/12.2303827>
- Benjamini, Y., & Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1): 289-300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>, URL <http://www.jstor.org/stable/2346101>
- Campbell, M., Hoane, A., & Hsu, F. (2002) Deep Blue. *Artificial Intelligence* 134(1-2): 57-83. [https://doi.org/10.1016/S0004-3702\(01\)00129-1](https://doi.org/10.1016/S0004-3702(01)00129-1). URL <https://linkinghub.elsevier.com/retrieve/pii/S000437020100129-1>
- Chen, J. Y. C., Barnes, M. J., Quinn, S. A., & Plew, W. (2011) Effectiveness of RoboLeader for Dynamic Re-Tasking in an Urban Environment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 55(1): 1501-1505. <https://doi.org/10.1177/1071181311551312>
- Chung, T. H., Hollinger, G. A., & Isler, V. (2011) Search and pursuit-evasion in mobile robotics A survey. *Autonomous Robots* 31(4): 299-316. <https://doi.org/10.1007/s10514-011-9241-4>
- Cribbie, R. A. (2007) Multiplicity Control in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal* 14(1): 98-112. <https://doi.org/10.1080/10705510709336738>
- DeCostanza, A. H., Marathe, A. R., Bohannon, A., Evans, A. W., Palazzolo, E. T., Metcalfe, J. S., & McDowell, K. (2018) Enhancing Human-Agent Teaming with Individualized, Adaptive Technologies: A Discussion of Critical Scientific Questions. *IEEE Brain (June)*: 38. <https://doi.org/10.13140/RG.2.2.12666.39364>. URL <https://apps.dtic.mil/docs/citations/AD1051552>
- Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013) Impact of robot failures and feedback on real-time trust. In: 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE. ISBN 978-1-4673-3101-2, pp. 251-258. <https://doi.org/10.1109/HRI.2013.6483596>. URL <http://ieeexplore.ieee.org/document/6483596/>
- Duncan, T. E., & Duncan, S. C. (2009) The ABC's of LGM: An Introductory Guide to Latent Variable Growth Curve Modeling. *Social and Personality Psychology Compass* 3(6): 979-991. <https://doi.org/10.1111/j.1751-9004.2009.00224.x>
- Endsley, M. R. (1988a) Design and Evaluation for Situation Awareness Enhancement. *Proceedings of the Human Factors Society Annual Meeting* 32(2): 97-101. <https://doi.org/10.1177/154193128803200221>
- Endsley, M. R. (1988b) Situation awareness global assessment technique (SAGAT). In: *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*. IEEE. ISBN 5856420187. <https://doi.org/10.1109/NAECON.1988.195097>. URL <http://ieeexplore.ieee.org/document/195097/>
- Endsley, M. R. (1995) Toward a Theory of Situation Awareness in Dynamic Systems. In: *Human Error in Aviation*, volume 37. Routledge. ISBN 9781315092898, pp. 217-249. <https://doi.org/10.4324/9781315092898-13>. URL <https://www.taylorfrancis.com/books/9781351563475/chapters/>
- Endsley, M. R. (2017) From Here to Autonomy: Lessons Learned from Human-Automation Research. *Human Factors* 59(1): 5-27. <https://doi.org/10.1177/0018720816681350>
- Endsley, M. R., & Kiris, E. O. (1995) The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37(2), 381, 394. <https://doi.org/10.1518/001872095779064555>
- Entin, E. E., & Serfaty, D. (1999) Adaptive Team Coordination. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 41(2): 312-325. <https://doi.org/10.1518/001872099779591196>
- Fan, X., McNeese, M., & Yen, J. (2010) NDM-Based Cognitive Agents for Supporting Decision-Making Teams. *Human-Computer Interaction* 25(3): 195-234. <https://doi.org/10.1080/07370020903586720>. URL <http://www.informaworld.com/openurl?genre=article>
- Fernandez, R., Zaroukian, E., Humann, J., Perelman, B., Dorothy, M., Rodriguez, S., & Asher, D. (2021) Emergent Heterogeneous Strategies from Homogeneous Capabilities in Multi-Agent Systems. *International Conference on Artificial Intelligence*.
- Foley, M., & Schmitendorf, W. (1974) A class of differential games with two pursuers versus one evader. *IEEE Transactions on Automatic Control* 19(3): 239-243. <https://doi.org/10.1109/TAC.1974.1100561>. URL <http://ieeexplore.ieee.org/document/1100561/>
- Gero, K. I., Ashktorab, Z., Dugan, C., Pan, Q., Johnson, J., Geyer, W., Ruiz, M., Miller, S., Millen, D. R., Campbell, M., Kumaravel, S., & Zhang, W. (2020) Mental Models of AI Agents in a Co-operative Game Setting. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM. ISBN 9781450367080, pp. 1-12. URL <https://dl.acm.org/doi/10.1145/3313831.3376316>
- Gibney, E. (2016) Google AI algorithm masters ancient game of Go. *Nature* 529(7587): 445-446. <https://doi.org/10.1038/529445a>. URL <http://www.nature.com/articles/529445a>
- Hancock, P. A. (2017) Imposing limits on autonomous systems. *Ergonomics* 60(2): 284-291. URL <https://www.tandfonline.com/doi/full/10.1080/00140139.2016.1190035>
- Hart, S. G., & Staveland, L. E. (1988) *Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research*. pp. 139-183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9). URL <https://linkinghub.elsevier.com/retrieve/pii/S016641150862386-9>
- Havrylov, S., & Titov, I. (2017) *Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols* URL <http://arxiv.org/abs/1705.11192>
- Hinds, P. J., Roberts, T. L., & Jones, H. (2004) Whose job is it anyway? A study of human-robot interaction in a collaborative task.
- Ho, Y., Bryson, A., & Baron, S. (1965) Differential games and optimal pursuit-evasion strategies. *IEEE Transactions on Automatic Control* 10(4): 385-389. <https://doi.org/10.1109/TAC.1965.1098197>. URL <http://ieeexplore.ieee.org/document/1098197/>
- Honig, S., & Oron-Gilad, T. (2018) Understanding and Resolving Failures in Human-Robot Interaction: Literature Review and Model Development. *Frontiers in Psychology* 9, 861, <https://doi.org/10.3389/fpsyg.2018.00861>. URL <https://www.frontiersin.org/article/10.3389/fpsyg.2018.00861/>
- Howell, B., Zaroukian, E., Asher, D., & Parker, L. (2021) *Identification of Emergent Collaborative Behaviors in Multi-Agent Systems*.

- Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019) *The Measurement of the Propensity to Trust Automation*. https://doi.org/10.1007/978-3-030-21565-1_32. URL <https://www.scopus.com/inward/record.uri?eid=2-s2>
- Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000) Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4(1): 53-71. https://doi.org/10.1207/S15327566IJCE0401_04
- Karau, S. J., & Williams, K. D. (1993) Social Loafing: A Meta-Analytic Review and Theoretical Integration. *Journal of Personality and Social Psychology* 65(4): 681-706. <https://doi.org/10.1037/0022-3514.65.4.681>
- Kolling, A., Walker, P., Chakraborty, N., Sycara, K., & Lewis, M. (2016) Human Interaction With Robot Swarms: A Survey. *IEEE Transactions on Human-Machine Systems* 46(1): 9-26. <https://doi.org/10.1109/THMS.2015.2480801>. URL <http://ieeexplore.ieee.org/document/7299280/>
- Korteling, J. E. H., van de Boer-Visschedijk, G. C., Blankendaal, R. A. M., Boonekamp, R. C., & Eikelboom, A. R. (2021) Human-versus Artificial Intelligence. *Frontiers in Artificial Intelligence* 4. <https://doi.org/10.3389/frai.2021.622364>. URL <https://www.frontiersin.org/articles/10.3389/frai.2021.622364/full>
- Kott, A. (2018) Intelligent Autonomous Agents are Key to Cyber Defense of the Future Army Networks. *The Cyber Defense Review* 3(3). URL <https://cyberdefensereview.army.mil/CDR-Content/Articles/Article-View/Article/1716477/>
- Langner, R., & Eickhoff, S. B. (2013) Sustaining attention to simple tasks: A meta-analytic review of the neural mechanisms of vigilant attention. *Psychological Bulletin* 139(4): 870-900. <https://doi.org/10.1037/a0030694>. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/a0030694>
- Latané, B., Williams, K., & Harkins, S. (2006) Many hands make light the work: The causes and consequences of social loafing. *Small Groups: Key Readings* 37(6): 297-308. <https://doi.org/10.4324/9780203647585>
- Lawson-Guidigbe, C., Louveton, N., Amokrane-Ferka, K., LeBlanc, B., & Andre, J. M. (2020) Impact of Visual Embodiment on Trust for a Self-driving Car Virtual Agent: A Survey Study and Design Recommendations. *Communications in Computer and Information Science* 1226: 382-389. https://doi.org/10.1007/978-3-030-50732-9_51
- Lazaridou, A., Peysakhovich, A., & Baroni, M. (2016) *Multi-Agent Cooperation and the Emergence of (Natural) Language* URL <http://arxiv.org/abs/1612.07182>
- Lee, J., Wickens, C., Liu, Y., & Boyle, L. (2017) *Designing for People: An Introduction to Human Factors Engineering*. 3rd edition. CreateSpace Independent Publishing Platform. ISBN 1539808009.
- Lee, J. D., & See, K. A. (2004) Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46(1): 50-80. https://doi.org/10.1518/hfes.46.1.50_30392. URL http://hfs.sagepub.com/cgi/doi/10.1518/hfes.46.1.50_30392
- Liang, C., Proft, J., Andersen, E., & Knepper, R. A. (2019) Implicit Communication of Actionable Information in Human-AI teams. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*: ACM. ISBN 9781450359702, pp. 1-13. <https://doi.org/10.1145/3290605.3300325>. URL <https://dl.acm.org/doi/10.1145/3290605.3300325>
- Lin, W., Qu, Z., & Simaan, M. A. (2011) A Design of Entrapment Strategies for the Distributed Pursuit-Evasion Game. *IFAC Proceedings Volumes* 44(1): 9334-9339. <https://doi.org/10.3182/20110828-6-IT-1002.00964>. URL <https://linkinghub.elsevier.com/retrieve/pii/S1474667016451116>
- Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017) *Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments* URL <http://arxiv.org/abs/1706.02275>
- Matsunaga, M. (2008) *Item Parceling in Structural Equation Modeling: A Primer*, volume 2. ISBN 1931245080245. <https://doi.org/10.1080/19312450802458935>
- McAuley, E., Duncan, T., & Tammen, V. V. (1989) Psychometric Properties of the Intrinsic Motivation Inventory in a Competitive Sport Setting: A Confirmatory Factor Analysis. *Research Quarterly for Exercise and Sport* 60(1): 48-58. <https://doi.org/10.1080/02701367.1989.10607413>. URL <http://www.tandfonline.com/doi/abs/10.1080/02701367.1989.10607413>
- McNeese, N. J., Demir, M., Chiou, E. K., & Cooke, N. J. (2021) Trust and Team Performance in Human-Autonomy Teaming. *International Journal of Electronic Commerce* 25(1): 51-72. <https://doi.org/10.1080/10864415.2021.1846854>. URL <https://www.tandfonline.com/doi/full/10.1080/10864415.2021.1846854>
- McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018) Teaming With a Synthetic Teammate: Insights into Human-Autonomy Teaming. *Human Factors* 60(2): 262-273. <https://doi.org/10.1177/0018720817743223>
- Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., & Shirase, L. (2019) Automation-induced complacency potential: Development and validation of a new scale. *Frontiers in Psychology* 10(FEB). <https://doi.org/10.3389/fpsyg.2019.00225>. URL <https://www.scopus.com/inward/record.uri?eid=2-s2>
- Muir, B., & Moray, N. (1996) Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39(3): 429-460. <https://doi.org/10.1080/00140139608964474>. URL <http://www.tandfonline.com/doi/abs/10.1080/00140139608964474>
- O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2020) Human-Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors (October)*. 64(5), 904-938, <https://doi.org/10.1177/0018720820960865>
- Parasuraman, R. (1997) Humans and Automation: Use, Misuse, Disuse, Abuse. In: *Technical Report 2*.
- Parasuraman, R., & Manzey, D. H. (2010) Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52(3): 381-410. <https://doi.org/10.1177/0018720810376055>. URL <http://journals.sagepub.com/doi/10.1177/0018720810376055>
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993) Performance Consequences of Automation-Induced 'Complacency'. *The International Journal of Aviation Psychology* 3(1): 1-23. https://doi.org/10.1207/s15327108ijap0301_1. URL http://www.tandfonline.com/doi/abs/10.1207/s15327108ijap0301_1
- Raykov, T. (1994) Studying Correlates and Predictors of Longitudinal Change Using Structural Equation Modeling. *Applied Psychological Measurement* 18(1): 63-77. <https://doi.org/10.1177/014662169401800106>
- Rodriguez, S. S., O'Donovan, J., Schaffer, J. A., & Hollerer, T. (2019) Knowledge Complacency and Decision Support Systems. In: 2019 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA). IEEE. ISBN 978-1-5386-9599-9, pp. 43-51. <https://doi.org/10.1109/COGSIMA.2019.8724175>. URL <https://ieeexplore.ieee.org/document/8724175/>
- Rosseel, Y. (2012) lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48(2). <https://doi.org/10.18637/jss.v048.i02>. URL <http://www.jstatsoft.org/v48/i02/>
- Ryan, R. M., & Deci, E. L. (2000) Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist* 55(1): 68-78. <https://doi.org/10.1037/0003-066X.55.1.68>. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/0003-066X.55.1.68>
- Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006) The Motivational Pull of Video Games: A Self-Determination Theory Approach. *Motivation and Emotion* 30(4): 344-360. <https://doi.org/10.1007/s11031-006-9051-8>. URL <http://link.springer.com/10.1007/s11031-006-9051-8>
- Schaffer, J., Humann, J., O'Donovan, J., & Höllerer, T. (2020) Quantitative Modeling of Dynamic Human-Agent Cognition. In: *Contemporary Research*. First edition: CRC Press, pp. 137-186. <https://doi.org/10.1201/9780429459733-7>. URL <https://www.taylorfrancis.com/books/9780429459733/chapters/10>
- Schaffer, J., O'Donovan, J., Marusch, L., Yu, M., Gonzalez, C., & Höllerer, T. (2018) A study of dynamic information display and

- decision-making in abstract trust games. *International Journal of Human-Computer Studies* 113: 1-14. <https://doi.org/10.1016/j.ijhcs.2018.01.002>. URL <https://linkinghub.elsevier.com/retrieve/pii/S107158191830002-8>
- Shepperd, J. A. (1993) Productivity loss in performance groups: A motivation analysis. *Psychological Bulletin* 113(1): 67-81. <https://doi.org/10.1037/0033-2909.113.1.67>
- Sheridan, T. B., & Verplank, W. L. (1978) Human and Computer Control of Undersea Teleoperators. In: *Technical report*, MASSACHUSETTS INST OF TECH CAMBRIDGE MAN-MACHINE SYSTEMS LAB
- Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018) A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362(6419): 1140-1144. <https://doi.org/10.1126/science.aar6404>. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.aar6404>
- Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008) Towards an Empirically Developed Scale for System Trust: Take Two. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52(19): 1335-1339. <https://doi.org/10.1177/154193120805201907>. URL <http://journals.sagepub.com/doi/10.1177/154193120805201907>
- Stumpf, S. (2016) Explanations Considered Harmful? User Interactions with Machine Learning Systems. *ACM SIGCHI Workshop on Human-Centered Machine Learning*. 1511-1524. <https://doi.org/10.1101/gad.377106>
- Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A. S., Yeo, M., Makhzani, A., Küttler, H., Agapiou, J., Schrittwieser, J., Quan, J., Gaffney, S., Petersen, S., Simonyan, K., Schaul, T., van Hasselt, H., Silver, D., Lillicrap, T., Calderone, K., Keet, P., Brunasso, A., Lawrence, D., Ekermo, A., Repp, J., & Tsing, R. (2017) *StarCraft II: A New Challenge for Reinforcement Learning* URL <http://arxiv.org/abs/1708.04782>
- Weintraub, I. E., Pachter, M., & Garcia, E. (2020) An Introduction to Pursuit-evasion Differential Games. arXiv URL <https://arxiv.org/abs/2003.05013>
- Wickens, C. D., & Dixon, S. R. (2007) The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science* 8(3): 201-212. <https://doi.org/10.1080/14639220500370105>. URL <http://www.tandfonline.com/doi/abs/10.1080/14639220500370105>
- Wright, J. L., Chen, J. Y., & Lakhmani, S. G. (2020) Agent Transparency and Reliability in Human-Robot Interaction: The Influence on User Confidence and Perceived Reliability. *IEEE Transactions on Human-Machine Systems* 50(3): 254-263. <https://doi.org/10.1109/THMS.2019.2925717>
- Zaroukian, E., Rodriguez, S., Barton, S., Schaffer, J., Perelman, B., Waytowich, N., Hoffman, B., & Asher, D. (2019) Algorithmically identifying strategies in multi-agent game-theoretic environments. In: T. Pham (ed.) *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, SPIE. ISBN 9781510626775, p. 38. <https://doi.org/10.1117/12.2518609>. URL <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11006/2518609/Algorithmically-identifying-strategies-in-multi-agent-game-theoretic-environments/10.1117/12.2518609.full>