Run Time Assurance and Human AI Fluency in Crewed Autonomous Intelligence Surveillance and Reconnaissance

Richard Agbeyibor*, Vedant Ruia†, Carmen Jimenez Cortes‡, Jack Kolb §
Georgia Institute of Technology, Atlanta GA 30332

Adan Vela¶
University of Central Florida, Orlando FL 32816

Samuel Coogan ‖ and Karen Feigh **
Georgia Institute of Technology, Atlanta GA 30332

The maturation of autonomy for electric vertical take-off and landing aircraft will soon make it possible to execute military intelligence, surveillance, reconnaissance (ISR) missions aboard crewed autonomous aerial vehicles. This research experimentally investigates factors that may influence the quality of interaction (i.e., team fluency) between a non-pilot human operator and the AI pilot responsible for autonomous flight, aboard a minimally crewed aircraft. In a flight simulator study with twenty-seven participants, various levels of workload and AI pilot capabilities are investigated including run time assurance through control barrier functions (CBFs). CBFs are used to enable pro-active collision avoidance behaviors by the AI pilot. Team fluency and mission effectiveness outcomes through trust, situation awareness, workload, interaction and performance show that task complexity and AI behavior are significant factors for the quality of human AI interaction in the autonomous ISR context.

I. Introduction

A. Motivation

The synchronous maturation of electric aircraft and autonomous aircraft technologies is opening up new applications and markets for new classes of vehicles collectively referred to as Advanced Air Mobility. The industry developing around Advanced Air Mobility promises to bring together autonomy and new forms of electric Vertical Take-Off and Landing (eVTOL) aircraft. This new class of aircraft would enable crew with little or no aviation training to operate aboard crewed autonomous eVTOLs. The U.S. military has expressed interest in using these new vehicles for specialized military aviation missions [1].

One such specialized military aviation mission – Intelligence, Surveillance, and Reconnaissance (ISR) – could leverage autonomous eVTOLs to reduce crew manning and training requirements while multiplying coverage. Given that the pace of commercial sector technology development in this emerging area far surpasses that of government R&D, many in the U.S. military are advocating for the adoption of off-the-shelf autonomous aerial vehicles for these military missions. Through personal communications with and interviews of U.S. Navy, AFWERX and DARPA personnel, ISR, medical evacuation, and cargo transportation were identified as missions that stand to benefit significantly from adoption of autonomous eVTOLs.

The ISR mission is split into crewed and uncrewed operations, with uncrewed operations being the predominant type. The uncrewed ISR mission is conducted using remotely piloted aircraft like the MQ-9. These vehicles are nominally operated through space-based satellite communication systems by a ground crew consisting of a pilot, a sensor operator, and an intelligence analyst. Crewed ISR is conducted with a similar crew composition to uncrewed ISR but usually

*Ph.D. Student, Robotics, School of Aerospace Engineering, AIAA Student Member.
†Research Assistant, School of Aerospace Engineering, AIAA Student Member.
‡Ph.D. Student, School of Electrical and Computer Engineering, AIAA Student Member.
§Ph.D. Student, Robotics, School of Aerospace Engineering, AIAA Student Member.
¶Associate Professor, Department of Industrial Engineering and Management Systems.
‖Associate Professor, School of Electrical and Computer Engineering.
**Professor and Associate Chair for Research, School of Aerospace Engineering.

on larger and more complex aircraft like the U.S. Navy's P-8 or the U.S. Air Force's RC-135. Crewed ISR aircraft usually require multiple pilots, multiple sensor operators and multiple intelligence analysts for missions that require high flexibility, versatility, sensitivity and immediate decision-making. It is a very costly mission to operate in terms of manning, training, and logistics. In the event of the unavailability of space-based satellite communication constellations used to remotely operate uncrewed ISR vehicles, it is possible that the U.S. military would send operators aboard eVTOLs to conduct distributed crewed autonomous ISR.

Autonomous aircraft capabilities could greatly reduce the manning requirements for ISR and increase operational coverage. With the piloting task in gross part delegated to an Artificial Intelligence (AI) pilot, the crew requirements could be significantly reduced for a single sortie. The same number of crew members could operate many more vehicles.

Autonomous uncrewed operations would not be acceptable if the humans on board face greater safety risk. Mechanisms to filter potentially unsafe primary controllers during the execution of the mission should be included. These strategies are commonly know as run time assurance (RTA) mechanisms. Introducing an RTA mechanism in the autonomy's capabilities would minimize risk and enforce safety throughout the mission. This study uses control barrier functions (CBFs) as the RTA strategy, as they enforce forward invariance of the constraint set so that no trajectory initialized within the constraint set ever leaves or violates the constraint set [2–5], i.e., never violates the safety specifications for the system.

B. Gaps

Collaboration is the process of two or more people, entities or agents working together to complete a task as a team. Fluent collaboration is the goal of any team, be it human-human or human-AI, as it leads to the best task and team outcomes [6].

Fluency is the "elusive yet palpable characteristic that exists when two agents collaborate at a high degree of coordination and adaptability, particularly when they are habituated to the work of one another" [7]. Fluency in collaboration has primarily been studied in the context of turn-by-turn manufacturing tasks which can be categorized according to Steiner's Taxonomy of Tasks as Divisible, Maximizing and Additive [8]. To date, there exists a gap in human-AI teaming research on collaboration and fluency in task contexts that are Unitary, Optimizing and Disjunctive.

There is emerging research interest in human AI teaming in autonomous vehicles. Motivated by the advent of self-driving cars, many researchers are now studying autonomy in cars. For decades, researchers in the field of aviation psychology have studied the evolution of automation in the cockpit, and many are now researching semi-autonomous aircraft operations.These researchers, however, are interested in how expert operators – licensed drivers and professional pilots – collaborate with autonomous technologies. At this time, there is no research in the literature studying specifically how non-operators aboard these vehicles interact with the autonomous agent driving or flying the vehicle.

In the context of military missions, the US military has funded research into human AI teaming for combat aviation [9], aerial refueling [10], cargo transport [11]– all of which are centered on expert pilots collaborating with the aircraft autonomy. Some of this research looked at supporting uncrewed ISR pilots with better autonomy, however, it still assumed the presence of well trained expert pilots [12].

This research is interested in addressing these gaps by researching factors that may affect the quality of interaction between a non-pilot human and an AI pilot dyad collaborating to accomplish a specialized mission aboard an autonomous aerial vehicle. Specifically, this study investigates the ISR mission and how it could be accomplished by non-pilots aboard autonomous aerial vehicles.

C. Research Questions & Hypotheses

In the context of ISR operators who are not trained in piloting or AI programming, collaborating with an AI pilot agent aboard a crewed autonomous aerial vehicle to accomplish a maritime ISR mission, how does task complexity and AI behavior affect team fluency?

Research Questions To address these gaps, the following questions are posed:

- RQ1**:** How do changes in task complexity affect situation awareness, workload and mission effectiveness?
- RQ2: How do various autonomy behaviors such as CBF-enabled run time assurance, affect fluency components - situation awareness, perceived performance, interaction and workload?
- RQ3: How do these fluency components — trust, situation awareness, perceived performance, interaction and workload — affect mission effectiveness?

Hypotheses The authors propose three hypotheses regarding the interaction between task complexity and team fluency, and their impact on aspects of mission effectiveness.

- H1: An increase in task complexity will decrease situation awareness, increase workload, and decrease mission effectiveness.
- H2a: Levels of autonomy that increase decision support, such as CBF-enabled run time assurance, will decrease workload and perceived performance.
- H2b: Levels of autonomy that share decision authority without transparency will increase workload, decrease situation awareness, and perceived performance.
- H3: A decrease in fluency,indicated by an increase in trust, situation awareness, and perceived performance, along with a decrease in interaction and workload, will decrease mission effectiveness.

II. Background

The assessment of fluency in human-robot collaboration was presented, mapped, and validated by Hoffman [13]. Its three main objectives were to give an archived list of subjective and objective fluency measurements, offer a preliminary theoretical analysis of those metrics, and systematically look into the correlation between the two. Objective measures quantitatively evaluate the level of fluency through various metrics in a particular interaction, and subjective measures gauge people's perceptions of the fluency of an interaction and associated features of the robot. In their preceding papers [7, 14, 15], Hoffman and Brazeal hypothesize that the fundamental key to achieving fluency may reside in the use of intelligent anticipatory action based on anticipation of one another's behavior.

Unhelkar et al. [16] offer a single human-aware robotic system that can anticipate human action and plan ahead to carry out effective and secure motions throughout the final assembly of an automobile. They compare the performance of their system in a simulation to three other approaches, including a baseline strategy that simulates the behavior of common safety systems used in factories along with planning with detection, and planning with prediction.

Gombolay et al. [17] investigated the effects of a robot worker's authority and capabilities on human worker's perception of the robot and their desire to work with it again in the future. Romat et al. [18] evaluate human-robot interaction using affordances and social cues as metrics. Affordances are attributes and characteristics of an object that determine its potential uses and suggest how it should be utilized.

In live-flight dog-fight experiments in fighter aircraft, the University of Iowa Operator Performance Laboratory used eye tracking, electrocardiogram (ECG) and other physiological measures to assess operator trust in an autonomous pilot agent. Highland, Schnell et al. [19] compared physiological measures of trust to self-reported measures and concluded that there is utility in a real-time machine learning framework. Napoli et al [20] warn of challenges with Machine Learning (ML) classification of cognitive states through physiological measures due to ambiguous ground truth, low samples, subject-to-subject variability, class imbalances, and wide data sets. They recommend a Naïve Adaptive Probabilistic Sensor ML framework to overcome these experimental data concerns.

Turning to more subjective measures, Paliga and Polak [21] devised a six-item evaluation to examine the subjective human-robot fluency from the human-oriented, robot-oriented, and team-oriented viewpoints. These included: trust in robot, robot's contribution / commitment, robot's performance, positive teammate traits, bond subscale and Working Alliance for HRI.

Schneider et al. [12] explore the impact of coordination, communication and intent on human-AI teams in the context of crewed ISR missions. This and the other works mentioned above inform the design of this study. Technical details of the run time assurance mechanism of the AI pilot's Collision Avoidance behavior are precented in [22].

III. Methods

The research questions were assessed via empirical analysis. A simulated cabin of a futuristic eVTOL aircraft similar to the CMV-22 Osprey was created, and 27 participants volunteered to complete an ISR task while interacting with the AI pilot over a series of scenarios which varied in task load and complexity.

A. Operationalization of HAI Fluency

Fluency in collaboration is assessed through subjective and objective measures. In the context of humans operating aboard an ISR autonomous aerial vehicle, the authors of this study define fluency as the combination of trust, situation awareness, workload, perceived performance, and interaction.

Trust is a complex construct spanning multiple categorizations. The trust construct of interest in this research is human-autonomy trust and it is measured through the elements of affective and capability-based trust [23]. A four question questionnaire rated on a 5-point Likert scale asks the participants to rate their perception of the AI's cooperation and support, as well as its contributions to their decision-making.

Situation awareness (SA) is defined as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" [24]. This study measures SA immediately post-scenario using one question out of a bank of questions that assess the participant's perception and comprehension of elements within the scenario such as the number, type, and position of enemy ships as well as the health of their aircraft.

The NASA Task Load Index (TLX) is a multi-dimensional rating procedure that assesses workload based on a weighted average of ratings on six subscales: mental demand, physical demand, temporal demand, effort, frustration and performance [25]. Unlike the other five which directly assess the affect of the participant, the performance subscale of the NASA TLX asks the participant how successful they think they were in accomplishing the goals of the task. In addition to workload, this study considers the performance subscale of the NASA TLX as a measure of perceived performance.

Perceived performance aims to capture the subjective perception of the human, the robot, and the team as perceived by the human operator. This study uses a six question questionnaire rated on a 5-point Likert scale to assess perceived performance, in addition to the performance subscale from the NASA TLX.

Interaction is the final element of team fluency as defined for this context. The interaction element captures the communication and action events between the human and the AI pilot through user interface logs. The user interface logs capture the alerts and notifications from the AI pilot and categorizes them by type and number. The user interface also logs user mouse clicks to measure the location and number of user waypoint inputs.

B. Participants

The study was approved by the Georgia Institute of Technology Institutional Review Board. Participants were recruited from the broader Atlanta area and the university community. Participation was limited to English speakers, between 18 and 65 years old, who are not color blind, and do not have a pacemaker or similar heart rate stabilization devices. The study lasted approximately two and a half hours. Participants were compensated $50.

Thirty participants completed the study. Data from three participants was discarded from the statistical analysis for incompleteness. Out of the remaining 27 participants, 19 were male, 6 female and 2 non-binary. Twenty-one had no AI experience, one was self-taught, one had undergraduate level coursework, and four had graduate level coursework in AI programming. Twenty-two participants had no flight experience, four had some (less than 10 hours), and one was licensed with 130 flight hours.

C. Experimental Apparatus

1. ISR Wargame

The SIR Wargame is an ISR mission simulator developed specifically for this research. The scenarios employed were developed in collaboration with an operational Navy P-8 Poseidon pilot. The graphical user interface for the ISR Operator Control Station is shown in Fig. 1a. It is presented on a 20 inch display inside the cabin and a computer mouse as the primary user input.

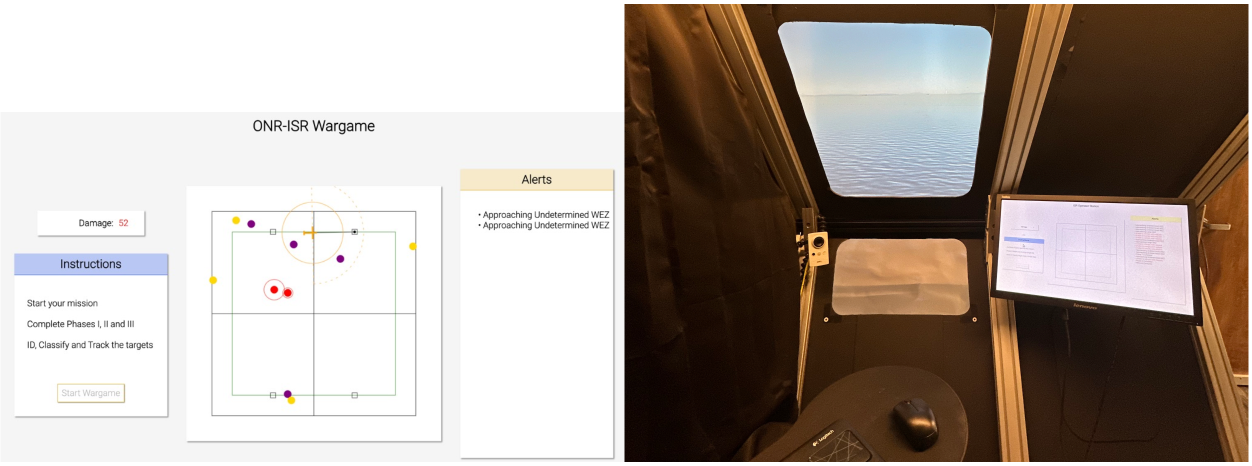
The behaviors of the AI pilot were designed to present various combinations of authority, responsibility and team dynamics. Analysis of the joint ISR task and the human-AI team construct are presented in [26].

2. Experiment Cabin

The experiment takes place inside of a cabin designed to emulate a smaller eVTOL version of an ISR vehicle like the CMV-22 Osprey currently used by the U.S. Navy for maritime ISR. An immersive experience is created through a projected view of Microsoft Flight Simulator™ where a CMV-22 Osprey is modeled and the user is flying off the coast of San Diego, CA. Users have a display to the right showing the graphical user interface (GUI) of the AI Agent and two large windows in front showing a view simulating the windshield and the vertical reference windows of the vehicle. Figure 1b shows the cabin from the participant's perspective.

3. Implementation of Control Barrier Functions in Simulator

CBFs were implemented using a second order unicycle model to simulate the aircraft. This CBF-enabled controller took as inputs the locations of all target ships in the simulated scenario and built barriers around each one of them based



Start your mission
Complete Phases I, II and III
ID, Classify and Track the targets

• Approaching Undetermined WEZ
• Approaching Undetermined WEZ

(a) ISR Operator Control Station Graphical User Interface **(b) Simulator Cabin from the Participant's Perspective**

Fig. 1 Simulator GUI and Cabin

on their latitude, longitude, and parameterized Weapon Employment Zone (WEZ) size. This CBF-controller was used as a run time assurance mechanism for collision avoidance for the simulated autonomous eVTOL.

The nominal trajectory of the aircraft was defined by the system's current latitude and longitude and the user's input waypoint. The system followed the shortest path to the target waypoint using a proportional-derivative controller. If during the navigation the nominal trajectory required the aircraft to fly over the enemies' WEZs, the CBF mechanism modified the nominal controller so that such collision could be avoided. By using this new safe-controller, the aircraft was able to circumnavigate the designated radius around the obstacle according to its WEZ size.

D. Task Design

The ISR simulator used enables the researcher to configure the number of targets, the speed of the targets, the search pattern, AI behavior mode, and the aircraft controller.

The AI can operate in four different modes or behaviors:

- **Waypoint Behavior (Level 0)** - The AI's baseline Waypoint behavior allows the AI to fly an automated search pattern. At any time, the user can override the AI's automated search pattern waypoint by clicking on a point on the screen to cast a vector to a new user designated waypoint.
- **Collaborative Behavior (Level 1**) - In the collaborative behavior, the AI flies a search pattern like in the baseline behavior. Unlike the baseline where user vectors are immediately executed, in this behavior, the user requests a waypoint when they click on a point in the surveillance area. The user's waypoint request is processed by the AI and when able, the AI executes the user's request. When unable to comply with the user request, the AI displays an alert and continues its automated search pattern.
- **Collision Avoidance (Level 2)** - In the collision avoidance behavior, the AI flies its search pattern waypoints while proactively avoiding targets on its path. The AI shows its planned collision avoidance path with green breadcrumbs. At any time, the user can override the AI's planned path by clicking on a point on the screen to cast a user waypoint vector. The AI flies to the user's waypoint while avoiding targets on its path.
- **Search Optimization (Level 3)** - In the search optimization behavior, The AI flies starts off flying its default search pattern and at any time, the user can request an optimization of the search pattern. The AI suggests an optimized search pattern that the user can accept, re-optimize, or deny. If the user accepts the suggested optimization, the AI starts flying the new search pattern. The user can cancel the optimized search pattern and return to the default at any time. At any time and with any search pattern, the user can override the AI's waypoint by clicking on a point on the screen to cast a user vector.

There are two task load levels in the experimental design: a Low level (Level A) and a High level (Level B). The Low task load level (Level A) has 10 targets, moving at 5 knots, with a simplified square search pattern. The High task load level (Level B) has 20 targets, moving at 15 knots, with a more complex ladder-style search pattern. The scenario factors and levels are summarized in Table 1.

AI Behavior	**Waypoint -**	**Collaborative -**	**Collision Avoidance -**	**Search Optimization -**
Task Load	**Level 0**	**Level 1**	**Level 2**	**Level 3**
Low - Level A	A0	A1	A2	A3
High - Level B	B0	B1	B2	B3

Table 1 Scenario Factors and Levels

E. Design of Experiments

A full-factorial, repeated measures, within-subjects design of experiments is employed. Each participant completed all combinations of task load and AI behavior factor levels shown in Table 1.

Fatigue and learning can have significant carry-over effects. To minimize those effects, a Latin Square design is utilized to counterbalance the order of scenarios that are presented to participants. A balanced Latin Square generator tool by Damien Masson at the University of Waterloo was used, following James Bradley's method [27] for complete counterbalancing. Balanced Latin Squares are special cases of Latin Squares which remove immediate carry-over effect in which a scenario will precede another exactly.

F. Control Barrier Functions for Collision Avoidance - AI Behavior Level 2

Control Barrier Functions were utilized as a collision avoidance mechanism in the pro-active AI behavior, Level 2. This section describes the foundations of CBFs as well as their implementation in the ISR mission.

Suppose the constraint set S is defined as $S = \{x \mid h(x) \geq 0\}$ for some continuously differentiable function $h(x)$. The boundary of S, denoted $\partial S = S \backslash \text{int}(S)$, is given by $\partial S = \{x \mid h(x) = 0\}$. $h(x)$ called a *Barrier Function*, and satisfies that $h(x) = 0$ implies $\nabla h(x) \neq 0$. The set S is *forward invariant* for the system $\dot{x} = f(x)$, with state vector $x \in \mathbb{R}^n$ if, for any initial condition within the set S, the system remains inside S for all time $t \geq 0$ [28]. If, further, f is Lipschitz continuous, it holds that

$$S \text{ is forward invariant} \iff \dot{h}(x) := \nabla h(x)^T f(x) \geq 0 \text{ for all } x \in \partial S, \tag{1}$$

which is classically known as Nagumo's Theorem. In the barrier function literature, the righthand condition is often strengthened to

$$\dot{h}(x) \geq -\alpha(h(x)) \quad \text{for all } x \in \mathbb{R}^n \tag{2}$$

for some locally Lipschitz function $\alpha : \mathbb{R} \to \mathbb{R}$ satisfying $\alpha(0) = 0$. Even though this new condition must hold for all x rather than only on ∂S, it more readily leads to control design techniques. For example, consider the controlled system $\dot{x} = f(x) + g(x)u$, now with input $u \in \mathbb{R}^m$, and the goal of designing a feedback controller $\sigma(x)$ such that S is forward invariant. Then, condition (2) leads to the design criterion that any Lipschitz continuous feedback controller $\sigma(x)$ satisfying $\sigma(x) \in U(x)$ where

$$U(x) := \{u \mid \nabla h(x)^T (f(x) + g(x)u) \geq -\alpha(h(x))\} \tag{3}$$

ensures forward invariance of S. The inequality in (3) is affine in u, which allows to include it in an optimization program to compute the feedback controller $\sigma(x)$ during run time. If it is possible to find this feedback controller, then $h(x)$ is called a *Control Barrier Function (CBF)*.

The use of CBFs for collision avoidance in the ISR mission is modeled as follows. Consider that each enemy target has a Weapon Engagement Zone (WEZ) characterized by a radius D^j_{WEZ}. Assume that each target also possesses a *proximity function* $d^j(x)$ characterizing a distance between the aerial vehicle position x and the enemy target z^j. For example, a choice for d^j could be $d^j(x) = \|x - z^j\|_2^2$. Then the system must satisfy the following safety constraint: the distance to all enemy targets should not be less than its WEZ, i.e., $d^j(x(t)) \geq D^j_{WEZ}$ for all t. To implement this collision avoidance mechanism, the following optimization problem needs to be solved at each time t:

$$\begin{aligned} &\underset{u}{\text{minimize}} \quad ||u - \hat{u}||^2 \\ s.t. \quad &\dot{h}_j \geq -\alpha_j(h_j) \quad \forall j = 1, ..., N. \end{aligned} \tag{4}$$

The nominal control strategy to command the aircraft is given by $\hat{u}$, whereas the safe controller, i.e., guaranteeing avoidance of all WEZs, is u. N is the total number of enemy targets and h_j is defined as:

$$h_j(x) = \|x - z^j\|_2^2 - (D^j_{WEZ})^2. \tag{5}$$

If the quadratic program (4) is feasible for all time t then u guarantees collision avoidance and the constraint set S is forward invariant [29, Thm. 2].

G. Data Collection

1. Demographic Data

Participants were asked to complete a pre-experiment demographics questionnaire. Within this questionnaire participants were asked to provide: age, gender, flight experience, and AI experience.

Participants indicated their AI experience between the five levels of: no experience, online course, undergraduate course, graduate course, or other. These data serve as cofounding variables within the data analysis.

Participants were also to indicate flight experience between None, Some, and Licensed and provide flight hours if experienced.

2. User-Interface Log Data

Various data were collected through user-interface logs. The number of user waypoint inputs and the number of AI alerts are recorded to understand the interaction between the user and AI. The user interface also logs the damage score and the mission duration from the timer and score displays at the top left corner of the GUI. The aircraft damage score and mission duration are used to determine the user's objective mission performance.

3. NASA Task Load Index (TLX) Data

Users completed a modified NASA TLX scale after each scenario of the experiment. Instead of the original 21 increments on paper, users completed the TLX on a sliding scale from 0-100 on a computer-based questionnaire. The performance subscale axis was reversed per the original NASA TLX scale design [25]. This data provided insight on the workload experienced by participants.

4. Situation Awareness Questionnaire Data

In order to gauge the situation awareness of participants, they were asked one multiple choice or free-response question about a specific attribute of their scenario. Examples of this include questions about how many of a certain target appeared, which quadrant of the surveillance area was the most populated, or what the aircraft damage score was at the end of a scenario. Questions were graded on a pass/fail criterion according a small margin of error for free-response questions.

5. Physiological Metrics

A BIOPAC Acqknowledge ECG was used to measure the heart rate variability (HRV) of each participant. The BIOPAC Acqknowledge HRV multi-epoch statistical analysis tool is used to calculate the mean root mean square of successive RR Interval differences. The root mean square of successive differences between normal heartbeats (RMSSD) is obtained by first calculating each successive time difference between heartbeats in milliseconds (ms). Then, each of the values is squared and the result is averaged before the square root of the total is obtained. The RMSSD reflects the beat-to-beat variance in HR and is the primary time-domain measure used to estimate the vagally mediated changes reflected in HRV. RMSSD reflects parasympathetic HR modulation. RMSSD characterizes short-term rapid changes in heart rate, which can only occur under the influence of the parasympathetic nervous system. When RMSSD ≤ 0.068, the heart rhythm is normal [30].

6. Post-Experiment Questionnaire Data

A post-experiment questionnaire is administered after a participant has completed all eight scenarios. The questionnaire is designed to understand the user's overall trust, communication, and perceived performance throughout the whole experiment. This questionnaire is based on the Interdependent Trust for Humans and Automation (I-THAu) Survey [31].

IV. Results and Analysis

This section presents results and analysis organized by research question. The analysis was conducted in R using generalized linear models (GLM) and mixed effects models. The mixed effects modelling used the listed independent variables as fixed effects and the listed confounding variables as random effects.

A. RQ1: Task Complexity vs. Fluency

To answer RQ1, the following data were analyzed using generalized linear models and mixed effects models:

- Independent Variables: Task Load, AI Behavior
- Dependent Variables:
  - Situation Awareness
  - Workload TLX: Mental Demand, Physical Demand, Temporal Demand, Effort, Frustration, Performance
  - Workload Physiology: Mean HR RMSSD
  - Mission effectiveness: Mission Duration, Damage

1. Situation Awareness

Figure 2 shows how situation awareness (SA) changes with task load throughout the 216 trials. In the 108 low task load scenario trials, participants had good SA in a higher number of trials than bad SA. In high task load scenario trials, there are also more trials with good SA than bad SA but the difference is smaller. Calculating the mean SA yields a mean of 0.79 under low task load and a mean of 0.61 under high task load.

A Chi-squared test is used to test for independence between situation awareness and task load. The Chi-squared test revealed a significant relationship ($\chi^2 = 6.786$, $p = 0.009$).

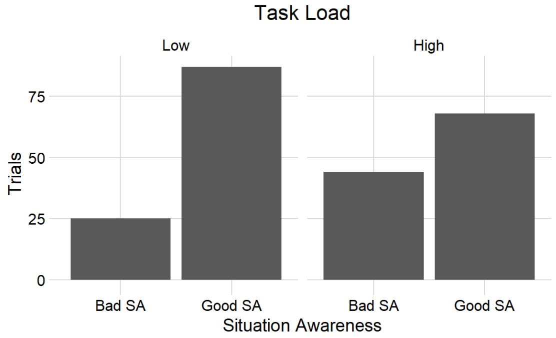


Fig. 2 Situation Awareness vs Task Load

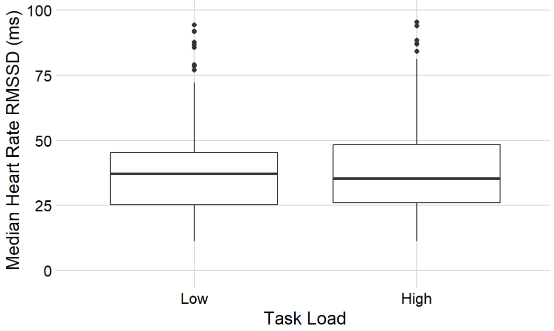


Fig. 3 Heart Rate Variability vs Task Load

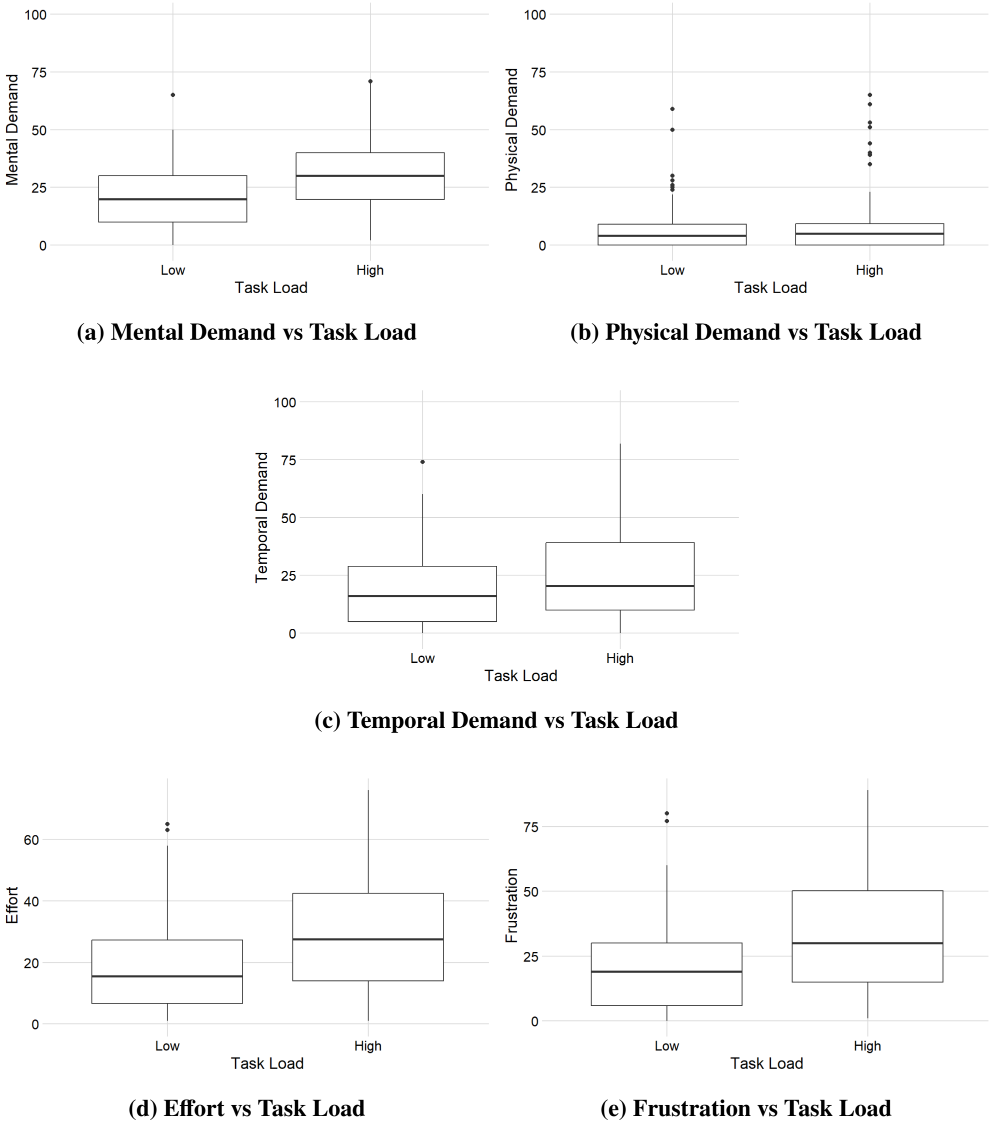
Since the SA question is graded pass / fail or good / bad, a binomial GLM is used to model its error distribution. An analysis of variance of a binomial GLM with task load as predictor for SA finds a highly significant difference between the two levels of task load with $p = 0.005$.

Analysis of the interaction between task load and AI behavior found two instances of significant interaction. When the task load is low, there is a significant difference in SA between the Search Optimization behavior and the default Waypoint behavior ($p = 0.0499$). The Search Optimization behavior requires that the operator review and approve / disapprove the AI's suggestions.

When the task load is high, there is a significant difference in SA between the Collision Avoidance behavior and the default Waypoint behavior ($p = 0.014$). The Collision Avoidance behavior offloads the task of avoiding enemy WEZs from the operator to the AI pilot.

2. Workload TLX

Figure 4 shows how NASA TLX subscales mental demand, physical demand, temporal demand, effort and frustration change with task load. The data are approximately normally distributed with a right skew. A linear mixed effects model is used to model the data with participant ID as a random effect. An analysis of variance indicates very highly significant differences in workload between the levels of task load, with the exception of physical demand. In the flight simulator environment, physical demand was not expected to vary. Table 2 shows the p-values obtained for all six NASA TLX subscales.



(a) Mental Demand vs Task Load

(b) Physical Demand vs Task Load

(c) Temporal Demand vs Task Load

(d) Effort vs Task Load

(e) Frustration vs Task Load

Fig. 4 NASA TLX subscales vs Task Load

In line with hypothesis H1, participants felt a higher mental demand with the higher task load. With a higher task load there is a marginally higher amount of perceived physical demand. Users felt a higher temporal demand i.e. time-pressure with an increase in task load. There was a higher amount of perceived effort as task load increased, as expected. Participants felt more frustration with an increase in the task load. Participants also perceived lower performance with an increase in task load.

TLX Subscale	Δ Low to High	p-value
Mental Demand	9.9	p < 0.0001***
Physical Demand	1.7	p = 0.018*
Temporal Demand	7.0	p = 0.00006***
Effort	11.5	p < 0.0001***
Frustration	13.1	p < 0.0001***
Performance	-16.4	p < 0.0001***

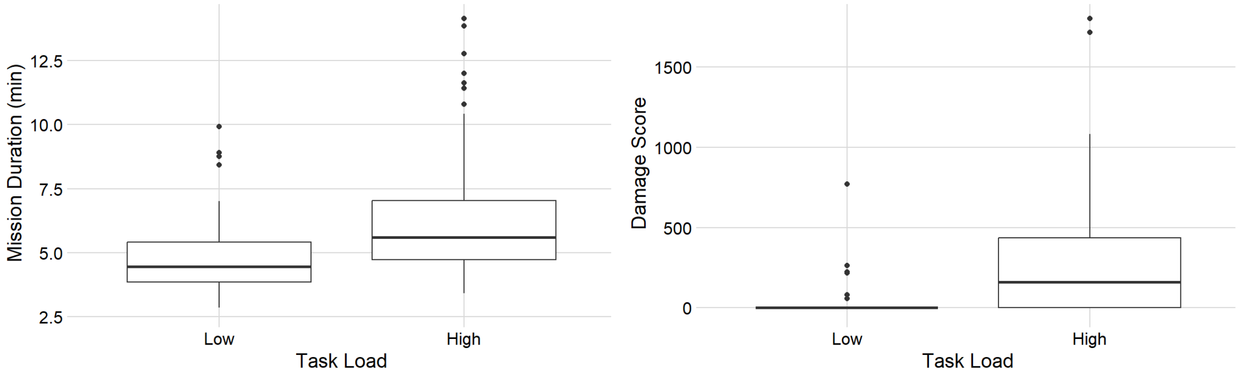
Table 2 Variance in Workload per Task Load

3. Workload Physiology

Figure 3 shows how Heart Rate Variability changes with task load. There was not a clear relationship between heart rate variability and task load. The Heart Rate Variability data did not support hypothesis H1.

4. Mission Effectiveness

Figure 5a shows how mission duration changes with task load. With a higher task load, mission duration increased. A linear mixed effects model best fit the data with heart rate variability as the response to task load as a fixed variable and participant ID as a random variable. Analysis of variance yielded a very highly significant difference between low and high task load with $p < 0.0001$.



(a) Mission Duration vs Task Load

(b) Aircraft Damage vs Task Load

Fig. 5 Mission Effectiveness vs Task Load

Figure 5b shows how aircraft damage changes with task load. The aircraft damage score increased with a higher task load. There was minimal damage in most of the lower task load scenarios. Analysis of variance of a linear mixed effects model fit to the data with damage as a fixed variable and participant ID as a random variable, found a very highly significant difference with $p < 0.0001$.

B. RQ2: AI Behavior vs. Fluency

To answer RQ2, the following data were analyzed using generalized linear models and mixed effects models:

- Independent Variables: AI Behavior
- Dependent Variables:
  - Situation Awareness: Situation Awareness
  - Perceived Performance: TLX Performance
  - Interaction: Number of User Waypoints
  - Workload TLX: Mental Demand, Physical Demand, Temporal Demand, Effort, Frustration
  - Workload Physiology: Heart Rate Variability RMSSD

1. Situation Awareness

Figure 6 shows how participant situation awareness changes with AI behavior. There are large differences in participant situation awareness across the AI behaviors in the low and high task load scenarios. In the low task load scenarios, the highest situation awareness passing scores were in the collaborative behavior. In the high task load scenarios, the search optimization mode yielded the highest situation awareness.

A Pearson's Chi-squared test was used to test for independence between situation awareness and task load. The Chi-squared test revealed a significant relationship ($\chi^2 = 9.6135$, $p = 0.02$). The relationship between situation awareness and AI behavior varied depending on the task load.

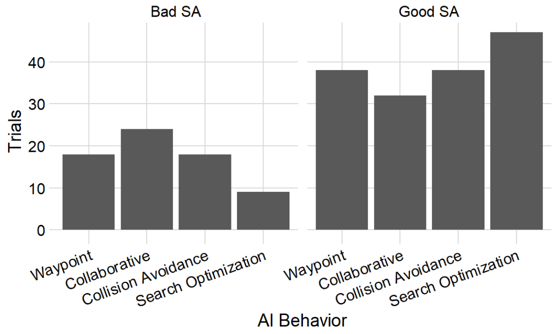


Fig. 6 Situation Awareness vs AI Behavior

TLX Subscale	p-value
Mental Demand	p > 0.05
Physical Demand	p > 0.05
Temporal Demand	p > 0.05
Effort	p > 0.05
Frustration	p = 0.0005**
Performance	p = 0.0019*

Fig. 7 Variance in Workload per AI Behavior

2. Perceived Performance

Figure 8 shows how the performance subscale of the NASA TLX changes with AI behavior. There are significant differences in participant self-reported performance ratings across the AI behaviors. On average, the collaborative AI behavior yielded the highest performance ratings and the waypoint behavior had the lowest performance rating. The data was fit with a linear mixed effects model for NASA TLX performance with AI behavior as a fixed effect and participant ID as a random effect. Analysis of variance found a very highly significant difference ($p = 0.0019$).

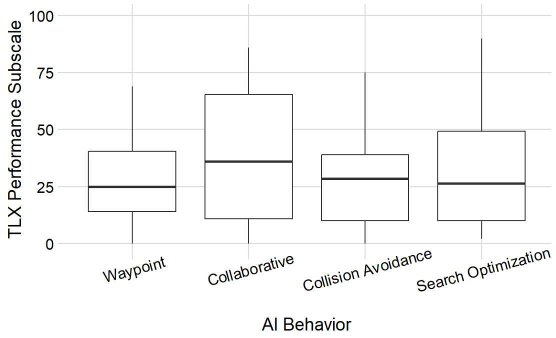


Fig. 8 TLX Performance Subscale vs AI Behavior

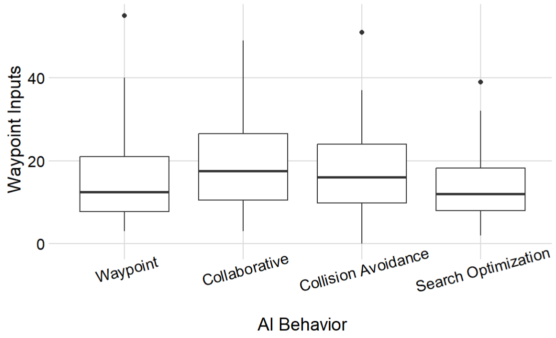


Fig. 9 Number of Waypoint Inputs vs AI Behavior

3. Interaction

Figure 9 shows how the number of user waypoint inputs change with AI behavior. On average, users clicked to cast waypoint vectors the most in the collaborative behavior and the least in the waypoint and search optimization mode.

A log-linear model was determined to best fit the data through comparison of AIC scores with a linear model, a Poisson GLM, and a negative binomial GLM. Analysis of variance of the log-linear model found a marginally significant difference ($p = 0.05066$).

4. Workload TLX

Figure 10 shows how mental demand, physical demand, temporal demand, effort and frustration change with AI Behavior. Table 7 summarizes the p-values. The frustration subscale showed a highly significant difference and the performance subscale showed a significant difference.

The collaborative and search optimization modes tended to have higher mental demand ratings than the other AI behaviors. The collision avoidance behavior has the lowest amount mental demand ratings on average. For physical demand, the highest variances in the data are observed in the search optimization AI behavior.

There is no significant difference in temporal demand (time-pressure) ratings amongst AI behaviors. Neither are there any significant relationships between behavior and effort, however, the largest variance in effort ratings is found in the collaborative behavior.

Figure 10e shows how frustration changes with AI behavior. Participants experienced a noticeably higher level of frustration with the collaborative AI behavior compared to the other behaviors. By a very slight margin the waypoint behavior had the lowest average frustration rating.

5. Workload Physiology

There were no significant relationships in the collected physiology data between heart rate variability and AI behavior.

C. RQ3: Fluency vs. Mission Effectiveness

To answer RQ3, the following data were analyzed using generalized linear models. There were differences in the average duration and score for each of the scenarios due to the differences in task load and AI behavior. To evaluate fluency and mission effectiveness across all scenarios, scenario results were summed for each participant:

- Independent Variables: Fluency
  - Trust: I-THAu Trust, Communication and Interaction
  - Situation Awareness: Situation Awareness
  - Perceived Performance: I-THAu Team Performance, TLX Performance
  - Interaction: Number of User Waypoints
  - Workload TLX: Mental Demand, Physical Demand, Temporal Demand, Effort, Frustration
  - Workload Physiology: Heart Rate Variability RMSSD
- Dependent Variables: Mission effectiveness– Mission Duration, Damage

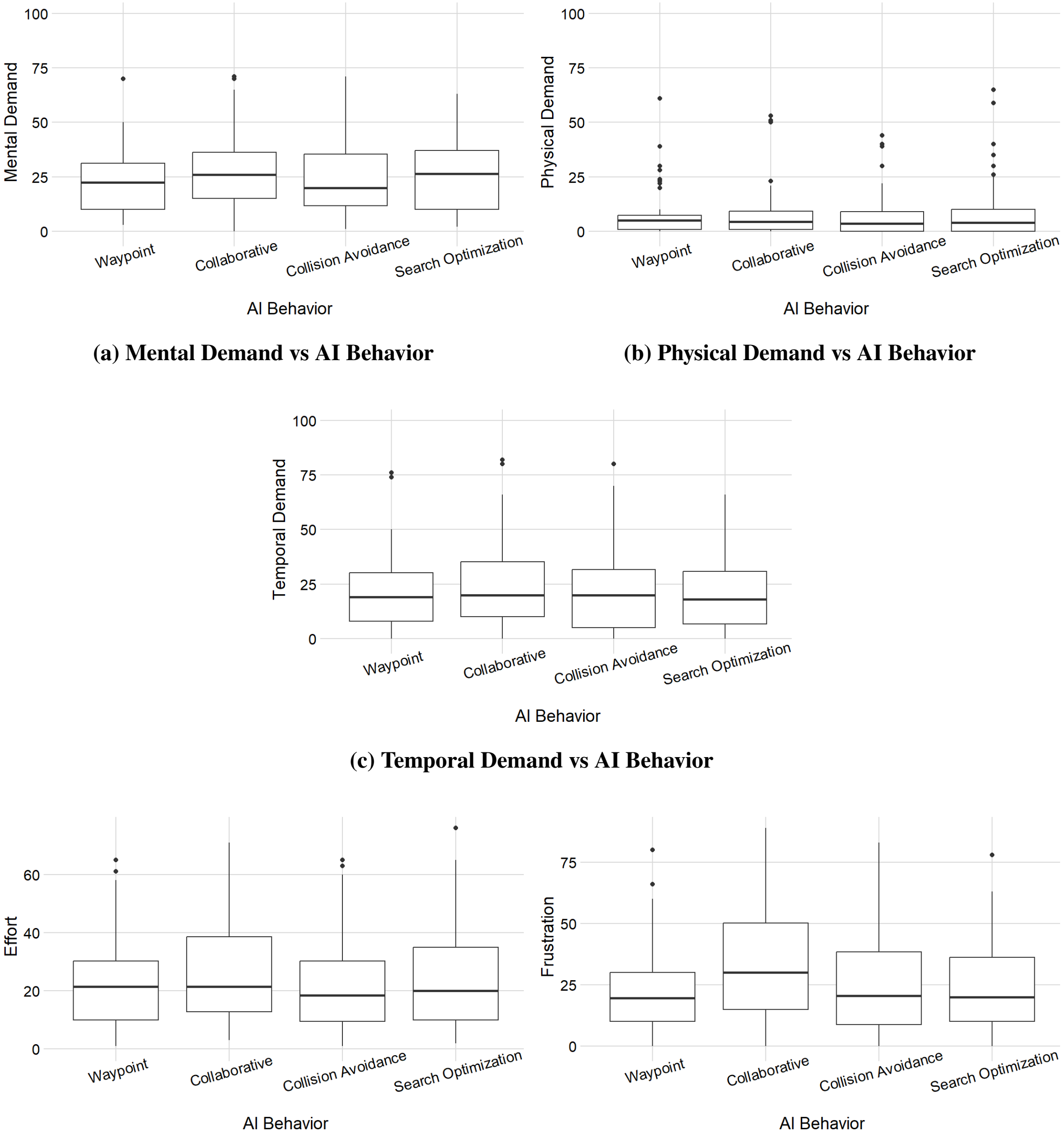
1. Trust, Situation Awareness, Perceived Performance, Interaction

No remarkable trends or statistical significance were found in the data for the fluency components of trust, situation awareness, perceived performance and interaction when analyzing their relationship with mission effectiveness.

2. Workload TLX

There were differences in the average duration and score for each of the scenarios due to the task load and the AI behavior. To evaluate fluency and mission effectiveness across all scenarios, NASA TLX ratings were summed for each participant to obtain total workload and total mission duration.

There are positive correlations between the TLX sub scales for mental (Fig. 11a, $r = 0.37$), physical (Fig. 11b, $r = 0.29$) and temporal demand (Fig. 11c, $r = 0.28$) with total mission duration. Mental demand exhibited a moderate positive correlation with total mission duration, indicating that higher levels of mental demand are associated with longer mission durations. This correlation is statistically significant ($p < 0.001$), suggesting a meaningful relationship between mental demand and mission duration.



(a) Mental Demand vs AI Behavior

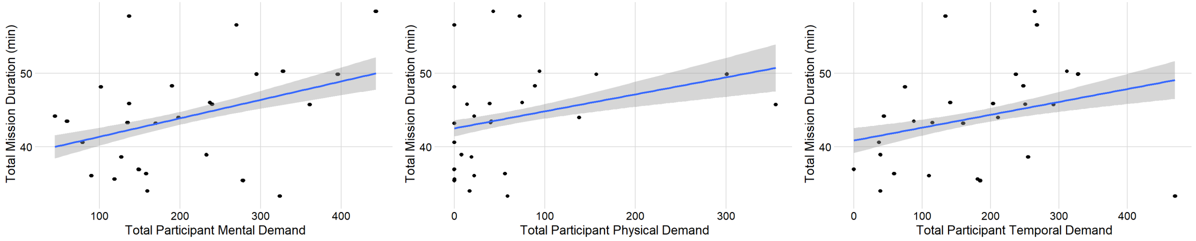
(b) Physical Demand vs AI Behavior

(c) Temporal Demand vs AI Behavior

(d) Effort vs AI Behavior

(e) Frustration vs AI Behavior

Fig. 10 TLX vs AI Behavior



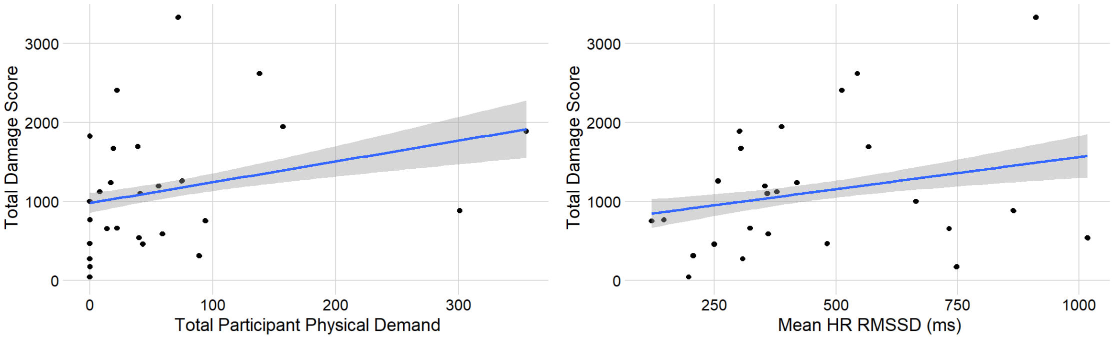
(a) Mental Demand

(b) Physical Demand

(c) Temporal Demand

Fig. 11 Total Mission Duration vs Workload

There is also a positive correlation between damage and perceived physical demand from the TLX scale. This relationship is in Fig. 12a but it is not statistically significant.



(a) Total Damage vs Physical Demand

(b) Total Damage vs Heart Rate Variability

Fig. 12 Total Damage vs Workload

3. Workload Physiology

There is an approximate association between the total damage score and the heart rate variability of participants. The Pearson correlation coefficient between the total damage score and heart rate variability (measured as RMSSD) is $r = 0.24$, indicating a weak positive linear relationship. This suggests that as heart rate variability increases, the total damage score tends to increase slightly as well. This relationship is plotted in Fig.12b but it is also not statistically significant.

V. Discussion

In the HAI literature, fluency is often defined in correlation with performance and time metrics for human-AI teammates collaborating on turn-by-turn tasks. The context of human-AI teammates collaborating in an autonomous aerial vehicles requires the accomplishment of tasks that are not strictly divisible, maximizing or additive [8]. This context required broadening of the definition of fluency to include workload, trust, situation awareness and interaction in addition to performance.

A. Task Complexity vs. Fluency

The exploration of the task complexity vs. workload served as a validation of experimental design in addition to studying the relationship between the workload of non-pilot human and the resulting fluency with the AI pilot.

With an increase of task load and complexity in the form of presenting the user more targets to surveil and a more complicated flight pattern to oversee, there was a significant increase in all levels of workload. This is evident by looking at the user's responses to the NASA TLX sub scales after all of the higher task load scenarios compared to the lower task load scenarios. This was not reflected as well in the physiological data, but there are general trends within the deviation of the heart rate that indicate the potential for classification using a neural network. These findings validate the hypothesis that an increase in task load would increase the overall workload of the user, which is a major component of fluency. With a higher workload, the demand for human AI team fluency rises.

There was also a statistically significant difference in the participant's situation awareness with a variation in task load and complexity. The higher the task load, the lower the user's situation awareness was. This is likely due to the increase in workload since the user's cognitive resources are more solicited and as such they have less bandwidth for overall situation awareness.

The participant's mission performance decreased with an increase in task complexity. These findings validate the experimental design and support hypothesis H1 that an increase in task complexity will decrease situation awareness, increase workload, and decrease mission effectiveness.

B. AI Behavior vs. Fluency

The search optimization AI behavior yielded the highest situation awareness under high task load. The search optimization behavior requires the user to evaluate AI suggestions before deciding whether to implement them or not. Although, there was no corollary under low task load, these results indicate that it becomes important to engage the user in the decision-making process when the operational tempo increases. These results partially support a corollary of hypothesis H2a – as relative workload decreases, situation awareness increases.

The study results indicate that amidst all the workload and other effects of AI behavior, user frustration was the most significant indicator. Specifically, the "collaborative" behavior in which the AI denied user requests without explanation caused the most frustration to the participants. These results partially support hypothesis 2b. In the debrief sessions, users indicated that the lack of transparency of the logic behind the AI's decision making was the primary source of their frustration. In addition to transparency and explain-ability, predictability was another important factor for the participants when it came to AI behavior.

Participants were most frustrated with the AI's collaborative behavior. User frustration is an indication of lack of team fluency. There was also an increase in frustration with task load which suggests a greater need for mechanisms to support team fluency during periods of increased workload. These results support hypothesis H2b that levels of autonomy that share decision authority without transparency such as the Collaborative AI behavior would increase workload, decrease situation awareness, perceived performance, and mission effectiveness.

Although it was not strongly indicated by the statistical results, debrief responses indicated that participants had strong positive affect for the control barrier function enabled collision avoidance behavior. Participants expressed that the assurance that the aircraft would take care of minimizing damage gave them much more trust and affinity for the AI pilot.

C. Fluency vs. Mission Effectiveness

There were general trends of a positively correlated relationship between fluency and performance on all scales. The data indicate that there is an increase in workload with a decrease in performance based on the TLX ratings vs. performance metrics such as mission duration and damage score. Although the differences were not always statistically significant, they support hypothesis H3 that a decrease in team fluency will likely worsen mission effectiveness. As such, it is important to design Human AI interactions that maximize team fluency.

D. Future Work

In future work, higher power and statistical significance could potentially be achieved with a higher sample size and a stronger demarcation of the task load levels. Furthermore, this research aims to further investigate physiological manifestations of workload through heart rate, pupillometry, hand movement, and facial expressions. Development of such real time measures of workload and cognitive state would enable the development of more adaptive autonomy for such safety critical missions.

Acknowledgments

This work was funded in part by the Office of Naval Research, Science of Autonomy grant N00014-21-1-2759 Human-AI Collaboration in Autonomous Aerial Vehicles. The views expressed in this article are those of the author and do not necessarily reflect the official policy or position of the Air Force, the Navy, the Department of Defense or the U.S. Government

References

[1] "Agility Prime - AFWERX," , 2024. URL https://afwerx.com/divisions/prime/agility-prime/.

[2] Ames, A. D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., and Tabuada, P., "Control barrier functions: Theory and applications," *2019 18th European control conference (ECC)*, IEEE, 2019, pp. 3420–3431.

[3] Ames, A. D., Grizzle, J. W., and Tabuada, P., "Control barrier function based quadratic programs with application to adaptive cruise control," *53rd IEEE Conference on Decision and Control*, IEEE, 2014, pp. 6271–6278.

[4] Prajna, S., Jadbabaie, A., and Pappas, G. J., "A framework for worst-case and stochastic safety verification using barrier certificates," *IEEE Transactions on Automatic Control*, Vol. 52, No. 8, 2007, pp. 1415–1428.

[5] Xiao, W., Cassandras, C. G., and Belta, C., *Safe Autonomy with Control Barrier Functions: Theory and Applications*, Springer Nature, 2023.

[6] National Academies of Sciences, E., *Human-AI Teaming: State-of-the-Art and Research Needs*, 2021. https://doi.org/10.17226/26355, URL https://nap.nationalacademies.org/catalog/26355/human-ai-teaming-state-of-the-art-and-research-needs.

[7] Hoffman, G., and Breazeal, C., "Effects of anticipatory action on human-robot teamwork: Efficiency, fluency, and perception of team," *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2007, pp. 1–8. https://doi.org/10.1145/1228716.1228718, iSSN: 2167-2148.

[8] Steiner, I. D., *Group process and productivity*, Academic press New York, 1972.

[9] "Collaborative Combat Aircraft (CCA)," 2023. URL https://www.airforce-technology.com/projects/collaborative-combat-aircraft-cca-usa/.

[10] "MQ-25â„¢ Stingray | NAVAIR," , 2024. URL https://www.navair.navy.mil/product/MQ-25tm-Stingray.

[11] Xwing, "Air Force Awards Xwing Military Approval to Fly Autonomous Air Force Cargo Missions Across California," , 2024. URL https://www.xwing.com/post/air-force-awards-xwing-military-approval-to-fly-autonomous-air-force-cargo-missions-across-californi.

[12] Schneider, M. F., Miller, M. E., and McGuirl, J., "Assessing Quality Goal Rankings as a Method for Communicating Operator Intent," *Journal of Cognitive Engineering and Decision Making*, Vol. 17, No. 1, 2023, pp. 26–48.

[13] Hoffman, G., "Evaluating Fluency in Human–Robot Collaboration," *IEEE Transactions on Human-Machine Systems*, Vol. 49, No. 3, 2019, pp. 209–218. https://doi.org/10.1109/THMS.2019.2904558, conference Name: IEEE Transactions on Human-Machine Systems.

[14] Hoffman, G., and Breazeal, C., "Cost-Based Anticipatory Action Selection for Human–Robot Fluency," *IEEE Transactions on Robotics*, Vol. 23, No. 5, 2007, pp. 952–961. https://doi.org/10.1109/TRO.2007.907483, conference Name: IEEE Transactions on Robotics.

[15] Hoffman, G., and Breazeal, C., "Achieving fluency through perceptual-symbol practice in human-robot collaboration," *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2008, pp. 1–8. https://doi.org/10.1145/1349822.1349824, iSSN: 2167-2148.

[16] Unhelkar, V. V., Lasota, P. A., Tyroller, Q., Buhai, R.-D., Marceau, L., Deml, B., and Shah, J. A., "Human-Aware Robotic Assistant for Collaborative Assembly: Integrating Human Motion Prediction With Planning in Time," *IEEE Robotics and Automation Letters*, Vol. 3, No. 3, 2018, pp. 2394–2401. https://doi.org/10.1109/LRA.2018.2812906, conference Name: IEEE Robotics and Automation Letters.

[17] Gombolay, M. C., Gutierrez, R. A., Clarke, S. G., Sturla, G. F., and Shah, J. A., "Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams," *Autonomous Robots*, Vol. 39, 2015, pp. 293–312.

[18] Romat, H., Williams, M.-A., Wang, X., Johnston, B., and Bard, H., "Natural human-robot interaction using social cues," *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2016, pp. 503–504.

[19] Highland, P., Schnell, T., Woodruff, K., and Avdic-McIntire, G., "Towards Human Objective Real-Time Trust of Autonomy Measures for Combat Aviation," *The International Journal of Aerospace Psychology*, Vol. 33, No. 1, 2023, pp. 1–34. https://doi.org/10.1080/24721840.2022.2127724, URL https://www.tandfonline.com/doi/abs/10.1080/24721840.2022.2127724, publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1080/24721840.2022.2127724.

[20] Napoli, N. J., Stephens, C. L., Kennedy, K. D., Barnes, L. E., Juarez Garcia, E., and Harrivel, A. R., "NAPS Fusion: A framework to overcome experimental data limitations to predict human performance and cognitive task outcomes," *Information Fusion*, Vol. 91, 2023, pp. 15–30. https://doi.org/10.1016/j.inffus.2022.09.016, URL https://www.sciencedirect.com/science/article/pii/S1566253522001488.

[21] Paliga, M., and Pollak, A., "Development and validation of the fluency in human-robot interaction scale. A two-wave study on three perspectives of fluency," *International Journal of Human-Computer Studies*, Vol. 155, 2021.

[22] Agbeyibor, R., Ruia, V., Kolb, J., Jimenez Cortes, C., Coogan, S., and Feigh, K. M., "Towards Safe Collaboration Between Autonomous Pilots and Human Crews for Intelligence, Surveillance, and Reconnaissance," *2024 IEEE/AIAA 43rd Digital Avionics Systems Conference (DASC)*, 2024.

[23] Razin, Y. S., and Feigh, K. M., "Converging Measures and an Emergent Model: A Meta-Analysis of Human-Automation Trust Questionnaires," , No. arXiv:2303.13799, 2023. https://doi.org/10.48550/arXiv.2303.13799, URL http://arxiv.org/abs/2303.13799.

[24] Endsley, M. R., "Design and Evaluation for Situation Awareness Enhancement," Vol. 32, No. 2, 1988, pp. 97–101. https://doi.org/10.1177/154193128803200221, URL https://doi.org/10.1177/154193128803200221, publisher: SAGE Publications.

[25] Hart, S. G., "NASA Task Load Index (TLX)," , 1986. URL https://ntrs.nasa.gov/citations/20000021487, NTRS Author Affiliations: NASA Ames Research Center NTRS Document ID: 20000021487 NTRS Research Center: Ames Research Center (ARC).

[26] Agbeyibor, R., Ruia, V., Kolb, J., and Feigh, K. M., "Joint Intelligence, Surveillance, and Reconnaissance Mission Collaboration with Autonomous Pilots," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 68, 2024. Publisher: SAGE Publications Inc.

[27] Bradley, J. V., "Complete counterbalancing of immediate sequential effects in a Latin square design," *Journal of the American Statistical Association*, Vol. 53, No. 282, 1958, pp. 525–528.

[28] Khalil, H. K., *Nonlinear systems; 3rd ed.*, Prentice-Hall, 2002.

[29] Ames, A. D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., and Tabuada, P., "Control Barrier Functions: Theory and Applications," *2019 18th European Control Conference (ECC)*, 2019, pp. 3420–3431.

[30] Shaffer, F., and Ginsberg, J. P., "An Overview of Heart Rate Variability Metrics and Norms," *Frontiers in Public Health*, Vol. 5, 2017, p. 258. https://doi.org/10.3389/fpubh.2017.00258, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5624990/.

[31] Razin, Y., "Interdependent Trust for Humans and Automation Survey," , 2020.