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

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

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

The maturation of autonomy and electric Vertical Take-Off and Landing aircraft will soon make it possible to execute military Intelligence, Surveillance, Reconnaissance missions aboard manned autonomous aerial vehicles. This research experimentally investigates factors that may influence the quality of interaction between a non-pilot human operator and the AI pilot aboard such an aircraft. In a flight simulator study with twenty-nine participants, various levels of workload and AI capabilities are investigated including run time assurance. Control Barrier Functions are used to enable pro-active collision avoidance behaviors by an AI agent. Team fluency and mission effectiveness outcomes show that trust, situation awareness, workload, perceived performance and user interface design are statistically significant factors for the quality of human AI interaction in this 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 (AAM). The industry developing around Urban 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 manned autonomous eVTOLs. The U.S. military has expressed interest in using these new vehicles for its own specialized military missions.

One such specialized military mission – Intelligence, Surveillance, and Reconnaissance (ISR) – could leverage autonomous eVTOLs to reduce 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 adaption of off-the-shelf autonomous aerial vehicles for military missions *.

The ISR mission is split into unmanned and manned operations, with unmanned operations being the predominant type. The unmanned ISR mission is conducted using remotely piloted aircraft like the MQ-9. These vehicles are currently operated through space-based satellite communication systems by a ground crew consisting of a pilot, a sensor operator, and an intelligence analyst. Manned ISR is conducted with a similar crew composition to unmanned ISR but at a larger scale on more complex aircraft like the P-8. Manned ISR aircraft usually require multiple pilots, multiple sensor operators and multiple intelligence analysts for missions of higher sensitivity and at greater distances from an operating base. 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 ISR vehicles, it is possible that the U.S. military would send operators aboard eVTOLs to conduct manned autonomous ISR. Autonomous aircraft capabilities alone could greatly reduce the manning requirements for ISR. With the piloting task in gross part delegated to an autonomous pilot, the crew requirements could be significantly reduced.

*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, School of Electrical and Computer Engineering.

‖Professor and Associate Chair for Research, School of Aerospace Engineering.

*Personal communications with and interviews of U.S. Navy, AFWERX, DARPA personnel.

No unmanned operation can succeed if either the autonomous aircraft or the human on board are put at risk. Run time assurance mechanisms should be included in the autonomy's behavior to minimize risk and to enforce safety throughout the mission. Safety of controlled systems can be modeled as a constraint on the system's state. Barrier functions or control barrier functions (CBFs) enforce forward invariance of the constraint set so that no trajectory initialized within the constraint set ever leaves or violates the constraint set [**? ? ? ?**], 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 [**?**].

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" [**?**]. 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 [**?**]. 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 at how expert operators – licensed drivers and professional pilots – collaborate with autonomous technologies. The authors did not find any specific study on 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, aerial refueling, both of which are centered on experts collaborating with autonomy. Some of this research looked at supporting unmanned ISR pilots with better autonomy, however, it still assumed the presence of well trained expert pilots [**?**].

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 teammate 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 manned 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 Control Barrier Functions 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 proffer three hypotheses as to how the task complexity and fluency will interact to impact aspects of mission performance.

- H1: Increase in task complexity will decrease situation awareness, increase workload, and decrease mission effectiveness.
- H2a: Levels of autonomy that increase decision support such as Control Barrier Function 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: Decrease in fluency — marked by increased trust, situation awareness, perceived performance; decreased interaction and workload – will decrease mission effectiveness.

II. Background

The assessment of fluency in human-robot collaboration was presented, mapped, and validated by Hoffman [?]. 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 are which statistically evaluate the level of fluency in a particular interaction and subjective metrics gauge people's perceptions of the fluency of an interaction and associated features of the robot. In their preceding papers [?], [?], [?], 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. [?] 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 nowadays along with planning with detection and planning with prediction.

Gombolay et al.[?] 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. [?] evaluate human-robot interaction using affordance and social cues as metrics. The attributes and characteristics of an object that determine its potential uses and suggest how it should be utilized are referred to as affordance.

In live-flight dog-fight experiments in fighter aircraft, the University of Iowa Operator Performance Laboratory used eye tracking, ECG and other physiological measures to assess operator trust in an autonomous pilot agent. Highland, Schnell et al. [?] 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 [?] 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 [?] 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. [?] explore the impact of coordination, communication and intent on human-AI teams in the context of manned ISR missions. This and the other works mentioned above inform the design of this study.

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 29 participants volunteered to complete an ISR task while interacting with the AI pilot over a series of scenarios which varied in task complexity and load.

A. Operationalization of HAI Fluency

Fluency in collaboration is assessed through subjective and objective measured. 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, perceived performance, interaction and workload.

B. Participants

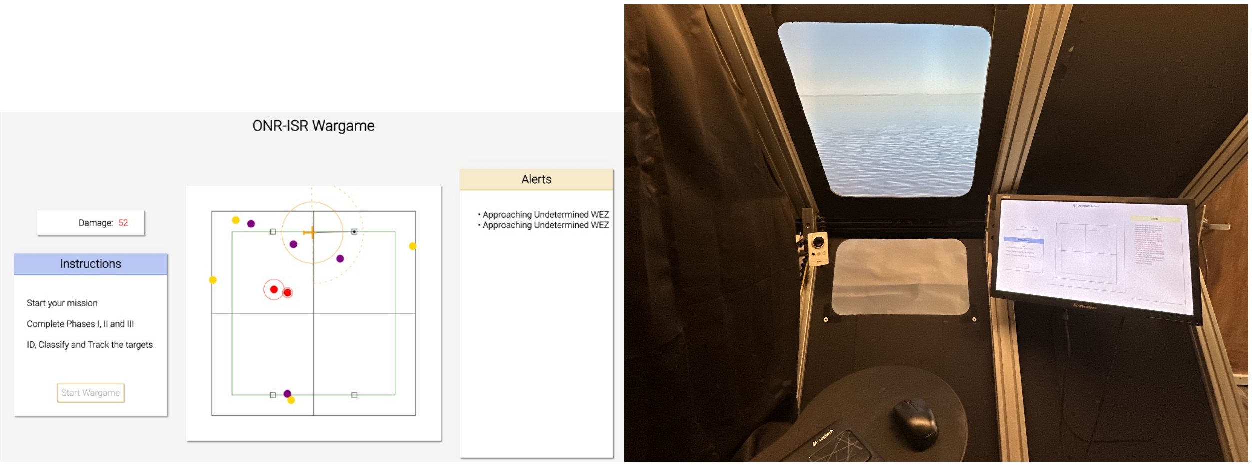
The study was approved by the Georgia Institute of Technology Institutional Review Board. Participants were recruited in the broader Atlanta area and Georgia Institute of Technology community. Participation was limited to english speaker, between 18 and 65 years old, who are not color blind, and do not have a pacemaker or similar heart rate stabilization devices.

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

C. Experimental Apparatus

1. AI Software

The ISR scenario employed was originally developed by an operational Navy P-8 Poseidon pilot based on US Navy training scenarios. The graphical user interface for this simulator is shown in Fig. 1a



Start your mission

Complete Phases I, II and III

ID, Classify and Track the targets

• Approaching Undetermined WEZ

• Approaching Undetermined WEZ

(a) Simulator Cabin from the participant's perspective **(b) Simulator Cabin from the participant's perspective**

Fig. 1 Simulator GUI and Cabin

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 ISR missions. 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 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

Control Barrier Functions were implemented using a second order unicycle model for the aircraft. This CBF enabled controller took as input the location of target ship in the simulated scenario and built barriers around them based on their parameterized Weapon Employment Zone (WEZ) size. This controller is used for run time assurance of collision avoidance for the simulated autonomous eVTOL.

If the trajectory of the aircraft was to head towards an enemy ship, the controller would slow down the aircraft and change its flight path 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 behaviors:

- **Way point Behavior (Level 0)** - The AI's way point behavior allows the AI to fly an automated search pattern. At any time, the user can override the AI's way point by clicking on a point on the screen to cast a user vector.
- **Collaborative Behavior (Level 1)** - In the collaborative behavior, the AI flies a search pattern and at any time, the user can request a different way point by clicking on a point in the surveillance area. The user's way point 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 search pattern.

AI Behavior	Waypoint -	Collaborative -	Collision Avoidance -	Search Optimization -
Taskload	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

- **Collision Avoidance (Level 2)** - In the collision avoidance behavior, the AI flies its search pattern way points 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 way point vector. The AI flies to the user's way point 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 way point by clicking on a point on the screen to cast a user vector.

There are two taskload levels in the experimental design: a Low level (Level A) and a High level (Level B). The Low taskload level (Level A) has 10 targets, moving at 5 knots, with a simplified square search pattern. The High taskload 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.

E. Design of Experiments

A repeated measures, within-subjects design of experiments is employed to minimize the effect of individual differences amongst the participants. Each participant completes each scenario.

Fatigue and learning over time can have significant effects on the results of the study. To control 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 [**?**] 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. In this section we now describe 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 $h(x)$, called a *Barrier Function*, with the property 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 will remain inside S for all time $t \geq 0$ [**?**]. Then the boundary of S, denoted $\partial S = S \backslash \text{int}(S)$, is given by $\partial S = \{x \mid h(x) = 0\}$. 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 relaxed to

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

for some locally Lipschitz function $\alpha : \mathbb{R} \longrightarrow \mathbb{R}$ satisfying $\alpha(0) = 0$. The advantage is that this condition, which must hold for all x rather than only on the boundary of S, 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. Notably, the inequality in the definition of $U(x)$ is affine in u and, therefore, can be included in convex optimization programs that compute a feedback controller $\sigma(x)$, possibly online at runtime. If such a feedback controller exists, then $h(x)$ is called a (classical) *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:

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

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 [**?** , 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 their Age, Gender, Aviation 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 way point clicks and the number of AI alerts are recorded to understand the interaction between the user and AI. The aircraft damage score and mission duration are used to determine the users objective mission performance.

3. NASA Task Load Index (TLX) Data

The NASA TLX scale divides the participant's overall workload into six categories including: mental demand, physical demand, temporal (time-pressure) demand, performance, effort, and frustration. Users completed a modified NASA TLX scale after each scenario of the experiment. Instead of the traditional 21 increments on paper, users completed the TLX on a sliding scale from 0-100 on a computer-based survey tool for convenience. This data provided insight on workload and emotional response of participants.

4. Situational Awareness Questionnaire Data

In order to gauge the situational 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, what area was the most populated, or what the participant's score was at the end of a scenario. Questions were graded with a dichotomous pass/fail criterion, incorporating a margin of error.

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 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. When RMSSD ≤ 0.068, the heart rhythm is normal [**?**].

6. Post-Experiment Questionnaire Data

A post-experiment questionnaire is administered to all participants at the end of all 8 scenarios. The questionnaire is designed to understand the user's overall trust, communication, and perceived performance throughout the whole experiment. This survey is based on the I-THAu Questionnaire developed in-house at Georgia Institute of Technology.

IV. Results and Analysis

At the time of the submission, this manuscript presents preliminary results. Detailed analysis is ongoing and will be included in the final version of this work.

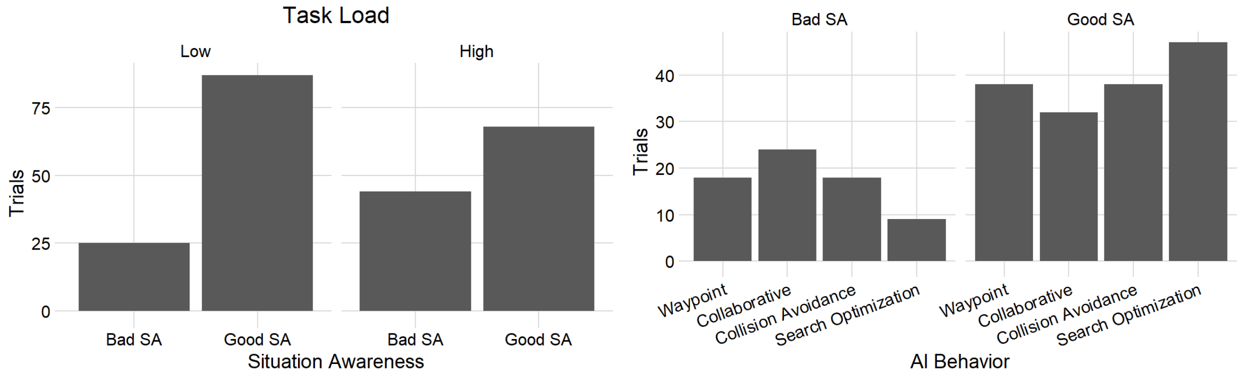
A. RQ1: Task Complexity vs. Fluency

To answer RQ1, the following data were analyzed using a Linear mixed model:

- Independent Variables (IVs): Task load, AI Behavior
- Confounding Variables (CVs): Participant Flight Experience, Participant AI Experience
- Dependent Variables (DVs):
  - Situation Awareness
  - Workload TLX: Mental Demand, Physical Demand, Temporal Demand, Effort, Frustration
  - Workload Physiology: Mean RMSSD, Pupillometry
  - Mission effectiveness: Mission Duration, Damage

1. Situation Awareness

Figure 2a shows how situation awareness changes with taskload. Pearson's Chi-squared test is used to quantify the correlation between situation awareness and taskload. There is a clear relationship between the user's situation awareness and task load: a lower task load typically resulted in higher situation awareness and higher task loads resulted in lower situation awareness. The Chi-squared test gives a significant $X - squared = 6.786$ with $p = 0.009$.



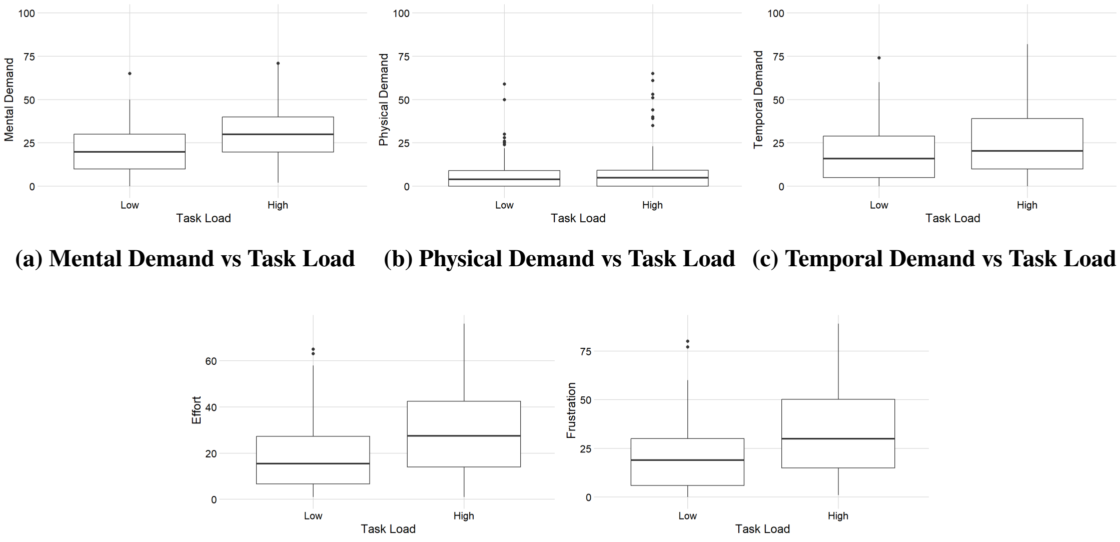
(a) Situation Awareness vs Task Load

(b) Situation Awareness vs AI Behavior

Fig. 2 Situation Awareness vs Scenario Factors

2. Workload TLX

Figure 3 shows how NASA TLX subscales mental demand, physical demand, temporal demand, effort and frustration change with task load. Table 2 shows the p-values obtained for each through linear mixed effects modeling.



(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. 3 NASA TLX subscales vs 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***
Perceived Performance	-16.4	p < 0.0001***

Table 2 Variance in Workload per Task Load

As expected, the users 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.

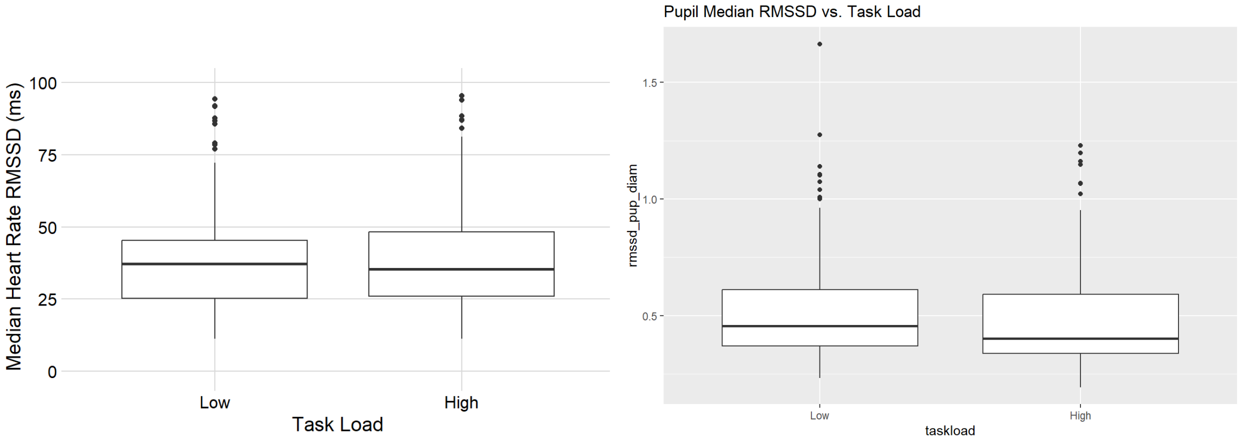
3. Workload Physiology

Figure 4 shows how Heart Rate Variability and Pupil Diameter change with task load. Statistical analysis did not find significance in the data, meaning that there is not a clear relationship between heart rate variability and task load. Statistical analysis did not find a model that could represent a relationship in the pupillometry data with statistical significance. Similar to heart rate, there is no relationship between pupil diameter variation and taskload.

4. Mission Effectiveness

Figure 5a shows how mission duration changes with taskload. With a higher task load, mission duration increased. A minimal linear mixed effects model analysis yields a very significant difference in a repeated measures ANOVA with $p < 0.0001 * **$.

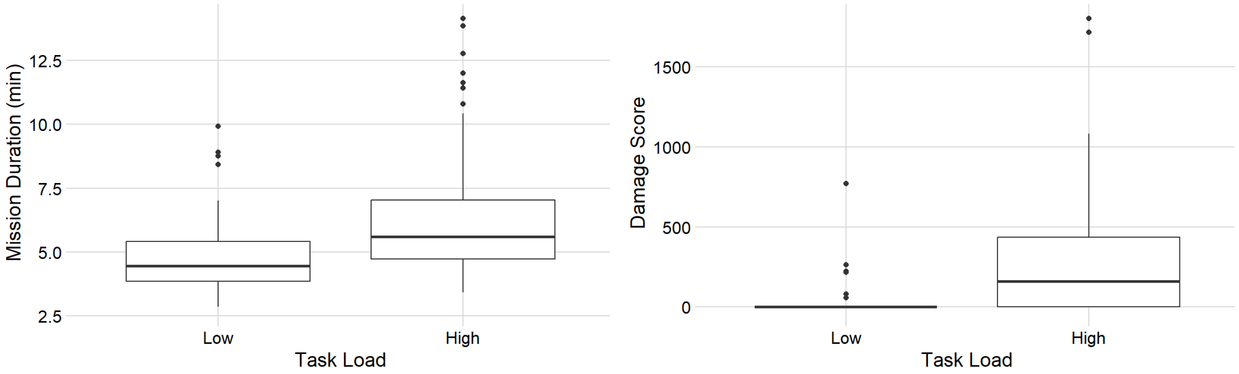
Figure 5b shows how aircraft damage changes with taskload. The aircraft damage score increased with a higher taskload. There was minimal damage in most of the lower task load scenarios. A minimal linear mixed effects model found a very significant difference in a repeated measures ANOVA with $p < 0.0001 * **$.



(a) Heart Rate Variability vs Task Load

(b) Pupillometry vs Task Load

Fig. 4 Physiology vs Task Load



(a) Mission Duration vs Task Load

(b) Aircraft Damage vs Task Load

Fig. 5 Mission Effectiveness vs Task Load

B. AI Behavior vs. Fluency

To answer RQ2, the following data were analyzed:

- IVs: AI Behavior
- CVs: Participant Gender, Participant Flight Experience, Participant AI Experience
- DVs:
  - Trust: I-THAu Trust, I-THAu 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, Pupillometry

No significant relationship was found for trust or physiological workload data when looking at their connection to the four different AI behaviors.

1. Situation Awareness

Figure 2b shows how situation awareness changes with AI behavior. There are large differences in participant situation awareness across all AI behaviors, with the highest situation awareness in the collaborative behavior in the low task load. The search optimization mode yielded the highest situation awareness in the high task load. The relationship between situation awareness and AI behavior varied depending on the task load. The Pearson's Chi-squared test was applied to quantify the correlation between situation awareness and taskload. The Chi-squared test gives a significant $X - squared = 9.6135$ with $p = 0.02$.

2. Perceived Performance

Figure 6 shows how Perceived Performance reported in the TLX questionnaire changes with AI behavior. There are significant differences in perceived performance across all of the AI behaviors. On average, the collaborative AI behavior yielded the highest perceived-performance and the way point behavior had the lowest perceived performance. A minimal linear mixed effects model analysis yields a very significant difference in a repeated measures ANOVA with $p = 0.0019$.

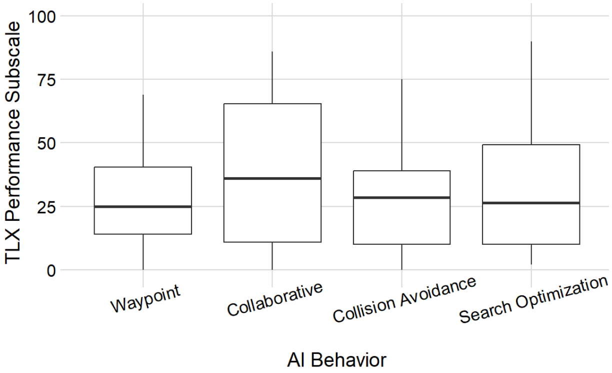


Fig. 6 Perceived Performance and AI behavior

3. Interaction

Figure 7 shows how the number of user way point inputs change with AI behavior. On average, users click to cast waypoint vectors the most in the collaborative behavior and the least in the waypoint and search optimization mode. A minimal linear mixed effects model found a very significant difference in a repeated measures ANOVA with $p < 0.0001 * **$.

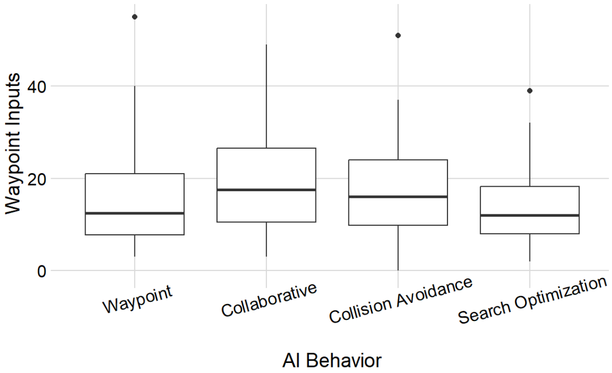


Fig. 7 Number of User Waypoints entered vs AI Behavior

4. Workload TLX

Figure 8 shows how mental demand, physical demand, temporal demand, effort and frustration change with AI Behavior. Table 3 summarizes the p-values.

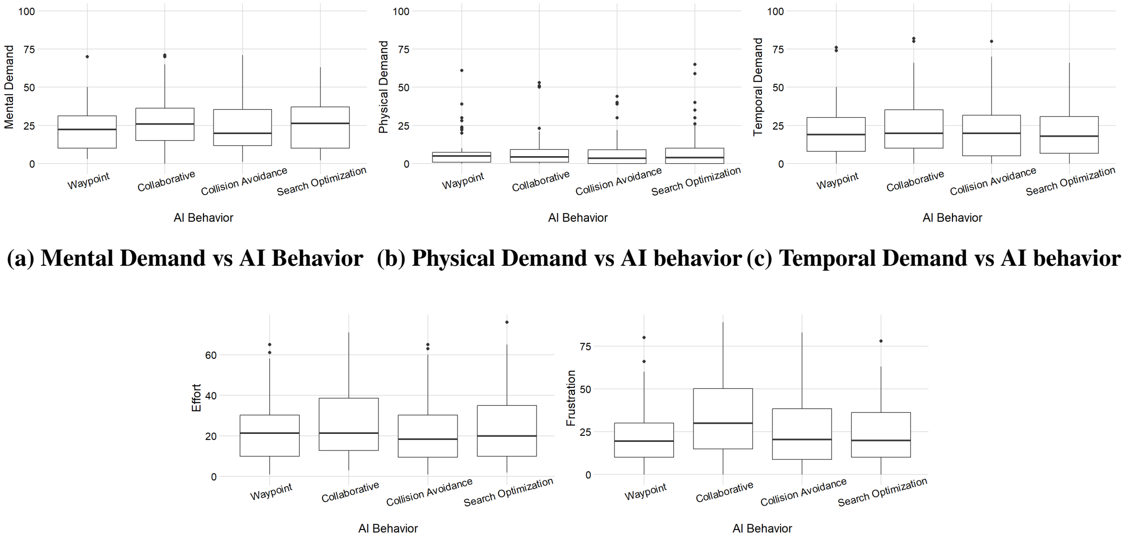
TLX Subscale	p-value
Mental Demand	p > 0.05
Physical Demand	p > 0.05
Temporal Demand	p > 0.05
Effort	p > 0.05
Frustration	p = 0.0005**
Perceived Performance	p = 0.0019*

Table 3 Variance in Workload per AI Behavior

The collaborative and search optimization modes tend to have more perceived mental demand than the other AI behaviors. The collision avoidance behavior has the smallest amount of perceived mental demand 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 i.e. time-pressure amongst AI behaviors. Neither are there any significant relationships between behavior and effort, however, the largest variance is in the collaborative mode.

Figure 8e shows how frustration changes with AI behavior. There is a much larger amount of perceived frustration with the collaborative AI behavior compared to the other behaviors. By a very slight margin the waypoint behavior had the smallest amount of frustration on average. A minimal linear mixed effects model found a very significant difference in a repeated measures ANOVA with $p = 0.0005$.



(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. 8 TLX vs AI Behavior

C. Fluency vs. Mission Effectiveness

To answer RQ3, the following data were analyzed:

- IVs:
  - Trust: I-THAu Trust
  - I-THAu Team Communication and Interaction
  - Situation Awareness: Situation Awareness
  - Perceived Performance: Mean 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, Pupillometry
- CVs: Participant Gender, Participant Flight Experience, Participant AI Experience
- DVs: Mission effectiveness: Mission Duration, Damage

No significance was found statistically or visually for trust, situation awareness, and interaction when looking at their relationship with fluency.

1. Perceived Performance

There is a slight negative correlation between perceived performance and damage score. This correlated relationship can be seen in 9 This result was not found to be statistically significant.

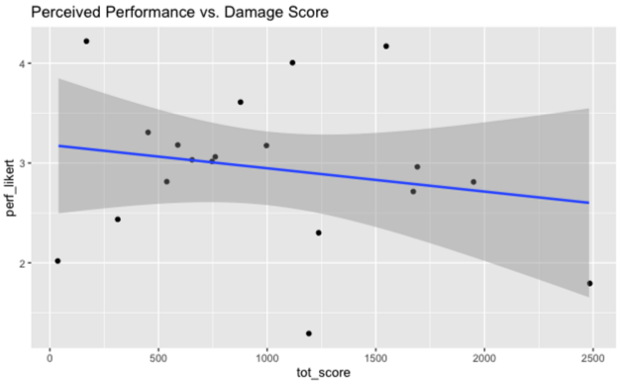
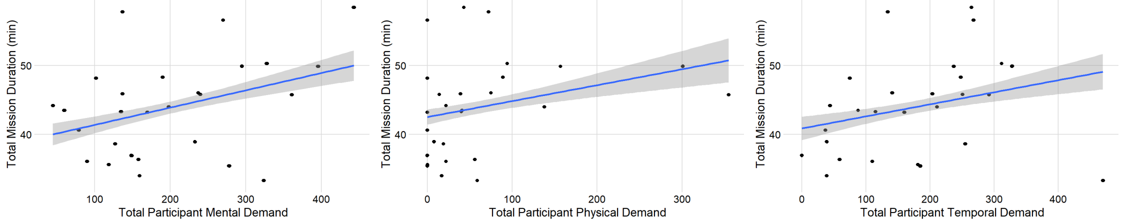


Fig. 9 Damage vs Perceived Performance

2. Workload TLX

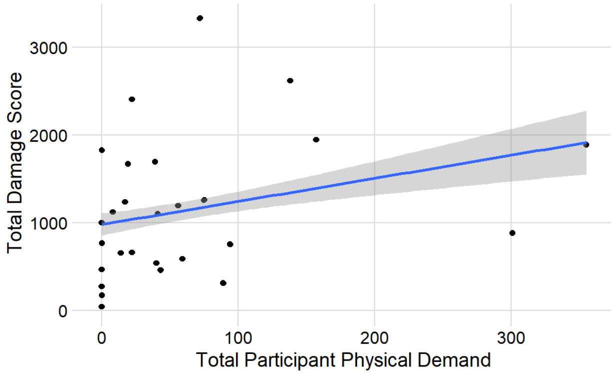
There is a rough positive correlation between the TLX sub scales for mental, physical, and time-pressure demand and the total participant mission duration. There was no statistical significance in any of the results between TLX perceived workload and total mission duration. Out of all of the TLX relationships, in Fig. 10a, Fig.10b, and Fig. 10c mental demand had the strongest coefficient of correlation with total mission duration.



(a) Mental Demand **(b) Physical Demand** **(c) Temporal Demand**

Fig. 10 Mission Duration vs Workload

There is also a subtle positive relationship between damage and perceived physical demand from the TLX scale. This relationship is in Fig. 11a but it is not statistically significant.



(a) Damage vs Physical Demand

Fig. 11 Damage vs Workload

3. Workload Physiology

There is an approximate association between the total mission duration and the heart rate variability of participants. This relationship is plotted in Fig.12 but it is also not statistically significant. The pupillometry data did not prove to be significant.

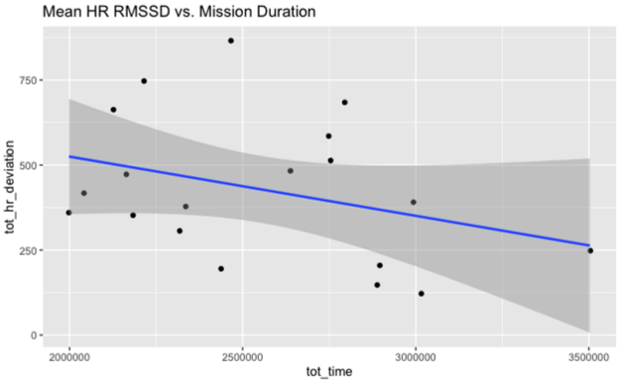


Fig. 12 Mission Duration vs Heart Rate Variability

V. Discussion

In the HAI literature, fluency is oft 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 [**?**]. This context required broadening of the definition of fluency to include workload, trust, situation 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 TLX rating 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 yield the highest situation awareness under high task load. The search optimization behavior requires the user to evaluate AI suggestions and choose 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 an unexpected 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 significantly indicator. Specifically, the "collaborative" behavior in which the AI denied user requests without explanation caused the most frustration to the participants. In the debrief sessions, users indicated that the lack of understanding 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 factors for the users when it came to AI behavior.

Users were most frustrated with the AI's collaborative behavior; when users are frustrated it is likely due to a lack of team fluency. There is also an increase in frustration with task complexity as mentioned earlier; this suggests a heightened need for team fluency during periods of increase workload. These results support hypothesis H2b that levels of autonomy that share decision authority without transparency such as the Collaborative AI 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.

The results did not support the hypothesis that AI behavior would significantly affect perceived performance.

C. Fluency vs. Mission Effectiveness

There are general trends of a positively correlated relationship between fluency and performance on all scales. The data indicate that there are 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 these results are not statistically significant they support the theory that an increase in team fluency will likely improve mission effectiveness.

In future work, statistical significance could potentially be achieved with a higher sample size and a stronger demarcation of the task load levels.

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.

References