RICE UNIVERSITY

By

Tewodros B Taffese

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE

Michael Byrne
Michael Byrne (Nov 29, 2023 18:51 CST)

Michael Byrne (Chair)

Eduardo Salas

Vaibhav Unhelkar

HOUSTON, TEXAS

December 2023

Workload-Based Subtask Selection for Automation

by

Tewodros B Taffese

Abstract

Workload is a construct used to understand and predict human-automation performance. For an effective human-autonomy collaboration, it is important to maintain workload at an optimum level. Excessive workload impedes performance, while low workload might lead to boredom, lack of vigilance, and lack of situation awareness. While there is existing research on workload and its impact on automation, there seems to be lack of guidelines regarding which components of a task should be subject to automation. The purpose of this dissertation was to address this gap by providing a framework for task selection in automation. Three experiment was conducted to investigate the efficacy of workload-based task selection for automation. Participants performed the Theatre Defense Task, which is a simulated military operation task. The first experiment compared subtask workloads while participants performed the task without automation, identifying high and low workload subtasks. Workload differences between subtasks were measured using subjective workload ratings and performance-based metrics. In the second experiment, these subtasks were automated and compared in different conditions. The third experiment introduced a design intervention to address observed strategies from the second experiment and to optimize the automation. Participants exhibited higher performance scores in high workload subtask automation compared to low workload subtask automation conditions, especially when the task difficulty was high.

The workload rating of different subtasks appeared to be influenced by the automation. The results showed that automation alters human-system interactions. Workload-based task

selection demonstrated performance improvement and reduced workload. More importantly, automating high workload subtasks significantly improved and reduced overall workload, while automating low workload components of a task had no impact on performance. This study underscores the significance of exploring the impact of subtask automation on both the overall task and individual subtask performances.

Acknowledgements

First, I would like to express my sincere gratitude to my mentor and advisor, Dr. Michael Byrne, for believing in me and providing me with this invaluable opportunity. Thank you for your support and understanding throughout this journey. I have faced various personal issues along the way, and I would not have coped if it was not for your help and understanding. I have gained a wealth of knowledge from our collaboration, and I will always be grateful. Second, I would like to extend heartfelt thanks to my wife and life partner, Zelalem Weldeselassie, for her love and unwavering support during my PhD journey. I am also grateful to Dr. Edwardo Salas and Dr. Vaibhav Unhelkar for their service on my committee. The valuable feedback they provided played a crucial role in shaping my thesis. I would also like to thank the Human Factors faculty and Psychology department staff at Rice University. You have all been very helpful and kind during this journey. Especially, I would like to express my sincere appreciation to Lanita Martin and Carrie Hodgeson. I cannot thank you enough for all your hard work and assistance. Special thanks to Dr. David Kaber and Dr. Melanie Wright for granting me access and permission to modify the software used in this research. I am also indebted to Kami Boutotte and Adi Zytek for their assistance with data collection, coding, and insightful observations. I want to express my appreciation for the friends I have gained at Rice, Dr. Shivam Pandey, Dr. Ian Robertson, Dr. Evan Mulfinger, Dr. Leo Alexander, Dr. Adam Braly, Dr. Brad Weaver, Dr. Xianni Wang, Dr. John Lindstedt, Dr. Meghan Davenport, Dr. Jennie Paoletti, Dr. Roslyn Shanklin, Sangeeth Jeevan, Will Mathews, Fabrizio Chavez, and Charlie Weeks. Thank you all for making my time at Rice a joyful experience. I could not have asked for better friends than you all. I would like to express my gratitude to my late father Belete Taffese, whom I lost during this journey. I know he was very proud of my pursuit and his role is significant in making me the

person I am today. I also extend my thanks to my siblings Bemenet, Biniam, and Mahlet for being a constant source of encouragement. Most importantly, I dedicate this achievement to my late mother Ejigayehu Assefa. Everything positive about me is a reflection of her influence. The significance of her role in my life cannot be expressed in simple words. Not a day goes by without thoughts of her, and I pursued this journey in her honor.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

Blended decision Making: BDM

Instantaneous Subjective Assessment: ISA

Theatre Defense Task: TDT

Air Commander: AC

Information Officer: IO

Levels of Automation: LOA

Degrees of Automation: DOA

Dynamic Function Allocation: DFA

Dynamic Difficulty Adjustment: DDA

Adaptive Automation: AA

Chapter 1

Introduction

The deployment of automation is driven by the objective of enhancing task performance, reducing cost or labor, bolstering safety, and increasing efficiency. Many autonomous systems work under supervision or in cooperation with human operators. These automations are applied in diverse contexts and varying levels. There are various Human Factors (HF) concerns regarding the design of autonomous systems that work collaboratively with an operator. Some of these issues include the out-of-the-loop performance problems, vigilance, monitoring (Endsley, 2017), calibration and measurement of trust (Yan et al., 2011; Okamura & Yamada, 2020; Endsley, 2017), shared mental model (Fan & Yen, 2011; Lorenzo, 2019), team coordination and communication, role allocation between human and systems, team performance improvements, and maintaining an appropriate operator workload (O'Neal, 2020; Endsley, 1987).

These concerns have some overlap and may also lead to subsequent challenges. For instance, the out-of-the-loop problem results in lack of situation awareness which leads to a task handover problem between operator and automation, performance degradation because the operator is not actively involved in execution of tasks, and communication and feedback problems (Endsley, 2017). High workload could lead to operator fatigue, frustration, lack of motivation, reduced efficiency and/or performance degradation. The lack of appropriate level of task demand or engagement with the work may also lead to issues such as operator complacency.

Automation induced operator complacency occurs when automation is highly reliable but not perfect (Parasuraman et al., 2000; Bailey 2004). In this case the operator plays a backup role, and their monitoring performance declines to a point where they fail to detect when the automation occasionally fails. Operator complacency usually occurs due to an over-trust and

heavy reliance in an automated system. Trust, situation awareness, and workload were described as valuable constructs of understanding and predicting human-automation performance (Parasuraman, Sheridan, & Wickens, 2008). A good automation system that works in collaboration with the human should be able to keep workload, situation awareness, and trust at optimum levels while improving task performance. The goal of this research is to investigate the efficacy of a workload-based task selection for automation.

Automation Approaches

Technological advancements have significantly expanded the scope and applications of automation. Some automated systems are now integrated to complex task such as flying an airplane, driving vehicles, and conducting medical diagnostics. Automation is applied in different levels, ranging from complete manual control to fully autonomous systems. Previously researched methods of automation include, the MABA-MABA (Men are better at-Machines are better at) approach (Fitts et al., 1951), the levels of automation framework (Sheridan & Verplank, 1978) where different levels define how much of the tasks the automation performs independently, and automation of system functions that aligns with the four-stage model of human information processing (Parasuraman, Sheridan, & Wickens, 2000; Wright & Kaber, 2005).

The MABA-MABA strategy suggests that roles should be assigned to humans and intelligent systems based on what they are comparatively good at (see Table 1). For example, humans are better at exercising judgement than machines are, hence such tasks should be assigned to humans. On the other hand, computational tasks should be allocated to machines. This distinction is based on a survey done at the time of the research and was expected to change

as technology and humans evolve through time. But the general idea is that role allocation should be done based on who is better at performing different types of tasks.

Table 1
The MABA-MABA list (Fitts 1951)

Men are better at	Machines are better at
Detecting small amounts of visual or acoustic energy	Responding quickly to control signals and applying great force smoothly and precisely.
Perceiving patterns of light or sound	Performing repetitive, routine tasks
Impoverishing and using flexible procedures	Storing information briefly, and then erasing it completely
Storing information for long periods of time and recalling appropriate parts at appropriate times	Reasoning deductively, including computational ability
Reasoning inductively	Handling complex operations, i.e., to do many different things at once
Exercising judgement	

Sheridan and Verplank (1978) designed a ten-point taxonomy that define the degrees of automation, ranging from 0 or no automation to 10 or full automation. This scale is intended to be used as a guideline to determine how much of a task should be automated (see Table 2). Autonomous systems that are in the range of levels 1 to 4 assist the operator by presenting a set of decision/action alternatives, narrowing down selections, and suggesting one alternative (Parasuraman et al., 2000) (see Table 2). At level 5, the computer could execute the suggestion it makes if a human approves it. However, in higher levels of automation such as level 7, the

computer performs an action and informs the human operator what it did. It is important to note that these recommendations are intended to provide a framework for design and should not be seen as fixed human and machine role assignments or substitutions (Parasuraman, Sheridan, & Wickens, 2008).

Table 2

Levels of automation of decision and action selection (Parasuraman, Sheridan, & Wickens, 2000), modified from Sheridan (1987)

High	10.	The computer decides everything, acts autonomously, ignoring the human
	9.	Computer executes an action and informs the human only if it, the computer, decides to
	8.	Computer executes an action and informs the human only if asked
	7.	Computer executes actions automatically, then necessarily informs the human what it did
	6.	Computer selects an action, informs the human, and allows the human a restricted time to veto before automatic execution
	5.	Computer suggests an action and executes that suggestion if the human approves
	4.	Computer suggests an action and the human may or may not do it
	3.	Computer narrows the selection down to a few; human may select one
	2.	The computer offers a complete set of decision/action alternatives
Low	1.	The computer offers no assistance: human must take all the decisions and actions

In automation levels 4 and below, even though there may be some interdependency between the human operator and the autonomous system, there will likely not be perceived agency of the automation. Automations at this level trim down their suggested alternatives while remaining subordinate to the human operator (O'Neal et al., 2020). In this case the autonomous system does not have a perceived agency but can work with and assist the human in completing tasks.

The MABA-MABA approach, the Levels of Automation (LOA) framework, and the information processing stages of system allocation methods are extensions or rather than

replacement of each other. The LOA framework can be used with task allocation methods such as MABA-MABA. Functions that machines are good at could also be automated to various levels of automation. For instance, the computer can be assigned computational tasks, but its decision selection and action performing authority could be limited to the level of automation. Parasuraman, Sheridan, and Wickens (2000) proposed a model that integrated types and levels of automation as a guide to make function allocation decisions. These system functions are drawn from the four-stage model of human information processing model (see Figure 1). The four groups of system functions are:

1. Information acquisition which is paralleled to sensory processing.
2. Information analysis which represents conscious perception and processing of information in the working memory.
3. Decision and action selection which is where decision is reached based on cognitive operations such as rehearsal, integration, and inference that occur in stage 2.
4. Action implementation which represents implementation of a response or an action (Parasuraman et al., 2000; see Figure 2)

Figure 1

Simple four-stage model of human information processing (Parasuraman et al., 2000)

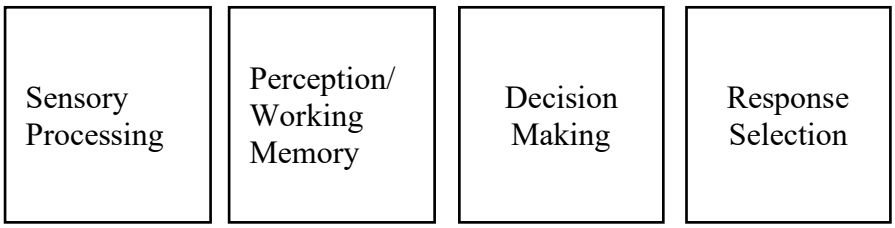
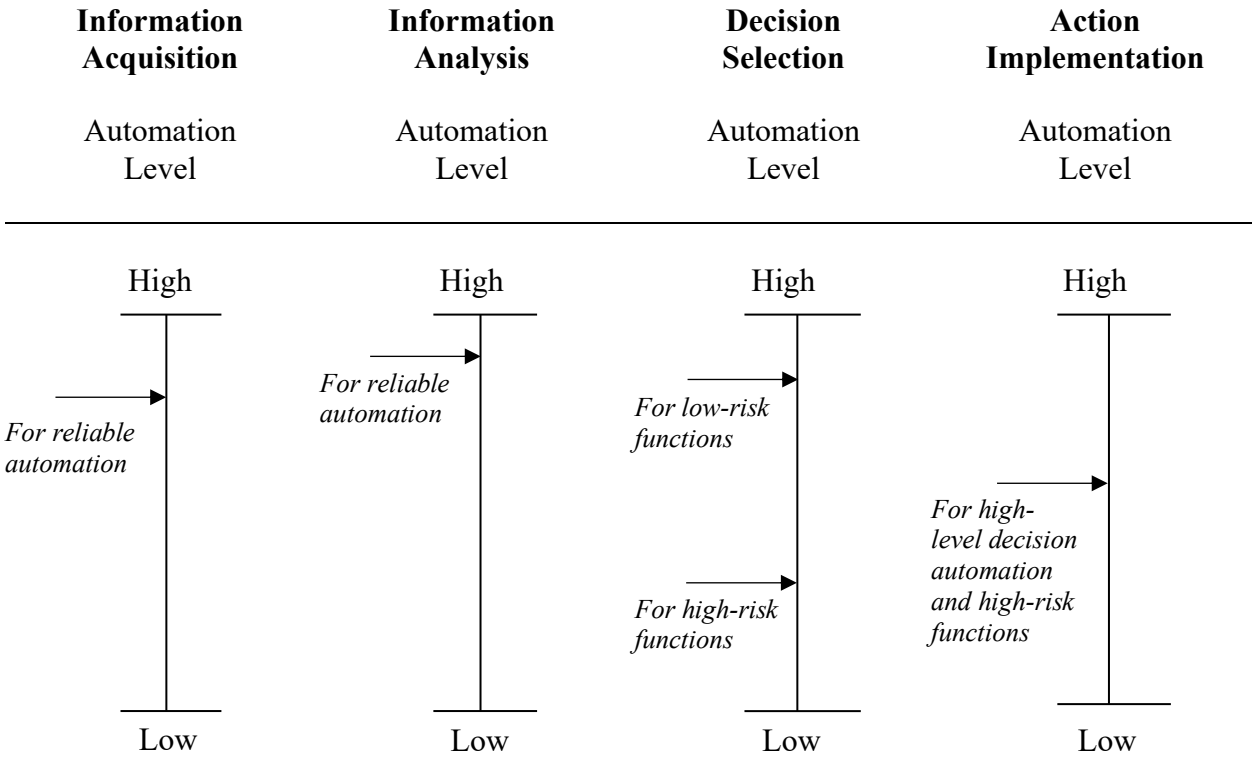


Figure 2
Combined implementation of levels of automation with four stages of system function allocation



The four-stage model is intended to supplement the 10-point level of automation taxonomy integrated with automation of system functions. With the combined approach each type or stage of automation can be automated to different levels (Parasuraman, Sheridan, & Wickens, 2000; see Figure 2). One of the objectives of such automation is to reduce workload of the human operator and improve efficiency in the context of human-autonomous system interaction. However, research shows that automation does not always reduce workload and can sometimes even increase workload on the operator (Parasuraman & Mouloua, 1997; Stapel, Mulakkal-Babu, Happee, 2018). Before delving into background on workload in automation, it is important to provide a brief explanation of theories and measurement related to workload.

Theory for Workload Measurement

Before talking about measurements of workload, it is important to define what is meant by workload. From the psychology and neuroscience perspective, the most common definition of workload is based on a resource processing theory (Kahneman 1973, Wickens, 2008; Norman & Bobrow, 1975). A resource in this case refers to the human mental capacity to process and manage information. Mental workload is a function of the demand imposed by a task on the limited human mental resources. Hence workload can be defined as the balance between task demand and mental resources (Young et al., 2015).

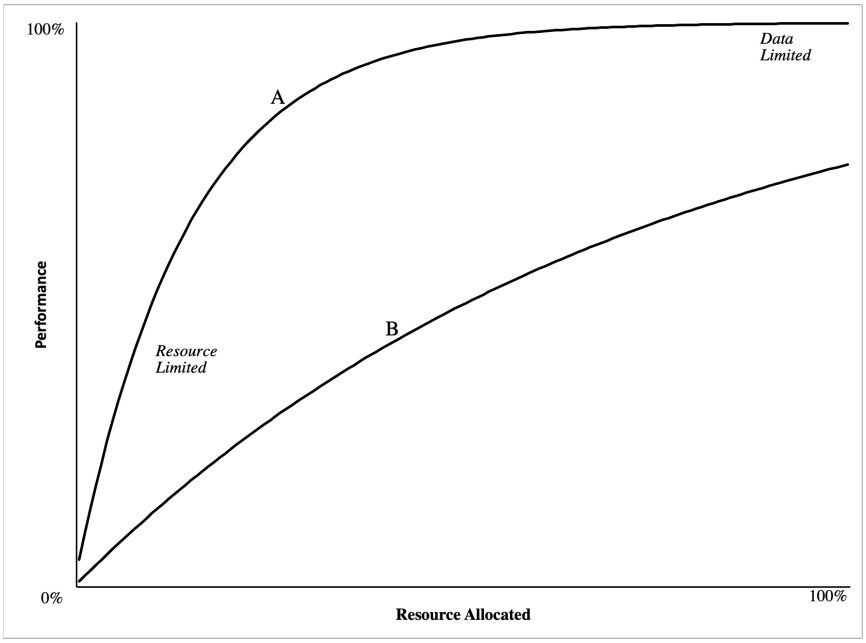
When a task demand is low, a person will have sufficient mental resources to deal with the task and the effect of workload is minimum. However, when tasks get more complicated, the human will need to allocate more resources to the task. If the task demand reaches a level where the human need to commit their full resources, then the high workload could lead to a deteriorating human performance. Norman and Bobrow (1975), introduced Performance Resource Function (PRF), which helps understand the relationship between task demand, mental resources, and performance.

Performance Resource Function (PRF)

The Performance Resource Function (PRF) model defines workload as the percentage of attentional resources invested to perform a task (Norma ad Bobrow, 1975). The performance axis represents the percentage of maximum performance possible (see Figure 3). Looking at curve A, the increasing part of the curve is resource limited. At this stage performance can be improved by adding more attentional resources to the task. The flatter section of the curve is data-limited, meaning that changes in resources allocated to the task have no effect on performance.

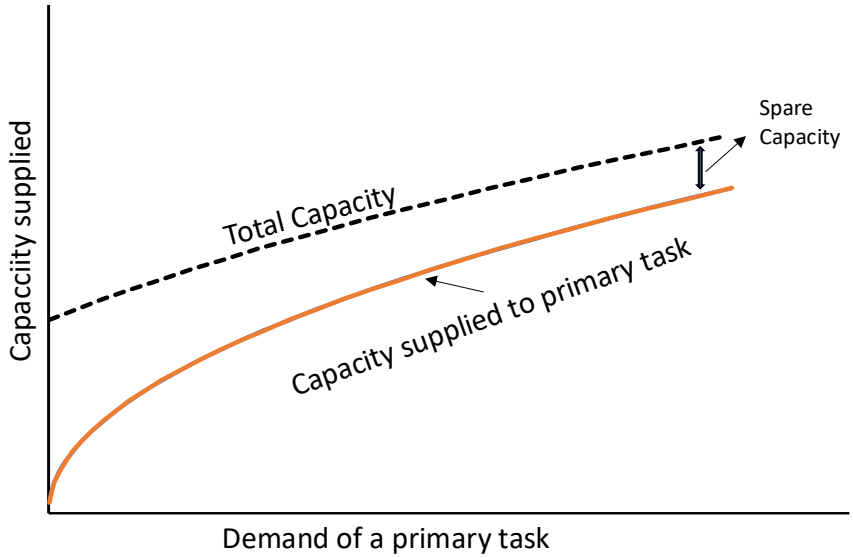
Figure 3

Example of the PRF Curve for Tasks That Differ in Level of Difficulty. (Matthews & Reinerman-Jones, 2017), modified from Norman & Bobrow (1975).



Depending on the task complexity, the curve might settle early with only 50% of resources allocated to the task (Curve A in Figure 3), could asymptote late, or it may not asymptote even when 100% of attentional resources allocated to the task (Curve B in Figure 3). The resource demand is low when the task is simple, but when the task is complex, it is expected that people will allocate more attentional resources. When performing a task, workload reflects allocated attentional resources to perform the task. As a result, performance measurement is one method used for measuring workload. In many cases, humans do not need to allocate all attentional resources to a single task (see Figure 4). Therefore, performance measure could also be applied by using measurement of spare capacity that is not allocated to a task. The spare capacity is usually measured by secondary task performance (Matthews & Reinerman-Jones, 2017). Performance degradation in the secondary task may indicate an increase in workload of the primary task.

Figure 4
Demand Vs Capacity Graph (Kahneman 1973)



Some tasks may consist of multiple subtasks or components that have different levels of complexity. For instance, in automotive driving, a person must maintain position of the vehicle, monitor traffic, and control the speed of the vehicle. Each component of the task will have its own PRF and gets some percentage of attentional resources allocated to it (Norman & Bobrow, 1975). For instance, monitoring might take 30%, speed control 20%, and maintaining position of the vehicle could take 30% of resources. The total workload of the task then adds up to 80%, which suggests a high workload. This shows that even if components of a task could have low workload, the combined effects of subtask might lead to a high workload. Depending on the task load, a person may allocate more attentional resources to component of the task that has a higher priority. For instance, when driving at an intersection, monitoring might take significant proportion of the resources allocated to the driving task.

Workload Measurement

Workload reflects the interaction between task demand and an effort to utilize human resources available to perform the task. There are different variations to resource theory

(Kahneman 1973, Wickens, 2008; Norman & Bobrow, 1975). Kahneman (1973) for instance suggested that resource allocation and workload increase might be due to different factors such as motivation, engagement with the task, and stress. These factors could determine the operator's willingness or effort to allocate resources to the task. In such case, it could be difficult to differentiate if the increase in workload come from the task itself. Instruments such as the NASA-TLX subjective assessments are designed to distinctly filter out task efforts from stress (Hart & Staveland, 1988).

Multiple Resource Theory defines the relationship between mental workload and performance in multitasking conditions (Wickens, 2008). Multiple resource theory suggests that task performance is associated with the degree of similarity between cognitive resources needed to perform primary and secondary tasks. When tasks are fighting for the same cognitive resources, performance declines. For instance, when people are driving a vehicle which has a high visual processing demand, a visually demanding secondary task could interfere with the task more than an auditory secondary task. Hence it is important to consider the task context when introducing secondary tasks to measure workload. To identify high and low workload subtasks, the foundational resource theory and the PRF model could be sufficient.

Subjective Ratings

Subjective ratings are widely favored as a method of measuring workload due to their cost effectiveness, simplicity, and ease of implementation (Matthews & Reinerman-Jones, 2017). These ratings utilize people's capacity to perceive the level of demand posed by a task they perform. Subjective measurements could be single rating scales where operators are expected to give single rating scale of their perceived workload (e.g., Instantaneous Subjective Assessment, ISA), or multicomponent scales where different sources of workload are measured (e.g., NASA-

TLX, Subjective Workload Assessment Technique), or they could be contextual measures designed for specific domains (e.g., Quantitative Workload inventory, QWI) (Leggatt, 2005; Hart & Staveland, 1988; Spector & Jex, 1998).

Most subjective ratings, exciding the ISA method, are typically administered following a task and may not be well-suited for continuous measurement of workload. On the other hand, the benefit of post-experience responses is that they do not interrupt or affect primary task performance. Other limitations of subjective ratings include vulnerability to biases, and participants' lack of conscious insight into their own mental processes (Matthews & Reinerman-Jones, 2017). In addition to cost, time efficiency, and ease of implementation, subjective workload ratings could also help measure amount of effort that is not directly observed in performance measures (Yeh & Wickens, 1998).

Objective Measures

There are two types of objective metrics used for measuring workload: performance measures and psychophysiological measures. Performance measures capitalize on the theories that posit that workload is a function of the task demand and the availability of internal resources to meet the demand (Matthews & Reinerman-Jones, 2017). As a result, human performance on a secondary task is used as a measure of workload. In many tasks, individuals do not need to utilize all their mental resources to perform a task. As a result, there is unused or spare capacity which could be utilized to perform additional tasks. Because of this, workload is typically measured by introducing a secondary task. For example, in the context of driving, drivers might engage in conversation with passengers while operating a vehicle. As the driving task becomes more demanding, such as at a busy intersection, the driver may stop the conversation, channeling

all attentional resources to the primary driving task. Consequently, decline in performance on the secondary task serves as a gauge for measuring the workload of the primary task.

The advantages of secondary task performance measures lie in their objectivity, alignment with theories of cognitive psychology, and suitability for continuous workload monitoring (Matthews & Reinerman-Jones, 2017). However, limitations exist, including challenges in distinguishing between performance and workload constructs, susceptibility to strategies of resource allocation, and difficulties in standardization (Yeh & Wickens, 1998; Matthews & Reinerman-Jones, 2017). Performance measures can be influenced by resource allocation strategies, posing a challenge when participants allocate resources based on strategies rather than the actual impact of workload.

Psychophysiological metrics utilize the change in aspects of the central nervous system due to variations in task demand (Afergan et al., 2014; Bailey & Iqbal, 2008; Nemeth, 2004, p. #238). With this method, variations of physiological activities are measured and correlated with changes in task demand. These measures exhibit variations in their precision, particularly concerning the identification of active brain areas during tasks (spatial resolution) and the timing of the involvement of these brain areas (temporal resolution) (Eysenck & Keane, 2015, p. #13). The choice of an appropriate tool to use depends on selecting a measurement that is aligned with the research question at hand. Psychophysiological measures offer several advantages, such as objectivity, ability to continuously measure workload, and support from neuroscience theories (Matthews & Reinerman-Jones, 2017). However, these metrics are challenging to implement, complex to interpret, and difficult to standardize.

Now that the theories and measurements of workload have been outlined, it is pertinent to review how automation impacts workload and explore efforts that have been undertaken to

address this issue. The following section will cover some of the existing research on workload and automation.

Workload in Automation

Automation is intended to improve efficiency by releasing some of the load from the operator, freeing their mental capacity to focus on the higher-level internalized schema-based functions of a task (Endsley, 1987). However, in some cases automation diverts the workload of human operators rather than reducing it. For instance, in automated driving, it was shown that monitoring automation increases mental demand compared to manual driving (Stapel, Mullakkal-Babu, & Happee, 2019). One of the reasons is that operators are expected to use increased attentional resources to monitor and manage automated systems. If the operator is not vigilant in monitoring the automation or is engaged in other tasks, they may not be able to take over or even notice the system when malfunction occurs.

In some cases, workload gets reduced in parts of the task where workload is already low (Parasuraman & Mouloua, 1997, p. #97). For instance, flight management systems reduce workload during cruising instead of high workload phases such as take-off and landing. In automobile driving, automations such as cruise control are deployed in low traffic density, highway driving conditions. These raises a question about how automated tasks are selected and implemented.

Endsley (1987) emphasized the importance of task selection as one of the key areas where human factors knowledge could be applied in the development of autonomous systems. The study emphasizes that the automation of systems should not exclusively center on technological capacities and system comprehension but should also incorporate a deep understanding of the human needs. For instance, in cockpit automation, methodologies such as

task analysis were employed to model the sequence of cognitive tasks undertaken by pilots. This approach was used to identify areas where pilots might face cognitive overload and could benefit from automation.

Despite these methodologies, there is a noticeable absence of studies where workload of operators is quantitatively measured to guide decisions on selection of tasks for automation. It seems that design choices for automation primarily aligns with technological advancements (Endsley and Kaber, 1999). In instances where workload of subtasks is measured, it often serves as a trigger mechanism for Adaptive Automation (AA) in high task difficulty scenarios. In addition, such workload measurements are utilized to determine optimal timings for operator notification (Kaber & Endsley 1997; Kaber & Riley, 1999; Afergan et al., 2014; Bailey & Iqbal, 2008; Cosenzo et al., 2009; Muñoz-de-Escalona et al., 2020).

This highlights a prevailing trend where task selection in automation appears predominantly driven by technological advancements rather than being rooted in comprehensive workload evaluations. The question persists: how can we design automation systems that improve efficiency and alleviate workload? And what components of a task should we automate to reduce overall workload? While there are studies that utilize workload measurement for specific purposes such as triggering adaptive automation, a gap remains in directly leveraging workload assessment for informed task selection in the context of automation. To understand this disparity, it is important to first explore existing methods used to reduce workload within the context of automation.

Degrees of Automation and Workload

Various studies have tried to understand how levels and degree of automation affects operator workload, situation awareness, trust, and performance (Parasuraman, Sheridan, &

Wickes, 2008). Onnasch, Wickens, Li, and Manzey (2014) conducted a meta-analysis of 18 experiments to examine the effects of Degree of Automation (DOA) on routine performance, return-to-manual control performance, workload, and situation awareness. Degrees of Automation (DOA) is a term used to define automations that use combination of levels and stages of automation as recommended in Parasuraman et al., (2000, see Figure 2). It is important to state that DOA was not an independent metric with which an automation can be clearly defined. DOA was used to rank order automations reviewed in the study. For instance, a high degree of automation is an implementation of the higher levels and later stages of automation.

The findings indicated there is a benefit of increasing DOA for improving situation awareness and reducing workload when the automation worked as expected (Onnasch et al., 2014). A negative effect of higher degrees of automation on situation awareness was observed when automations failed and the system had to transition to manual control. In addition, negative effects of automation were more apparent when automation transitions between information analysis and action selection. Most of the papers reviewed by Onnasch et al., (2014) showed a negative correlation between workload and Degrees of Automation. However, some studies showed positive correlation or no correlation between subjective measure of workload and DOA.

Overall, effects of degrees of automation on workload, situation awareness, and performance showed a cost-benefit trade-off. Workload and performance benefited from an increase in degree of automation, but situation awareness suffered. Intermediate levels of automation were recommended to provide an optimum choice with respect to primary task performance, workload reduction, and loss of situation awareness was supported. The mixed result on effects of DOA on workload is another piece of evidence that more effort must be made

to examine how workload is affected by automation. It also showed that it is important to consider how a system is automated.

Adaptive Automation

There is a growing body of research on measuring workload in the context of its effect on automation. One such context is the measurement of workload for adaptive automation. In adaptive automation, a system changes its behavior based on user's state. Performance based adaptive automation has been proposed to improve performance, reduce workload, and improve situation awareness (Calhoun, Ward, & Ruff, 2011; Parasuraman, Cosenzo, & Visser, 2009).

Calhoun, Ward, and Ruff (2011) examined image analysis performance on a multi autonomous Unmanned Aerial Vehicle (UAV) control simulation system where automation was adjusted based on operator performance. Participants performed five tasks with and without the application of Adaptive Automation. The image analysis task was a primary task, and the remaining four tasks were used as secondary tasks to measure workload. With Adaptive Automation, the level of automation was adjusted based on average response time score of participants on the secondary tasks. Response times and accuracies were measured on the image analysis task. Participants showed a faster completion time with adaptive rather than static automation. However, accuracy did not show a significant difference between the two automation conditions. The lack of significant effect in accuracy performance was unexpected given the automation had a 100% reliability. Post-experiment subjective responses showed that participants felt the task was less demanding with adaptive automation (Calhoun, Ward, & Ruff, 2011). However, the secondary task performance showed mixed results. Participants showed improved performance with adaptive automation in just one of the four secondary tasks.

Efficacy of Adaptive Automation on Unmanned Vehicles (UVs) was compared in manual, static automation, and adaptive automation conditions (Parasuraman, Cosenzo, and Visser 2009). Participants performed UAV target identification and Unmanned Ground Vehicle (UGV) rerouting, change detection, and call sign acknowledgement tasks. The change detection task performance was used to invoke the automated target recognition system that detected and identified targets. Situation awareness was measured by the call sign acknowledgement task and by responses to verbal situation awareness probes. Overall workload and situation awareness were measured by post experience questionnaires adopted from the NASA Task Load Index (NASA-TLX) and the Cognitive Compatibility Situation Awareness Technique Questionnaires.

In comparison to the manual condition, change detection accuracy and response frequency were significantly higher in both adaptive and static automation. Adaptive automation outperformed static automation in change detection accuracy. However, there was no notable difference in UGV route planning task. Response time for call signs was faster in adaptive automation compared to both static automation and manual conditions, although there was no significant difference between manual and static conditions. More notably, participants reported significantly lower subjective workload with adaptive automation than with static automation and manual conditions. However, there was no significant difference in response to situation awareness between static and adaptive automation.

Research in adaptive automation primarily concentrates on reducing workload and enhancing situation awareness after specific components of a task are automated. However, a significant challenge lies in determining which parts of a task should be automated in the first place. The current focus on adaptive automation appears to be more oriented towards

determining when to activate automation rather than addressing the crucial question of what aspects of a task should be automated.

Dynamic Function Allocation and Task Difficulty Adjustment

Dynamic task Difficulty Adjustment (DDA) is another approach proposed to optimize workload (Afergan et al., 2014). Afergan et al., (2014) measured performance on a path planning UAV task in DDA and static conditions. In the dynamic difficulty adjustment condition, workload was measured in real time using a Functional Near-Infrared Spectroscopy (fNIRS) tool. Depending on the participant workload, the task difficulty was adjusted by adding or removing a UAV. Success was measured by number of UAVs that reached the target without entering a no-fly zone.

There was no significant difference in the measure of success. Participants controlled roughly the same number of aircraft in both DDA and non-adaptive conditions. Participants committed less navigation errors in the DDA than in the static automation condition. Moreover, participants entered fewer no-fly zones in the DDA than the non-adaptive condition. In addition, participants recovered from errors more quicky in DDA condition than they did in the non-adaptive condition. Overall, there seemed to be a performance improvement in terms of error rates and error recovery. However, it is important to note that the purpose of the automation utilized in DDA does not focus on performing a specific component of a task and but rather on managing the overall workload of a system. The approach does not inform us much about what parts of a task should be automated. In complex applications, it will be difficult to experimentally control the difficulty of a task.

Deng et al., (2020) analyzed the effect of automation on Situation Awareness (SA) and workload between static low-level of automation (Static low LOA), Dynamic Automation

Function Allocation (DFA), and Static high-level of automation (Static high LOA) conditions using Unmanned Aerial Vehicle (UAV) control tasks. Participants completed a simulation of military style reconnaissance scenarios. The simulation included an aircraft detection, a fight arrival time calculation, and an alarm handling tasks. The detection task was used as a measure of situation awareness, and the flight time calculation was used as an objective measure of workload. In addition, subjective SA and workload questions were administered between blocks of trials. Automation was applied to the alarm handling task where participants were instructed to select an appropriate control action for emergency scenarios. In DFA the automaton was only applied when the task difficulty was high. In static high LOA, the computer suggested the three most relevant control actions and participants were instructed to choose from the three options. In static low LOA, there was no decision-making automation support.

The results showed that there was no significant effect of mode of automation on subjective SA and workload ratings. Participants were more accurate on the aircraft detection task in the static high LOA condition than they were in the low LOA condition. This indicated a better SA in the static high LOA than the low LOA condition. There was no significant improvement on the aircraft detection task DFA and both static automation conditions. In addition, there was a significant effect of automation on fight time projection accuracy, which was an objective measure of workload. Participants were significantly less accurate in the static high LOA condition than both the Static low LOA and the DFA. This was unexpected because high LOA was expected to reduce workload than static low LOA, and DFA was supposed to reduce workloads than both static conditions. This shows that the approach has mixed outcomes in terms of optimizing workload.

Similar to Adaptive Automation (AA), DFA is a potential method to optimize workload after the decision on task selection for automation is made. Unlike AA, DFA was not adaptive to performance; instead, it is consistently applied when the task difficulty is high. In a controlled experiment, it is possible to control the level of task difficulty. However, such control may not be available in other settings. Regardless of its application, this approach does not inform much about what components of a task should be automated.

Problem Statement

So far, the importance of selecting appropriate tasks for automation has been emphasized, and a gap in research concerning the decision-making process involved in task selection has been highlighted. Methods like adaptive automation, dynamic task difficulty adjustment, and dynamic automation function allocation are valuable tools in deciding when automation should be employed. More importantly, an adaptive approach may help optimize workload and situation awareness. However, such methods provide limited guidance regarding what specific components of a task should be automated. This task selection decision should be based on measurement of the human and system interaction.

The objective of the current research is to formulate a guideline comprising task selection steps that should precede decisions regarding levels and timings of automation. Before decisions are made on how much of the task an automation should perform, there needs to be a guideline on what parts of a task should be automated. Once a task selection decision is made, automaton strategies such as LOAs and adaptive automation could be used to optimize the human automation interaction. Hence, the main goal of the current research was to a investigate the effectiveness of a Workload-based task selection for automation. Three primary questions were formulated to see the efficacy of such an approach.

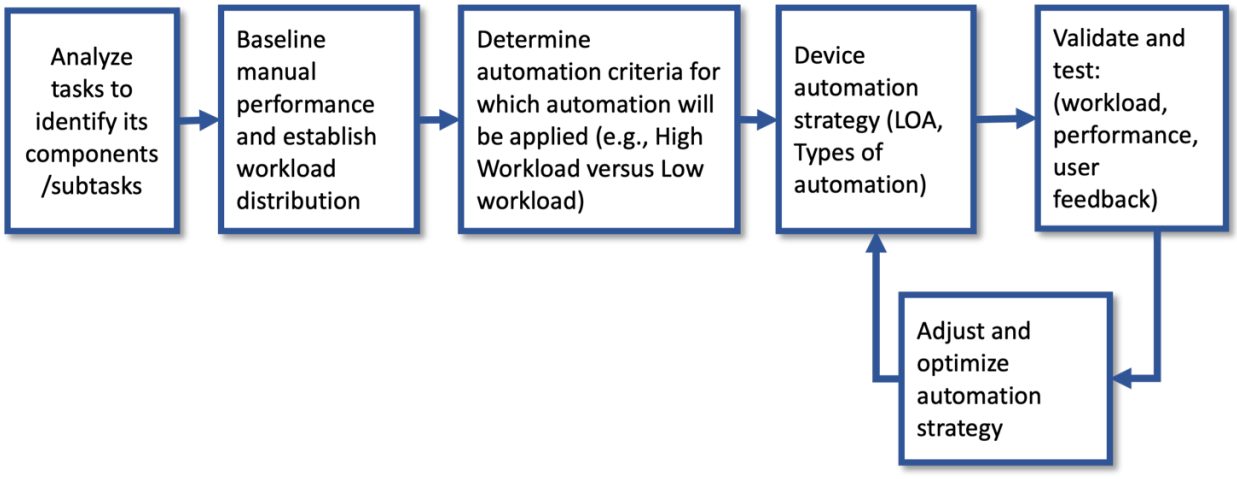
1. Can we select subtasks for automation based on workload measurement?
2. Does workload-based task automation reduce workload?
3. How does workload-based automation affect performance?

Framework for Workload-Based Task Selection

To implement a workload-based task selection, it is first important to understand the characteristics of the task and its components. Hence, the first step is to identify the primary task components of a complex task (see Figure 5). This could be achieved by using methods like task analysis to model the sequence of cognitive tasks undertaken by operators (Endsley, 1987). Once the main components of a task are identified, the next step should be to understand the workload distribution of the task. The workload distribution refers to the mental and physical demands associated with the task across different components or subtasks. Analyzing the workload distribution helps in identifying which components of a task are more demanding, providing insights into what or how to automate aspects of a task for improved performance and efficiency.

Figure 5

Framework for Workload-based task selection. This framework provides a step-by-step guideline to make task selection decision based on measurement of workload of task components.



Establishing baseline performance and determining workload distribution can be achieved by having participants perform a task with no automation and measuring the variations

in workload. This approach has been employed in various contexts before. For instance, Bailey and Iqbal (2008) measured workload variations during execution of goal-oriented tasks to identify points where notifications could be introduced with minimal intrusion. They used pupil dilation measures to identify high and low workload moments of a task. The study showed that notifications are best introduced during moments of low workload, minimizing intrusion during task execution. Workload distribution could be measured by either physiological, objective, or continuous subjective measures.

The third step is establishing automation criteria based on workload characteristics. The decision depends on the intended goal or purpose of the automation. In the second experiment of the current research, subtasks with low and high workloads were automated and subsequently compared. The objective of the research is to assess which automation method leads to the most significant performance improvement. If the primary goal of the automation is to alleviate workload, it is logical to automate the high workload components of a task. Therefore, at this stage, practitioners have the flexibility to decide whether to automate the high or low workloads, depending on the goal of the automation. The key point here is that this stage is where tasks are selected for automation.

Following the establishment of an automation criteria, the next or fourth step involves devising an automation strategy. During this stage, choices regarding the level or degree of automation and the implementation of adaptive or dynamic automation can be made. Different levels and modes of automation can be tested and compared to identify which one leads to substantial performance improvement. Once the automation is implemented, it can be continuously validated and tested in iterative cycles until the desired level of performance is

achieved. This iterative testing and optimization is crucial to understand effects of the automation on the users and tailoring it to meet their needs and expectations effectively.

In the current study, the framework was applied to investigate the efficacy of workload-based task selection in automation. Initially, the task was thoroughly analyzed to identify the main subtasks. Subsequently, in Experiment 1, a workload analysis was conducted to understand the workload distribution within the task. Third, in Experiment 2, specific automation criteria and approaches were selected, implemented, and compared. The automation was further optimized and tested in Experiment 3, based on the insights obtained from Experiment 2.

The Task

The task utilized in this experiment was a decision making and target elimination task named the Theatre Defense Task (TDT). Different versions of this task have been used in various studies. The initial version, developed by Kaber and Endsley (1997), was an individual control task. Later, Bolstad and Endsley (1999) modified the task into a two-person team task. The most recent version, also a two-person task, was adapted by Wright and Kaber (2005). This was the version customized and used in the current research.

The choice of this task was deliberate for several reasons. First, its implementation by Wright and Kaber (2005) occurred within the framework of human-automation collaborative work. They specifically examined the effectiveness of different system function automations in measuring workload and situation awareness, aligning with the objectives of this research. Second, the task complexities and subtasks mirror real life scenarios where tasks encompass components eliciting diverse types and levels of demand. Thirdly, Wright and Kaber (2005) provided the source code of the task and granted permission to modify it for this research.

Measurement Selection

As mentioned in the workload measurement section, various workload measurement techniques have distinct advantages and limitations. However, prior research indicates that these metrics tend to be most effective when employed in combination (Longo & Leiva, 2020; Yeh & Wickens, 1998). In the first experiment of the current research, the primary objective was to gauge workload distribution of subtasks and identify high and low workload subtasks. Subsequently, these subtasks were automated in Experiment 2 and three where performance was compared between different automation conditions. It was important to select a measurement method that is sensitive enough to assess workload of subtasks without impeding task performance.

The utilization of psychophysiological measures posed a challenge due to the task characteristics and the limited expertise in the domain. The Theatre Defense Task, characterized by its fast-paced and dynamic nature, posed a significant challenge in correlating physiological responses to specific subtasks. In the TDT, multiple events happen rapidly, and operators may need to switch between tasks quickly. For instance, a user could be monitoring the radar screen while switching between tasks. Moreover, multiple events happen within fraction of a second. In such cases it would be very challenging to distinguish the precise association between subtasks and physiological responses.

Prior research has demonstrated the effectiveness of employing combinations of two or more workload measurements (Yeh & Wickens, 1998, Hancock & Mathews 2018). In such instances, associations and dissociations may occur. Caution is advised when interpreting dissociations between measurement outcomes. Nevertheless, the use of combined measures of cognitive load can aid in distinguish between different forms of task demand. Specifically, Yeh

and Wickens (1998) suggested that practitioners should combine subjective and performance measures when evaluating performance and choosing between systems.

In the current research, Instantaneous Subjective Assessment (ISA) and performance-based measures were used to measure workload. ISA was chosen over other subjective methods such NASA-TLX due to its ability to continuously measure momentary workload of subtasks (Leggat, 2005, Long & Leiva, 2020). In addition, previous research showed that ISA is minimally intrusive to task performance and is easy to implement (Stevens et al., 2018, p. #245). The combination of ISA with a performance-based measure was believed to provide an effective means of measuring workload associated with subtasks and discriminating performance differences between various automation techniques.

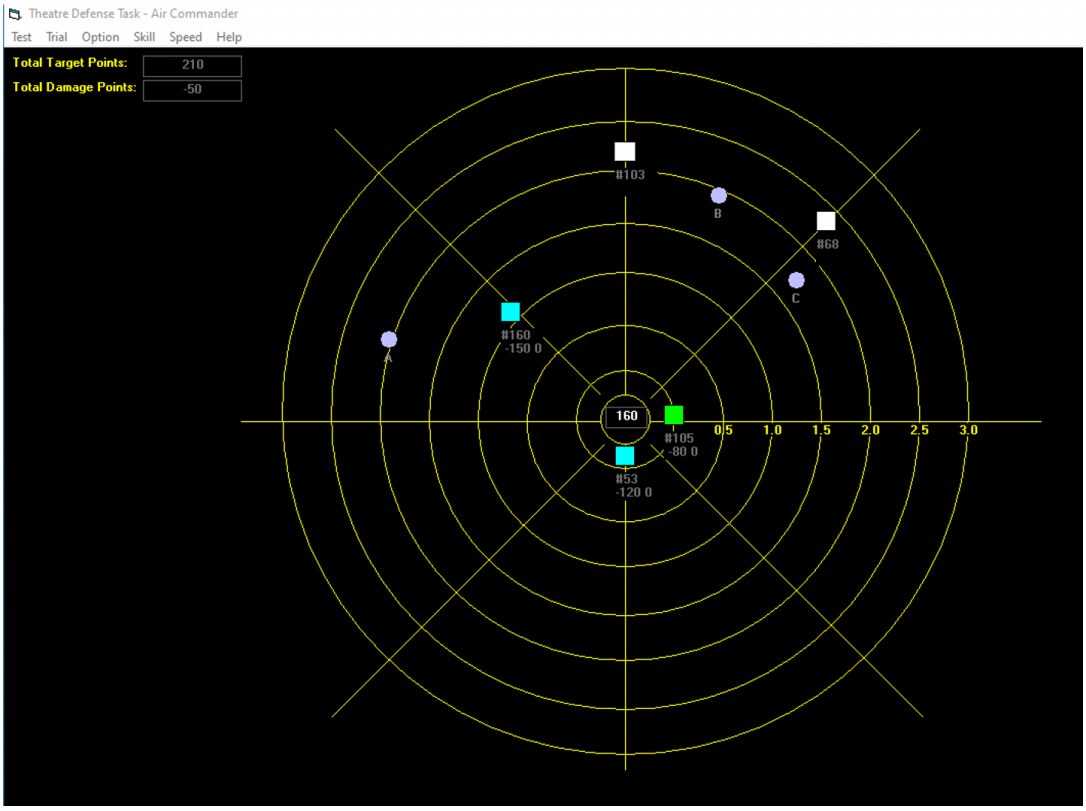
The Theatre Defense Task and its Components

The Wright and Kaber (2005) version TDT was a team decision making and target elimination task that includes two human operators and an automation. The two members of the team represented an Intelligence Officer (IO) and an Air Commander (AC), who worked collaboratively on target elimination tasks. Depending on the condition, different parts of the information officer tasks were automated.

The role of the AC was to protect the base from enemy aircraft, and the role of the IO was to classify incoming targets as enemy or friendly aircraft and notify the AC. The AC monitored the radar screen where aircraft appeared from outside the radar and moved towards the home base (see Figure 6). There were also Airborne Warning Control Systems (AWACS) that traveled around the airspace in a random pattern and reported their prediction of incoming aircraft (see Figure 6). First the AC sent target location and sensor reliability information to the IO.

Figure 6

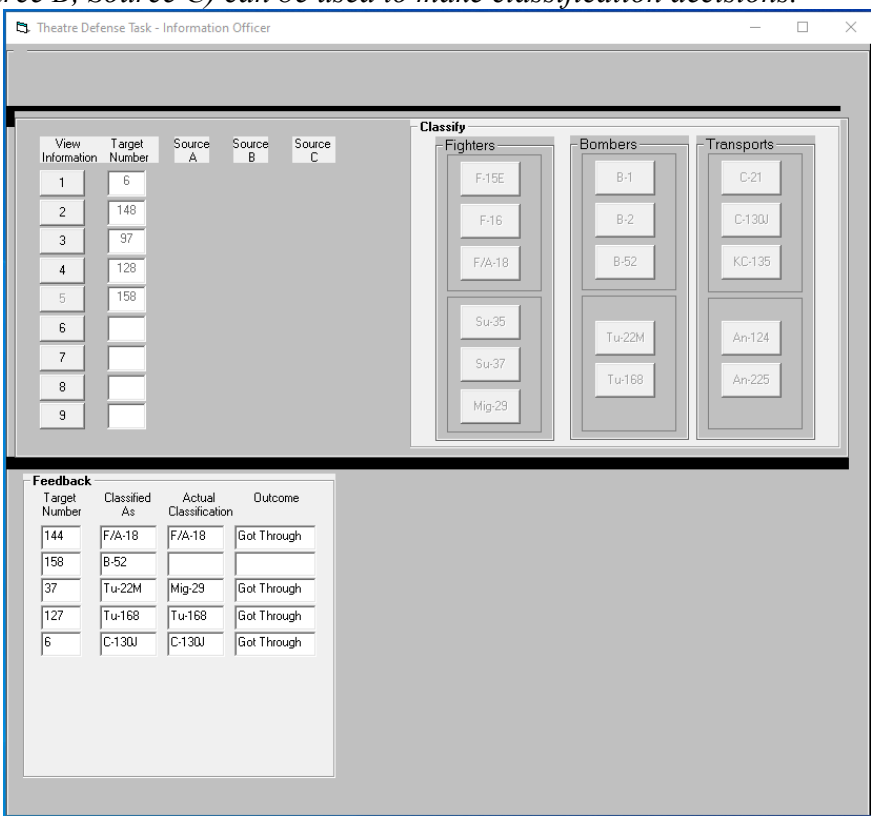
Air commander radar screen. Square objects are moving targets coming towards base, and the circular object marked as A, B, and C are AWACS sensor aircraft. Unclassified object are colored white and classified objects turn to various colors representing aircraft category



Then the Intelligence officer (IO) classified the target as enemy or friendly and informed the Air Commander (AC) about the type of aircraft. The IO made classification decisions based on information they received from the AWACS (see Figure 7). The reliability of the AWACS was dependent on their proximity to the home base at the center of the radar screen. The closer AWACS were to the base, the higher their reliability. Hence, the IO needed to utilize AWACS position information they received from the AC, and account for their reliability. Based on classification from the IO, the AC then decided to either eliminate a target or allowed it to land. The AC needed to use an appropriate missile to destroy enemy aircraft.

Figure 7

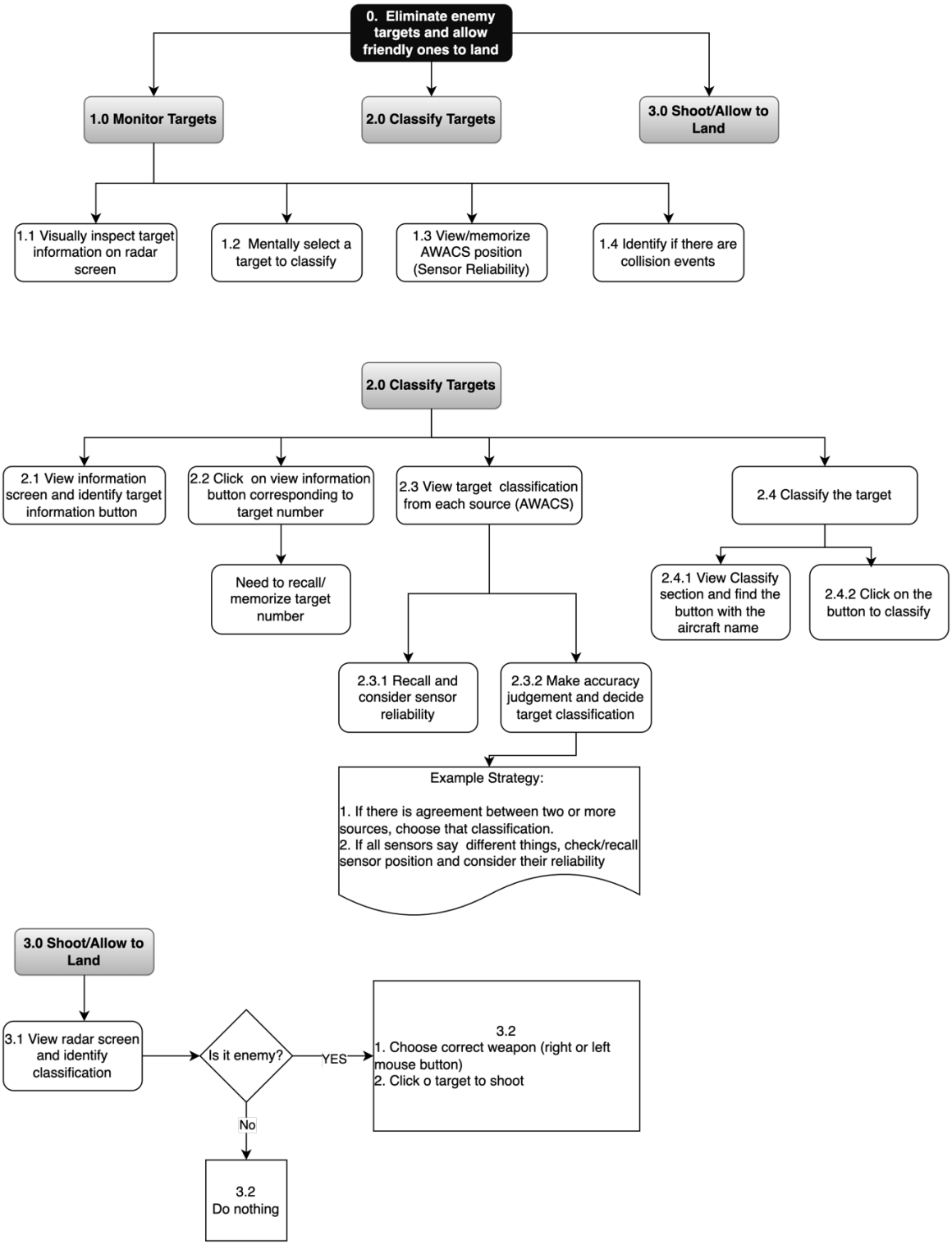
TDT Intelligence Officer (IO) window. Here information about targets from the three AWACS (Source A, Source B, Source C) can be used to make classification decisions.



In the current research, the task was modified into a single-person format where an individual interacted with two windows, undertaking both the AC and IO tasks. This modification was done to create a sufficiently complex task encompassing various subtasks with varying levels of difficulty. A detailed task analysis was conducted to breakdown the task into sequence of smaller subtasks (see Figure 8).

Figure 8

Theater Defense Task (TDT) flow diagram. Similar levels of subtasks are identifiable by highlight and number



The main goal of the task is to protect home base by eliminating enemy targets and allowing friendly targets to land. There were three primary subtasks that the operator needed to perform. The operator is responsible for monitoring the radar screen, classifying targets, and acting on targets based on classification. Once these primary components of the task were established, the subsequent step was to baseline performance and establish the workload distribution. Experiment 1 was specifically designed to fulfil this objective.

Chapter 2

Experiment 1: Baseline Performance and Establish Workload Distribution

The main goal of this experiment was to establish workload distribution of the Theatre Defense Task and identify subtasks that produce high and low workload. The output of this phase was used to automate high and low workload subtasks and compare performance in Experiment 2. Participants performed an air traffic control task named the theater defense task (TDT) (details of the task are explained in the Task section). Workload was measured by subjective responses and performance-based measures. Instantaneous subjective assessment (ISA) questions were administered to participants during the task. Participants rated their perceived workload for each subtask twice per trial. In addition, participants were asked debriefing questions that included a question about which subtask they felt was more demanding.

The theatre defense task included collision scenarios where aircraft could collide before they reach home base. Participants were expected to monitor such situations, identify aircraft types, and avoid or allow collisions. Since this task was not the primary task, it was used as a performance measure for workload of subtasks. It was expected that in high workload scenarios, participants would concede more collisions. Subtasks that participants were engaged with during collision events were marked and compared.

Method

Participants

The number of subjects needed for the study was determined by an *a priori* power analysis using G*Power for an ANOVA with a significance level of .05. A total of 28 subjects was sufficient to detect a medium sized effect of 0.25 with 80% power (see Table 3). There were 31 participants recruited for the study. Participants were recruited from Rice University

undergraduate subject pool, and they were compensated with credit towards a course requirement. There were 20 female and 11 male participants with age ranging from 18 to 21 (M = 19.19).

Table 3
Power Analysis

	Power			
Cohen's *f*	.8	.85	.9	.95
.10	163	184	213	259
.25	28	31	36	43
.40	12	13	15	18

Task

The task used in this experiment was the TDT task explained in the preceding section. The workload analysis here focused on three primary subtasks: Monitoring, classifying, and action (shooting). As stated in the previous section, the main goal of the task was to eliminate enemy aircraft and allow friendly ones to land. The operator received positive points for eliminating enemy targets and got penalized (earned negative points) if they destroyed a friendly aircraft. The number of points assigned depended on the type of enemy aircraft.

During the task, there were scenarios where some targets collided with each other. Depending on the colliding aircraft type, the operator could gain or lose points. If two enemy aircraft collided, the highest target point of the two aircraft was awarded. Participants could also maximize their gain by shooting both enemy aircraft and earning target points for both aircraft. If two friendly aircraft collided, then no points were accrued. However, if a friendly and an enemy aircraft collided, the operator received double the penalty points allocated for allowing an enemy

aircraft to land. Hence, it was in the operator's best interest to shoot the enemy aircraft before it collided with a friendly aircraft. The task was designed to allow the operator to make strategic decisions.

Task difficulty was determined by the number of targets present on the radar screen at a time. Different combinations of number of targets and target movement speed were tested during a pilot study to determine high and low difficulty conditions. As a result, the number of targets for low and high difficulty conditions were set to 4 and 6 targets respectively. The speed of the aircraft motion was set to level 2 in all conditions. Each level of speed has a range with minimum and maximum values. Hence, some targets could move faster than others within the range of speed in a level.

Design

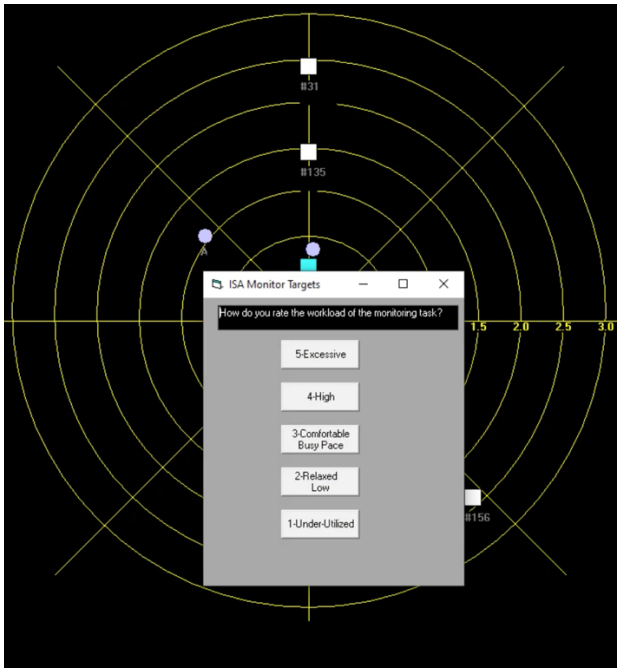
A 2 x 3 within-subjects design was used in this experiment, with two levels of task difficulty and three subtasks representing 3 within-subjects levels. The order of task difficulty was randomized where half of the participants performed the high difficulty task first, while the other half started with low task difficulty. It was not feasible to randomize the presentation of subtasks due to potential disruption in the natural task process. For instance, if the operator did not monitor the radar first, accurate target classifications could not be made. Subsequently, the ability to take appropriate action on the target would be hindered.

The workload of the subtasks was assessed using a combination of objective and subjective measures. The objective measure was based on collision monitoring and avoidance, with collision avoidance serving as a secondary task to measure workload. There were two measures of collision: the effect of different subtasks on collisions and collision rate. Collision rate was calculated as a ratio of number of collisions incurred to the total number of possible

collisions. The effect of different subtasks on the chances of collision was measured by observing participants' activities during the event. At any given moment during a task, participants were either monitoring events, classifying targets, or taking actions on targets. Subtasks that kept a participant's attention away from handling collision events were registered and counted. The main goal of this measurement was to find out the effect of different subtasks on participant's workload.

The second workload measure used was a subjective measure named Instantaneous Subjective Assessment (ISA). Instantaneous subjective performance is a common subjective measure of workload in various domains, including air traffic control (Muñoz-de-Escalona, Cañas, Leva, & Longo, 2020; Leggatt, 2005). Operators responded to five-point scale questions presented to them at regular intervals while they were performing tasks (see Figure 9). ISA is minimally intrusive, flexible, and requires minimal resources, which makes it a useful tool to measure operators perceived workload of subtasks in real time (Stevens et al., 2018, p. #245).

Figure 9
ISA Rating Prompt



The recommended timing of ISA is every two minutes (Stevens et al., 2018). However, the TDT task is dynamic and multiple events occur within two minutes. Timings of half a minute, one minute, and 90 seconds were evaluated in the pilot study. Participants stated that 30 second was too fast and intrusive to the task. As a result, ISA questions were administered every 60 seconds and two responses were collected for each subtask per trial.

Materials

A windows desktop computer that ran the TDT software was used to present the task to participants. The computer was connected to two monitors the AC displayed in one, and the IO on the other. The two windows were placed at the right and left corners of the two screens and placed at the center where the two screens meet (see Figure 10). This was done to minimize left and right head movements by participants. Interaction with the software was performed using a mouse where participants classified and took actions on targets by clicking left and right mouse buttons. ISA questionnaires were presented as prompts on the radar screen every minute (see Figure 9). Prompts remained on the screen until participants responded to them, and the task was suspended. Participants were trained on the task for fifteen minutes and each trial lasted 7 minutes. The timing of the practice session was determined by a previous study that showed 15 minutes was the duration needed to produce an asymptotic performance (Kaber & Riley, 1999).

Figure 10
Experiment 1 Monitor setup

Results

For repeated measures ANOVAs, Huynh-Feldt correction was applied when the assumption of sphericity was violated ($\varepsilon < .9$). In such cases, fractional degrees of freedom will be reported.

Subjective Workload Measures

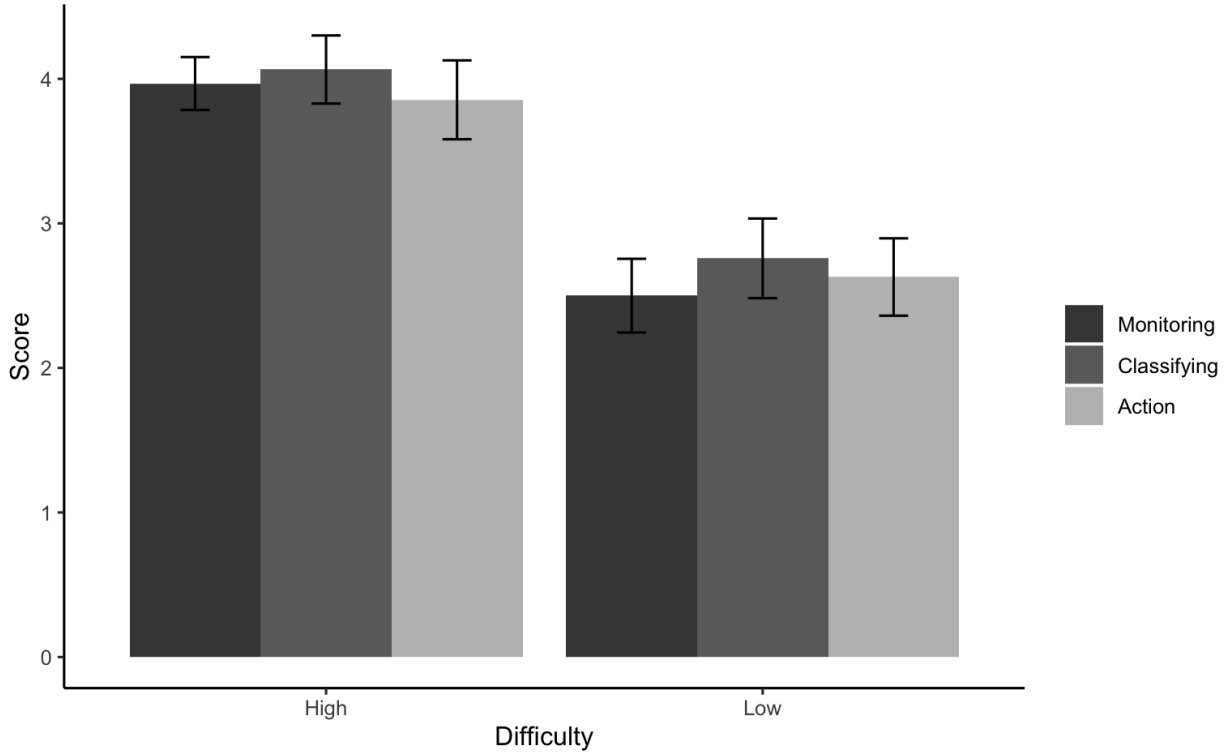
Instantaneous Subjective Assessment (ISA)

A 2 x 3 repeated measures ANOVA was used to compare workload ratings of the three subtasks in two levels of task difficulty. The ratings were on a scale for 1 to 5 where a rating of 1 indicated a very low task demand where participants felt under-utilized and a rating of 5 meant the task was excessively demanding or overwhelming. One participant's data was discarded

because they stated that they did not correctly perform the task and accurately responded to the rating questions. As a result, data from 30 participants was used to perform the analysis.

Figure 11

Instantaneous Subjective Assessment (ISA) scores of Subtasks. Error bars represent 95% CI.

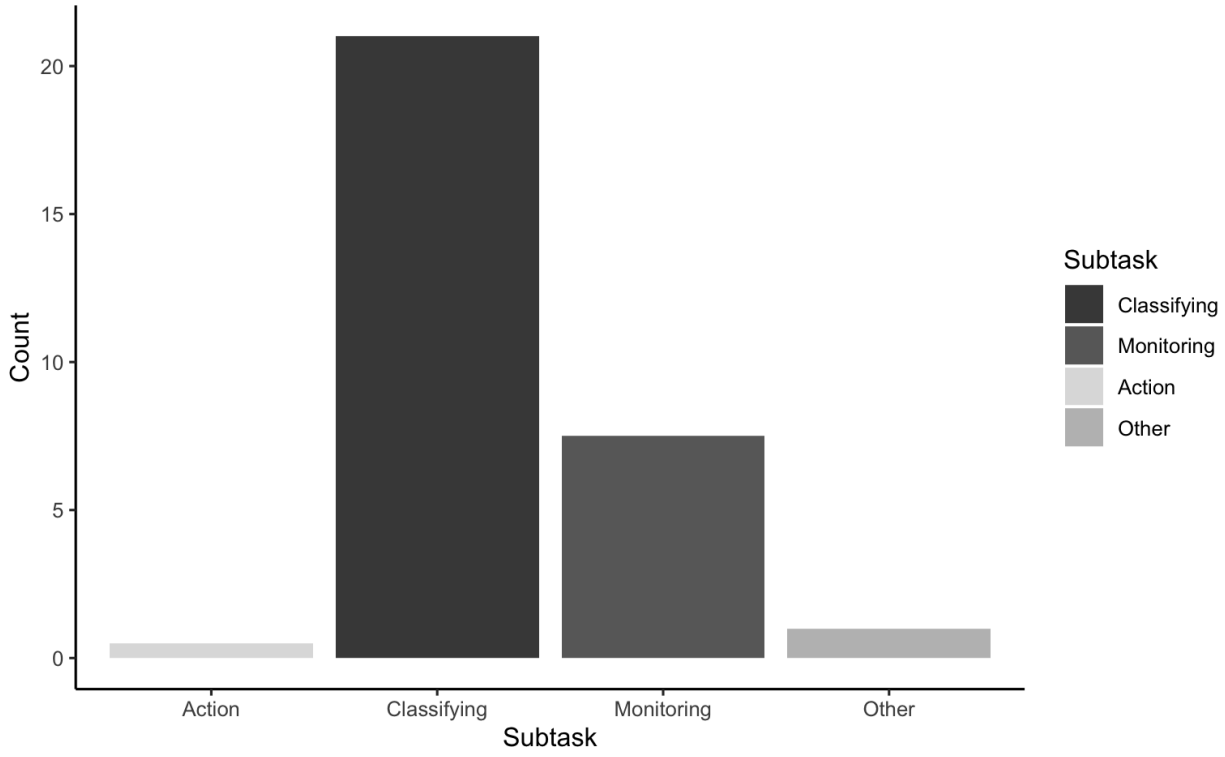


There was a significant main effect of task difficulty $F(1, 29) = 184$, $MSE = 0.43$, $p < .001$, Cohen's $f = 0.56$ (see Figure 11). The main effect of subtask was not significant $F(1.66, 48.14) = 2.58$, $MSE = 0.273$, $p = .096$, Cohen's $f = 0.01$. The interaction between subtask and task difficulty was also not significant $F(2.00, 58.00) = 1.42$, $MSE = 0.167$, $p = .25$, Cohen's $f = 0.01$. Even though the ISA measure was sensitive to task difficulty, it did not clearly illustrate workload distribution of the subtasks.

Subjective Response (Post-trial)

The post-trial subjective response showed that participants felt classifying was the most demanding subtask followed by monitoring (see Figure 12).

Figure 12
Post-Trial Subjective Response



For analysis, the subtask variable was classified into four levels with classifying, monitoring, action, and other (see Figure 12). In cases where participants selected two subtasks, a half point (0.5) count was given for each of the subtasks. For instance, when participants respond that both monitoring and classifying were difficult, the response was split as 0.5 for each subtask (see Table 4).

Table 4
Frequency of participant responses

	Monitoring	Classifying	Action	Task Switching
Frequency	7.5	21	0.5	1

A chi-square goodness of fit test was performed to determine whether the proportion of subtasks was equal between the four groups. The proportion of ratings differed by Subtask, X^2(3,

30) = 36.47, $p < .001$. The results showed that participants perceive classifying as the most demanding subtask and action (shooting targets) as the least demanding subtask. The monitoring subtask showed an intermediate level of difficulty.

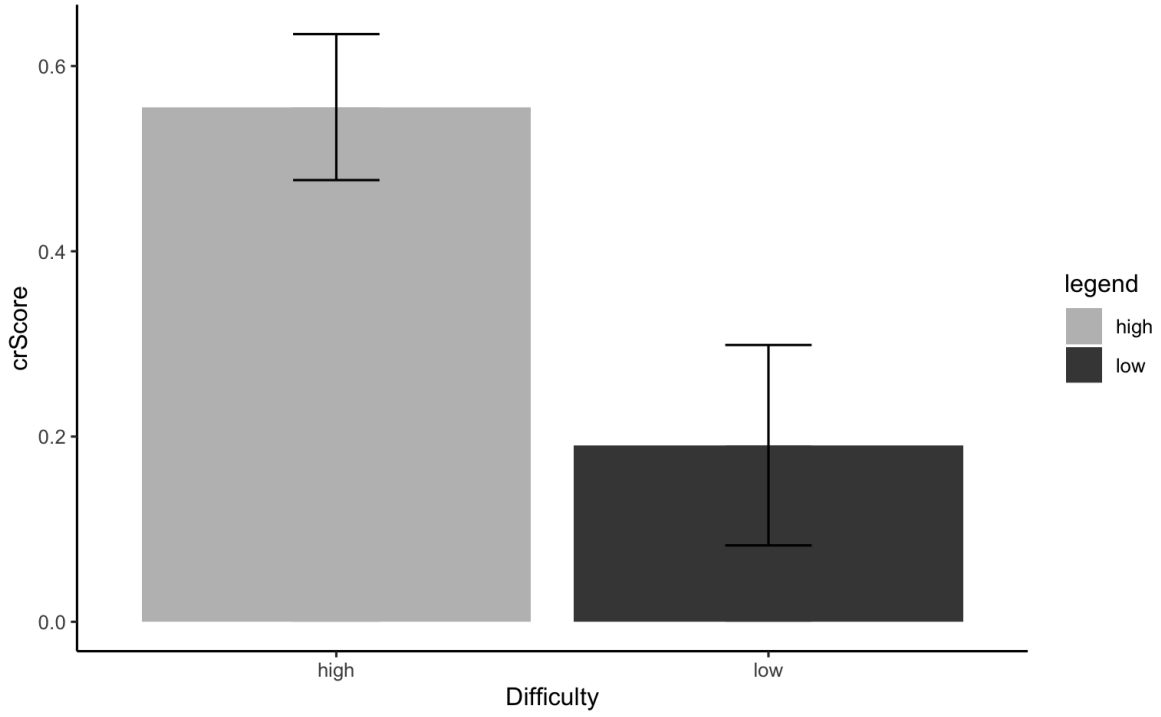
Collision Performance

Collision Rate

Collision rate was used to compare performance between high and low difficulty conditions. A higher rate of collision was observed in high task difficulty condition than in the low difficulty condition (see Figure 13).

Figure 13

Collision Rate by task difficulty. Error bars represent 95% confidence interval.



A paired-samples *t*-test indicated that collision rate scores were significantly higher for the high level task difficulty than for the low level task difficulty, $t(29) = 5.68$, $p < .001$, $d = 1.04$.

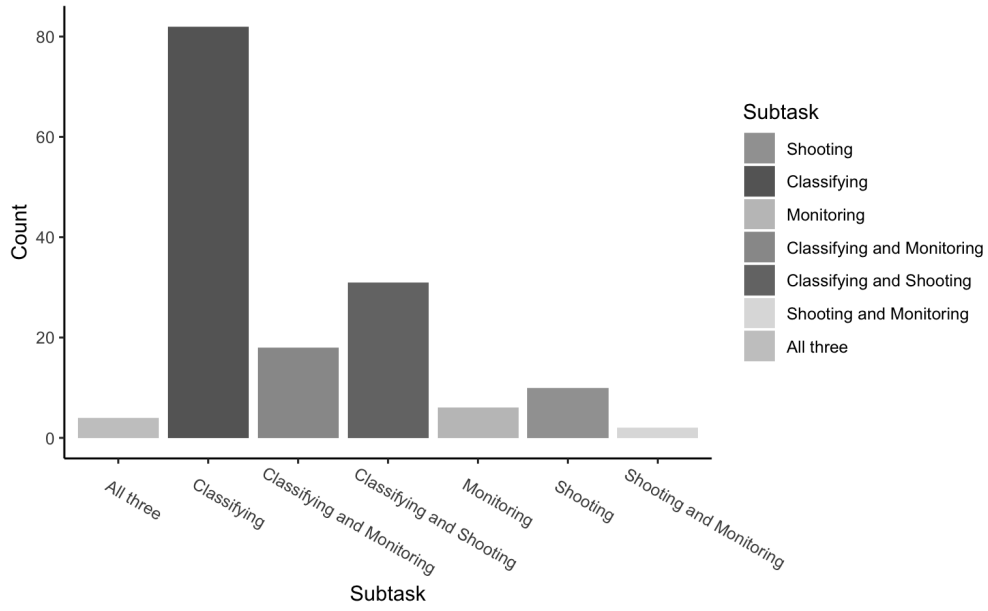
Subtasks and Collisions

One participant's trial was not screen recorded which limited the ability to categorize subtasks that captured participants attention during moments of collision. As a result, data from 29 participants were used for analysis. There was a total of 153 collisions that occurred across all trials of 29 participants. Subtasks that participants were engaged with during moments of collision were reviewed and counted by two researchers independently. Counts from the two researchers were compared. It is important to note here that monitoring task was very difficult for proctors to discriminate from other tasks. Monitoring events were recorded as moments when participants were neither actively shooting nor classifying. However, participants could be monitoring events while they were performing other tasks.

Initially, there were seven categories of subtasks and subtask combinations that participants were engaged with during moments of collision (see Figure 14). Participants were predominantly occupied with the classifying subtask, followed by action (shooting) and monitoring. There were various overlaps where participants could be performing multiple tasks during events of collision.

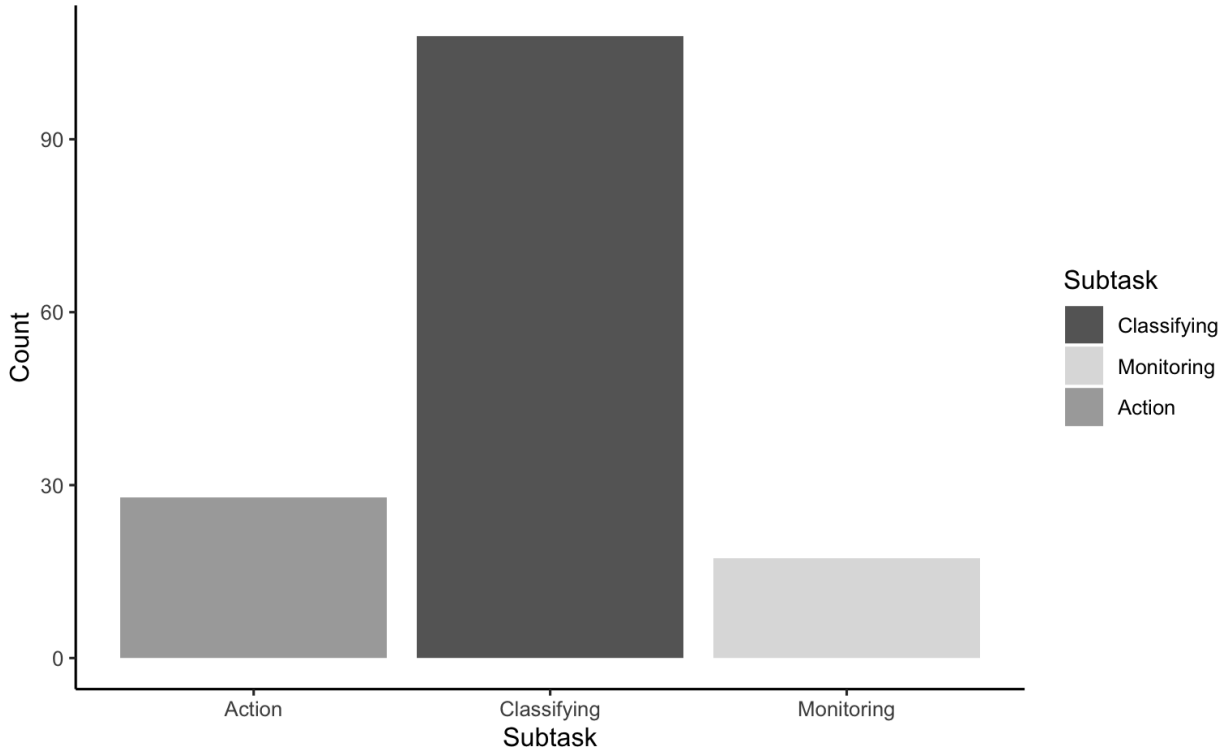
Figure 14

Subtasks Participants Were Performing During Collision



The main goal of this experiment was to identify the workload generated by the three subtasks. Hence, the overlaps were divided and assigned to either of the three subtasks. For instance, if participants were performing three tasks, the total number of such events divided by 3 was assigned to each subtask. If there were two tasks involved, then half count was assigned to each subtask. The summarized outcome shows that participants were performing classifying subtasks during approximately 108 collision events, monitoring during 17 events, and were shooting targets during 28 events (see Figure 15)

Figure 15
Modified Subtasks Engagement During Collision Events



A chi-square goodness of fit test was performed to determine whether the proportion of subtasks was equal between the three subtask groups. The proportion of ratings differed by Subtask, $X^2(2, 29) = 96.08$, $p < .001$. The results indicated that participants were engaged with the classifying subtask during most of the collision events. In other words, classifying was the most demanding subtask for participants. Contrary to the post-experience responses, action subtask appeared to be slightly more demanding than monitoring task.

Discussion

The goal of this experiment was to illustrate the workload distribution of the monitoring, classifying, and action (shooting) subtasks. The Instantaneous Subjective Assessment (ISA) measurement was sensitive to task difficulty but did not show a significant difference between subtasks. One reason could be that participants were unable to discriminate their perceived

workload due to each subtask. The other reason could be that participants were affected by the overall task demand when they respond to rating questions specific to subtasks. Third, participants may have responded to the overall task demand instead of the specific subtasks. Previous studies also showed that subjective responses could be dominated by the total task resource demand on working memory (Yeh & Wickens, 1988).

Results from the post-trial subjective responses and the collision measurement showed that classifying was the most demanding subtask. Although the effect of subtask was subtle in the ISA score, the outcome of the combined measures from post experience response, and collision performance implied that classifying has the highest workload. The difference between action and monitoring subtasks seems to vary across measurements. In the subjective responses, participants rated monitoring subtask higher than the action task.

Analysis from the collision performance showed the action subtask was more demanding than the monitoring subtask. This could be due to the method subtask engagement was scored. Monitoring events were recorded as moments where participants were neither classifying nor shooting. However, participants might have been monitoring while performing classification or shooting tasks. As a result, the effect of monitoring subtask on collisions might have been undercounted. Participant's responses when asked which task was more demanding clearly showed that monitoring was perceived as a more demanding subtask than shooting. In addition, when participants were asked why they felt classifying was more demanding, some of them stated that it was because they must monitor at the same time. Some participants also stated that they had difficulties monitoring collisions when the task was more difficult.

The combined results led to the conclusion that classifying was the most demanding subtask, followed by monitoring as the second most demanding, and action as the least

demanding subtask. After determining the workload distribution, the next step was to determine criteria for applying automation. The primary objective of this research was to investigate the impact of workload-based automation on task performance and workload. As a result, the chosen automation criteria were to automate both high and low workload subtasks and compare their performance.

In practical applications, practitioners can decide which types of tasks to automate based on their specific goals. However, for this study, both the high workload task of classifying and the low workload subtask of action were selected for automation. In addition, to minimize the effect of monitoring on the classifying subtask, an automation support was integrated to the monitoring component of the task. In experiment 2, high and low workload automations were validated and tested to determine which subtask automation yielded the most significant performance improvement.

Chapter 3

Experiment 2: Device Automation Strategy, then Validate and Test

The goal of this experiment was to identify which type of automation improves performance and reduces workload. Performances were compared across different subtask automation conditions. The experiment included two types of subtask automation and two levels of task difficulty variables. The action subtask was automated as the low workload subtask and the classifying subtask was automated as the high workload subtask. In Experiment 1, it was observed that the difficulty conditions were less demanding than expected. This was expected to be alleviated with the addition of automation support. As a result, different settings were tested in pilot study and the number of targets for low and high difficulty conditions were changed to 5 and 7 respectively.

Automation Strategy

Research indicates that intermediate level automations are most effective in maintaining a desirable operator situation awareness (SA) and increasing team efficiency (Endsley & Kaber, 1999). While situation awareness was not the primary focus of this research, it remains an integral consideration among the three human factors constructs that predict human-automation performance (Parasuraman, Sheridan, & Wickens, 2008). System design must be strategically positioned in the range of desirable levels to ensure optimal outcomes. From the workload perspective, both overload and underload are not desirable. Previous studies have shown that an intermediate level of automation strikes a balance in keeping human operators sufficiently engaged with a task, while offering enough autonomy to enhance overall task performance. An intermediate level automation is between levels 4 to 6 (see Table 2). As a result, an intermediate level of automation named Blended Decision Making (level 5) was selected for this experiment.

In addition, this type of automation was also previously used in automaton and teamwork research (Wright & Kaber 2005). Details of the automaton's functions in the high and ow workload conditions is explained in the following section.

Method

Participants

The number of subjects needed for the study was determined by an *a priori* power analysis using G*Power for a repeated measures ANOVA with a significance level of .05. A total of 24 subjects was sufficient to detect a medium sized effect of 0.25 with 80% power (see Table 5). There were 33 participants recruited for the study. Participants were recruited from Rice University undergraduate subject pool, and they were compensated with credit towards a course requirement. There were 21 female and 12 male participants with age ranging from 18 to 21 (M =19.03).

Table 5
Power Analysis

	Power			
Cohen's *f*	.8	.85	.9	.95
.10	138	156	179	216
.25	24	26	30	36
.40	10	11	13	15

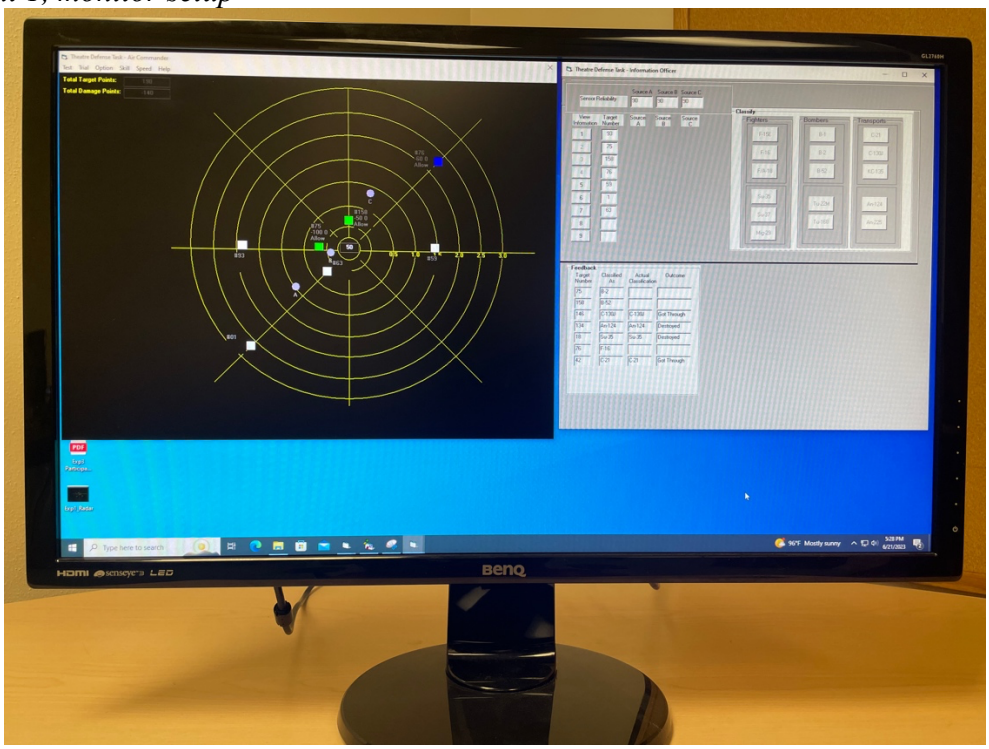
Task

The task used in this experiment was the same as Experiment 1 with two types of automations applied in different conditions. Hence, in this section, I only explain the changes from the first experiment. In Experiment 1, some participants stated that it was difficult and

tiresome to perform the task on two monitors. As a result, the task was changed to fit into one monitor with radar and information windows presented side by side (see Figure 16). The frequency of ISA prompt presentation minimized to reduce frustration of participants and to encourage attentive responses. As a result, ISA questions were presented once every two minutes and each subtask was rated once per trial.

Figure 16

Experiment 1, monitor setup



Subtask Automation

There were two types of automation used in this experiment, Classification Automation and Action Automation. The level of automation used for the subtasks was Blended Decision Making (BDM) which represented a system that provides high-level decision support and could make decisions (Wright and Kaber, 2005; Endsley & Kaber, 1999). In BDM, automation is applied to the information analysis and decision selection of the four-stage human information processing model (Parasuraman et al., 2000; Endsley, 2017). The automation analyzes information, selects a decision, and presents that information to the operator. The automation did not act on the decision it made and the human operator could agree with or override the automated decision.

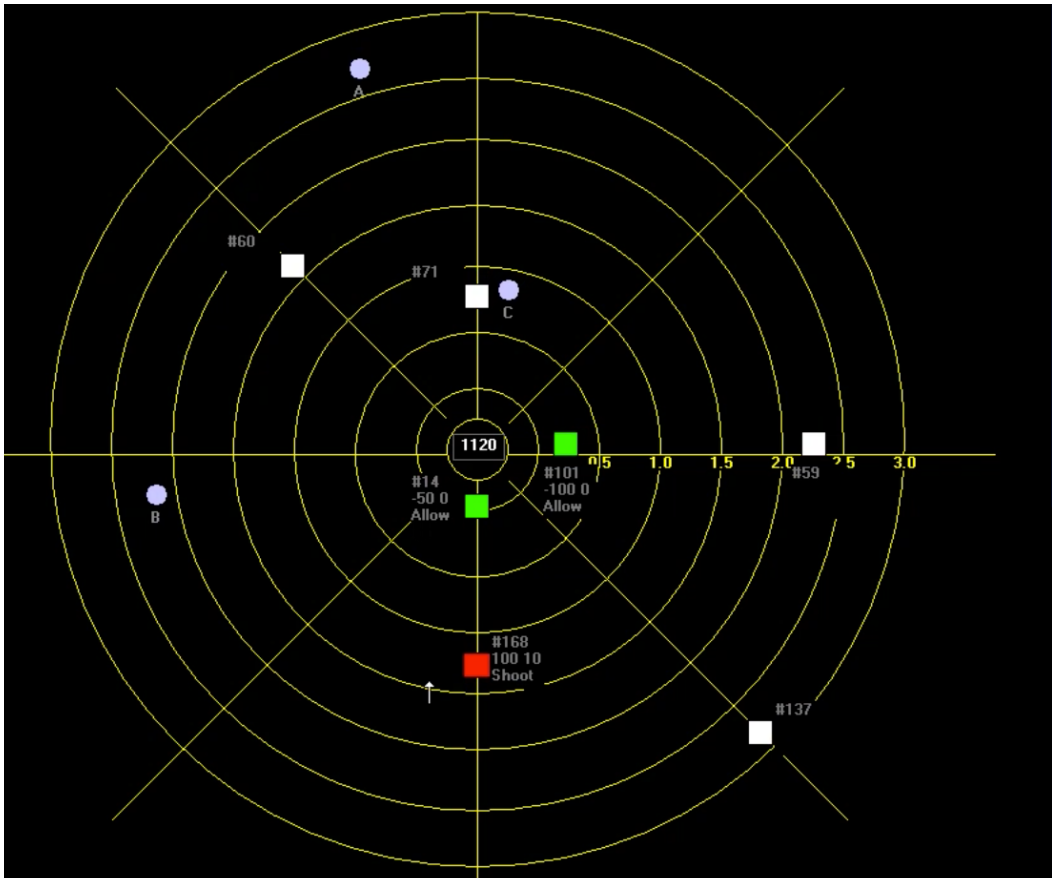
Action Automation

Action Automation was intended to support the operator making decisions on whether to shoot a target or allow it to land. This was managed in two forms.

1. The reward and penalty points for allowing or shooting targets was presented on the target label (see Figure 17). This information was based on classification made by participants, and it was intended to help the human make better shooting decisions when they were unsure about their classifications. Hence, participants were expected to make decisions about shooting based on how confident they were about the classification decision, and the reward/penalty tradeoff of shooting or allowing a plane to land. This part of the automation was also used in Wright and Kaber (2005) and was termed as Action Support.
2. The automation provided recommendations about shooting a target or allowing it to land. Each classified target had a text that informed the human either to shoot a plane or allow it to land (see Figure 17). This was more useful when the reliability of all three AWACS were the same, but their classifications of targets were different. For instance, if all the AWACS are at a 30% zone and they all classified targets differently, then classification reliability was low because the target could be any of the three aircraft. The automation always checked the sensor reliability information, the agreement between the sensor classification, the classification by the human, and the reward-penalty tradeoff to inform the human if they should shoot a target or allow it to land.

Figure 17

Action Automation Radar Screen View. Classified targets, green and red objects, had labels that include reward and damage points, and automation recommendation to allow or shoot the target.



Classification Automation

The goal of the Classification Automation (CA) condition was to help make classifying decisions. This was the same automation used in Wright and Kaber (2005) and was named Blended Decision Making (BDM). In their study, only one subtask was automated, and the name BDM did not create any ambiguity. However, in this experiment, BDM was used for different subtasks, and automations were named based on the automated subtask. Similar to Action Automation, Classification Automation supported the human operator in two forms.

1. The automation made target classification decision and informed the human operator (see Figure 18). This information was presented on the "Auto" tab on the right side of

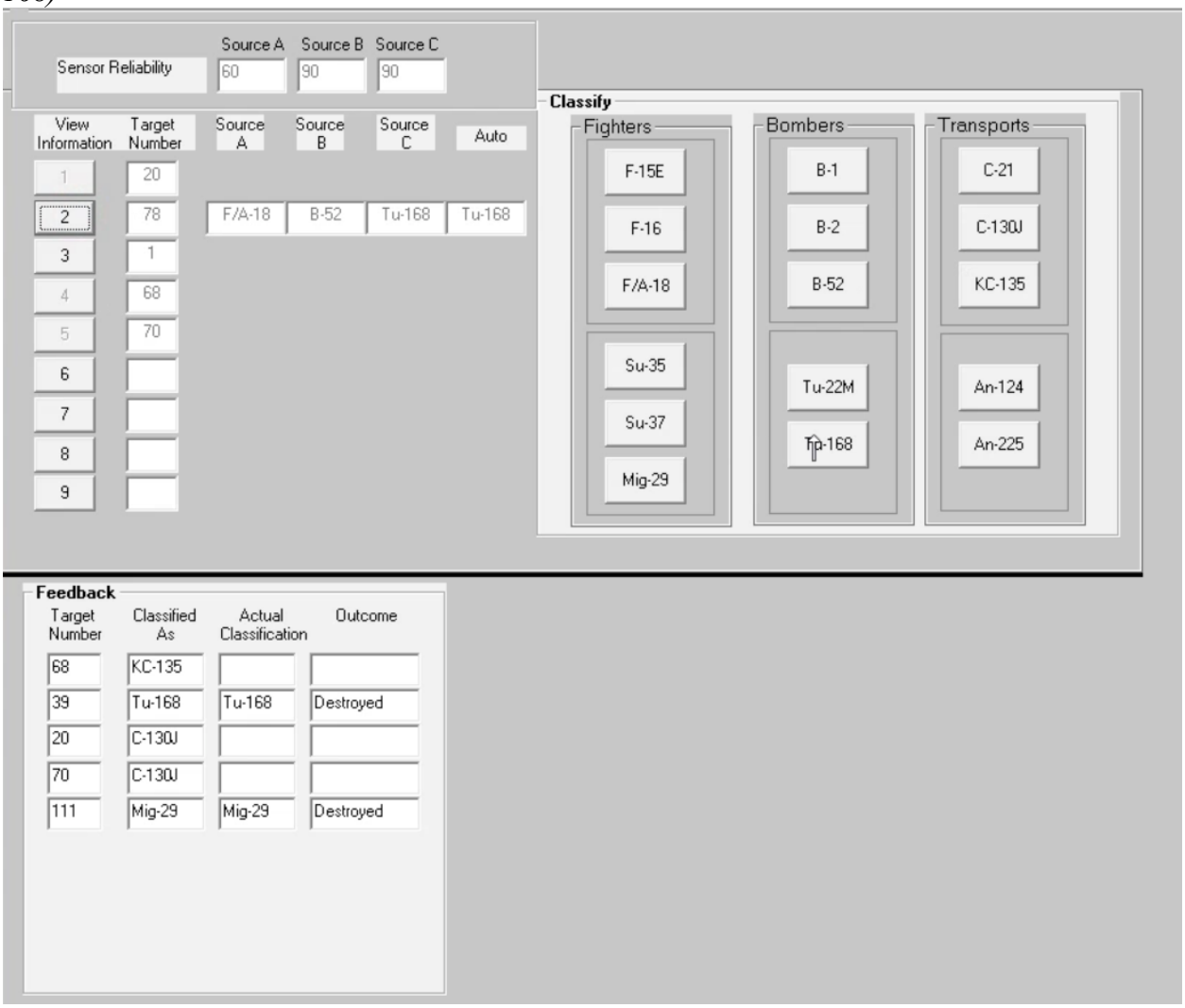
the information from the three AWACS. The automation made classification decisions based on sensor information, how reliable sensors were, and agreement between the sensor classifications. The human could agree with the automation and classify the target as the automation suggests, or they could override the automation classification.

2. The automation sorted the list of targets in the information window (View information tab in Figure 18) based on classifications from sensors and the target distance from the home base. The automation prioritized what it perceived as enemy planes at the top of the list. In other words, the list was sorted with the nearest enemy targets at the top of the list.

In both automation conditions, there was a monitoring support automation. This automation supported human operators by displaying sensor reliability information on the information screen. The sensor reliability information was presented on top-left corner of the information screen, above the "View information" area (see Figure 18). The sensor reliability was dynamically updated as the AWACS moved around the air space. Instead of tracking and memorizing the sensor positions on the radar screen, operators could focus on target position information. The rationale for adding a monitoring support for automation is to limit the effect of working memory load on subjective responses. When resources are dominated by the demand of the task on working memory, subjective measures may be affected (Yeh & Wickens, 1988). In addition, the monitoring automation support could help filter out the effect of automating high and low workload subtask by reducing one of the sources of working memory demands.

Figure 18

Classification Automation Information Screen. The "Auto" tab to the right of "Source C" tab shows the automation decision (e.g., The Automation would classify target number 78 as a Tu-168)



Design

A 2 x 2 x 3 within-subjects design with two levels of task difficulty, two types of automation, and three subtasks was conducted. There were four conditions of automation type and task difficulty pairs: Classification Automation-High task difficulty (CAHigh), Classification Automation-Low task difficulty (CALow), Action Automation-High task difficulty (AAHigh), and Action Automation-Low task difficulty (AALow). To counterbalance any order effects, the

participants performed sequence of conditions in Latin square. Similar to Experiment 1, it was not feasible to randomize the presentation of subtasks due to potential disruption in the natural task process. Overall net score, target processing and shooting score, classification accuracy, and collision performance scores were used as dependent variables. Each of these measurements are explained in the following section.

Net Score

Assessment of overall performance was measured by the final net score participants earned during the task. This score was calculated from target score, damage score, and the total number of points attainable per trial. Target score was calculated by summing the positive points gained for shooting down enemy aircraft and subtracting the negative scores incurred from shooting friendly aircraft. Damage score was a result of adding the penalty points received for allowing enemy aircraft to land unharmed. In the high-difficulty conditions, a greater number of targets were present, providing more opportunities to accumulate points. Consequently, the net score utilized in the analysis was derived by dividing the final score (Target score minus damage score) by the total possible points attainable per session.

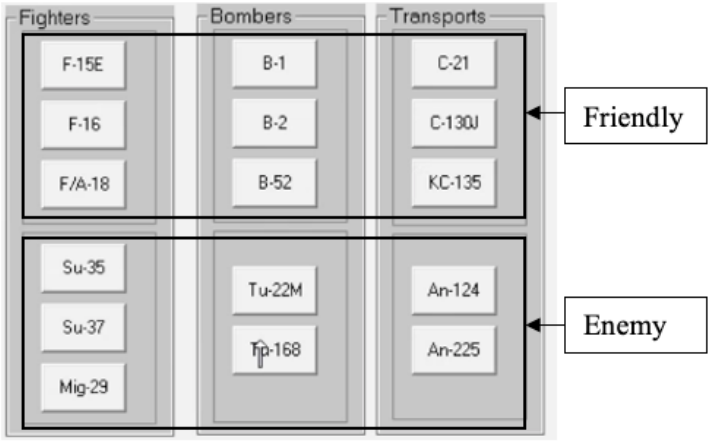
$$Net\ Score = \frac{Target\ score - Damage\ score}{Total\ possible\ points\ attainable\ per\ session} \quad (1)$$

Classification Performance

Net score performance was intended to measure overall performance differences between Classification and Action Automation conditions. However, more measurements were needed to understand how different types of automation affected performance. It was important to understand which part of the task performance was improved due to the automation support. As a result, classification performance was analyzed to assess the classification part of the task.

Classification performance was analyzed using two measures: proportions classified, and classification accuracy. The first measure used was to understand if there were differences in the proportion of targets classified. This measure was a ratio of number of targets classified by the total number of targets processed per trial. There were 15 types of aircraft in two categories of friendly and enemy (see Figure 19). Classification accuracy was measured in two ways. First, classifying accuracy was measured by specific match between participant's classification and the aircraft. In this measure an accurate classification was when participants classified targets by their specific aircraft name. For instance, enemy bomber Tu-22M had to be classified as a Tu-22M to be counted as an accurate classification.

Figure 19
Aircraft Categories on IO Screen



The second classification accuracy measure used was classifying by category (Friendly or Enemy). This measure was used because it was observed that some participants classified targets as enemy or friendly without considering the exact aircraft. In other words, participants for example classified an enemy bomber as an enemy fighter and shot it down. They still received the reward/target point for shooting an enemy aircraft regardless of the classification accuracy. Participants may have perceived that classifying by category, they could save some time earned by avoiding looking for the specific aircraft name. In this measure, accurate classification was

when a target was classified within the same category regardless of the specific aircraft. For instance, if a friendly bomber B-1 was classified as a friendly transport C-21, it was considered as an accurate classification.

Number of Targets Processed and Shot

In the TDT task the more quickly participants classified and acted on targets, the more targets they could process within a trial. Shooting down an enemy aircraft or allowing a friendly plane to land would initiate an addition of a new target to the radar screen. For friendly targets, participants waited for the targets to land before another target was added to the screen. However, with enemy aircraft, they could speed up the process by classifying and shooting the aircraft quickly. Hence, number of targets processed per trial was an indicator of how fast people processed targets and could show action performance differences between conditions. Number of targets shot was also an indicator of participants shooting performance.

Workload Measures

Similar to Experiment 1, workload of the subtasks was measured by Instantaneous Subjective Assessment (ISA) ratings, post-trial subjective responses, and collision performance. Based on lessons from Experiment 1, the post-trial subjective questions were administered after each trial. Post-trial questions were intended to help understand participants perceived workload ratings, and the reasons for their decision.

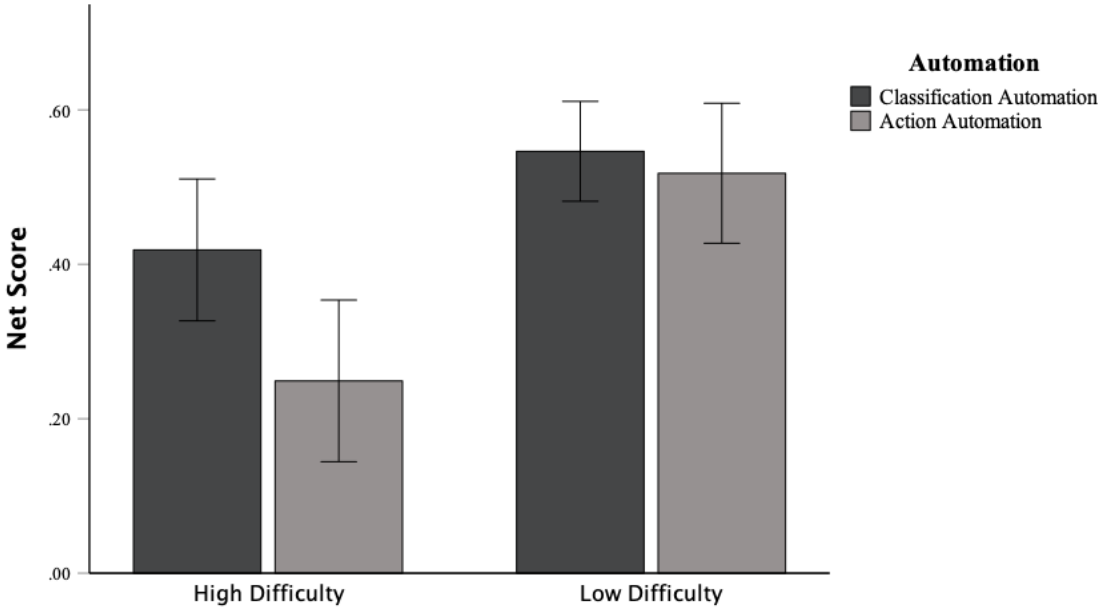
Results

Net Score

A 2 x 2 repeated measures ANOVA was used to compare overall performance of two automation conditions in two levels of task difficulty. The net score was calculated by Equation 1. Data from a total of 32 participants was used in this analysis.

Figure 20

Net Score Performance in Two Automation and Two Difficulty Conditions. Error bars represent 95% confidence interval.



Net score performance showed that participants performed much better on the Classification Automation condition than in the Action Automation condition when the task difficulty was high (see Figure 20). When task difficulty was low, effect of automation was not significant. There was a significant interaction effect between automation type and task difficulty subtask and task difficulty $F(1, 31) = 5.22$, $MSE = 0.031$, $p = .029$, Cohen's $f = 0.15$. To get more insight on the interaction, a pairwise analysis was performed.

A paired-samples *t*-test indicated that scores were significantly higher in Classification Automation than in Action Automation, $t(31) = 3.6$, $p = .001$, $d = 0.64$, in the high task difficulty condition. In the low task difficulty condition, scores were not significantly different between Classification Automation and Action Automation, $t(31) = .54$, $p = .30$, $d = 0.10$. There was a significant main effect of automation type $F(1, 31) = 6.39$, $MSE = 0.049$, $p = .017$, Cohen's $f =$

0.17. The main effect of task difficulty was also significant $F(1, 31) = 17.15$, $MSE = 0.073$, $p < .001$, Cohen's $f = 0.38$.

Classification Performance

Net score performance showed that Classification Automation improved performance relative to Action Automation when the task difficulty was high. To gain insight into elements that contributed to the performance improvement, further analysis was done comparing classifying performances.

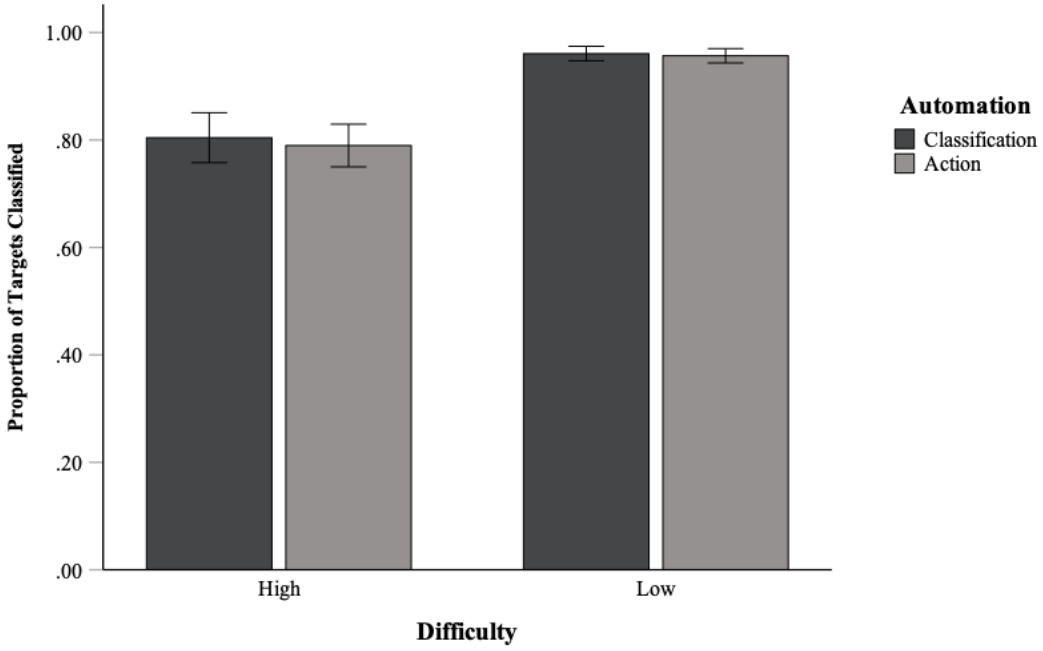
During the task it was observed that one participant developed a clever strategy to maximize their performance. The strategy was classifying enemy targets and neglecting to classify friendly targets. In other words, the participant viewed sensor information about targets and only classified the targets if they saw it as an enemy aircraft. It appeared that the participant intended to save time by skipping a step in the task. The rationale was that if a target was friendly there was no action to take on it but to allow it to land. Leaving it unclassified did not have a clear consequence for the participant. As a result, the participant focused on enemy aircraft which they had to classify to take actions on. As a result, data from the participant was discarded in the classification performance analyses. Data collected from 31 participants was used in the analysis of classification performances.

Proportion of Targets Classified

A 2 x 2 repeated measures ANOVA was used to compare classifying performance of two automation conditions and two levels of task difficulty. The goal here was to see if there were performance differences in percentage of targets between conditions.

Figure 21

Proportion of Targets Classified. Error bars represent 95% confidence interval.



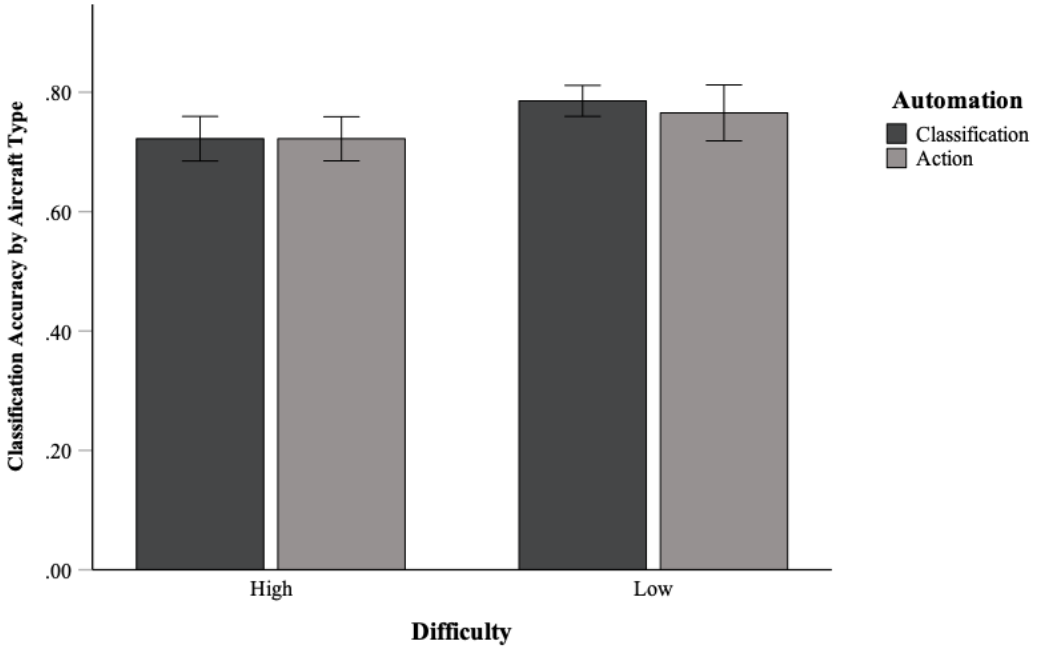
There was a significant main effect of task difficulty $F(1, 30) = 87.54$, $MSE = 0.009$, $p < .001$, Cohen's $f = 1.12$. The main effect of automation type was not significant $F(1, 30) = 1.16$, $MSE = 0.002$, $p = .29$, Cohen's $f = 0.04$. In addition, the interaction between automation type and task difficulty was not significant $F(1, 30) = 0.53$, $MSE = 0.002$, $p = .47$, Cohen's $f = 0.02$. The main difference in this classifying performance was only observed between high and low difficulty conditions (see Figure 21). Participants classify a lower percentage of targets in the high task difficulty condition than they do in low difficulty condition.

Classifying by Aircraft

A 2 x 2 repeated measures ANOVA was used to compare classifying performance of two automation conditions and two levels of task difficulty. The classifying score was a proportion score with number of matches as a numerator and the total number of targets as a denominator. An accurate classification was when a classification was an exact match to the aircraft name.

Figure 22

Classifying Accuracy by Aircraft (Exact Match). Error bars represent 95% confidence interval.



The main difference in this classifying performance was observed between high and low difficulty conditions. Participants were less accurate in the high task difficulty condition than they were in the low difficulty condition (see figure 22). There was a significant main effect of task difficulty $F(1, 30) = 10.6$, $MSE = 0.008$, $p = .003$, Cohen's $f = 0.27$. The main effect of automation type was not significant $F(1, 30) = 0.24$, $MSE = 0.013$, $p = .63$, Cohen's $f = 0.01$. In addition, the interaction between automation type and task difficulty subtask and task difficulty was not significant $F(1, 30) = 0.38$, $MSE = 0.008$, $p = .54$, and Cohen's $f = 0.012$.

Classifying by Category

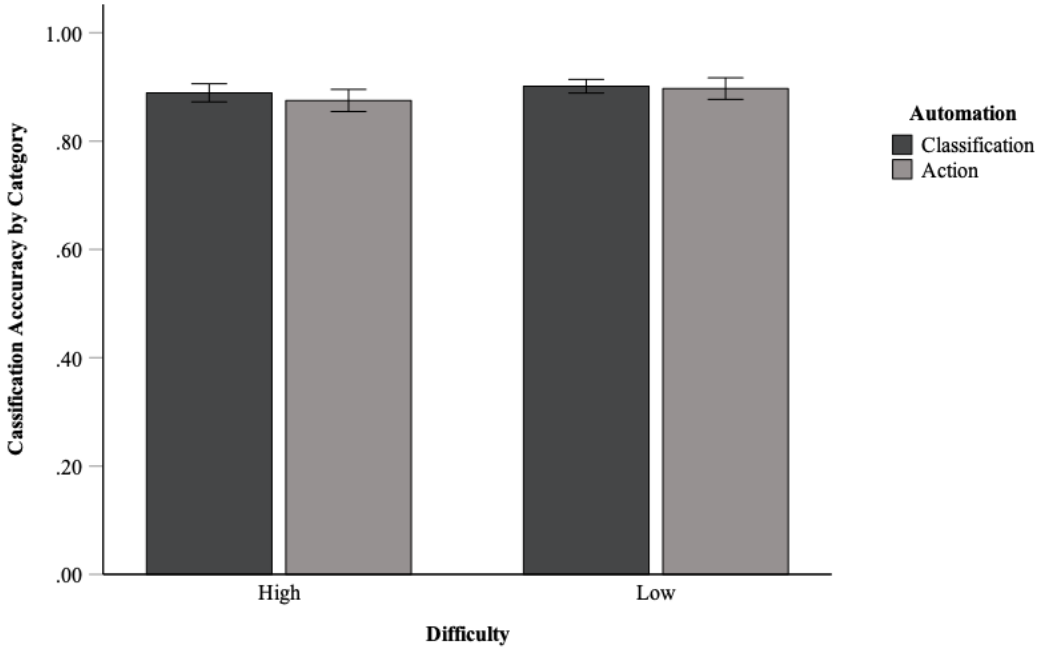
The main goal of the task was to eliminate enemy targets and allow friendly planes reach homebase. As a result, participants were observed classifying by group and saving time than matching the aircraft type. If a target was enemy, they should shoot it down regardless of what type of enemy it was. The caveat of this strategy was that participants often used the wrong weapon to shoot targets. This effect may not be big because participants were able to quickly

alternate between weapons and shoot targets. As a result, it was important to review *classification performance by category* performance.

A 2 x 2 repeated measures ANOVA was used to compare classifying performance of two automation conditions and two levels of task difficulty. The classifying score was a proportion score with number of matches as a numerator and the total number of targets as a denominator.

Figure 23

Classifying by Category Score (Enemy as Enemy, and Friendly as Friendly). Error bars represent 95% confidence interval.



The main difference in this classifying performance was observed between high and low difficulty conditions. Participants were less accurate in the high task difficulty condition than they were in the low difficulty condition (see Figure 23). There interaction was not significant $F(1, 30) = 0.448$, $MSE = 0.002$, $p = .51$, and Cohen's $f = 0.02$. There was a significant main effect of task difficulty $F(1, 30) = 4.66$, $MSE = 0.002$, $p = .039$, Cohen's $f = 0.14$. In addition, the main effect of automation type was not significant $F(1, 30) = 0.926$, $MSE = 0.003$, $p = .34$, Cohen's $f = 0.03$.

Subjective Workload Measures

Instantaneous Subjective Assessment (ISA)

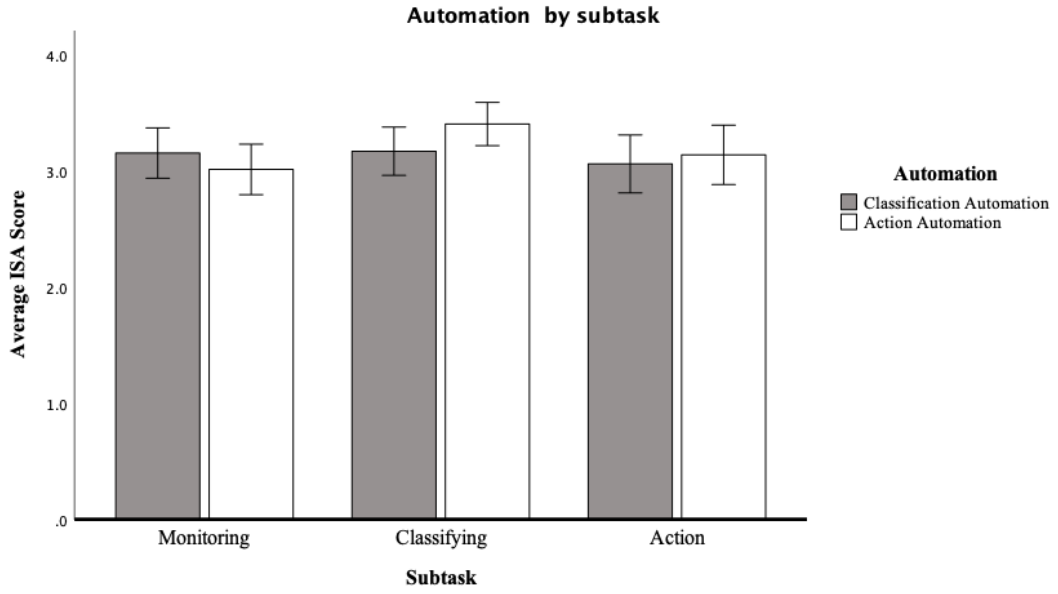
A 2x2x3 repeated measures ANOVA with two levels of task difficulty, two types of automation, and three subtasks was conducted to measure the change in workload of subtasks due to automation. Responses from a total of 32 participants was used for this analysis. Huynh-Feldt correction was applied when the assumption of sphericity was violated ($\varepsilon < .9$).

There was no three-way interaction of automation, task difficulty, and subtask $F(1.9, 61.8) = 0.05$, $MSE = 0.194$, $p = .947$, Cohen's $f < 0.01$. There was also no interaction effect of automation and task difficulty $F(1, 31) = 0.45$, $MSE = 0.708$, $p = .509$, Cohen's $f = 0.01$. The interaction effect of difficulty and subtask was significant $F(1.5, 47.5) = 4.71$, $MSE = 0.383$, $p = .021$, Cohen's $f = 0.13$. There was also an interaction effect of automation and subtask $F(1.9, 60.6) = 4.89$, $MSE = 0.238$, $p = .011$, Cohen's $f = 0.14$.

First, to evaluate the interaction effect of automation by subtask, pairwise comparisons were carried out between subtasks within and across automation conditions (see Figure 24).

Figure 24

Automation by subtask interaction. Error bars represent 95% confidence interval.



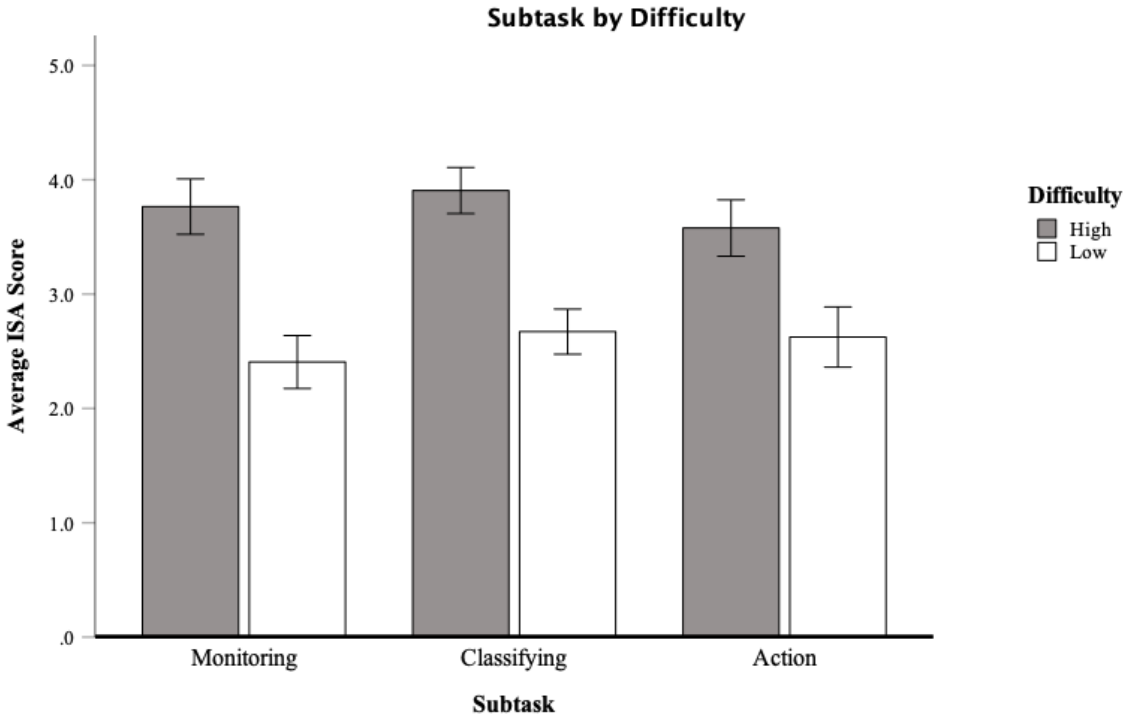
A paired sample *t*-test showed that the classifying subtask was rated as less difficult in Classification Automation than in Action Automation, $t(31) = 2.4$, $p = .023$, $d = 0.42$. Within Action Automation, the classifying subtask was rated as significantly more difficult than both the monitoring, $t(31) = 4.7$, $p < .001$, $d = 0.83$, and the action subtasks, $t(31) = 2.24$, $p = .033$, $d = 0.4$. Whereas within Classification Automation, there was no significant difference between the subtasks. In addition, the main effect of automation on subtask ratings was not significant $F(1, 31) = 0.54$, $MSE = 0.59$, $p = .47$, Cohen's $f = 0.02$.

These results showed that perceived workload of the classifying subtask decreased in Classification Automation condition (see Figure 24). The classifying subtask was perceived as less difficult in CA than in AA, and it was perceived as a more difficult subtask than the monitoring and the action subtasks in the AA condition. There were no significant rating differences between subtasks within the Classification Automation condition.

To evaluate the interaction effect of task difficulty by subtask, pairwise comparisons were made between subtasks within and across levels of task difficulty (see Figure 25).

Figure 25

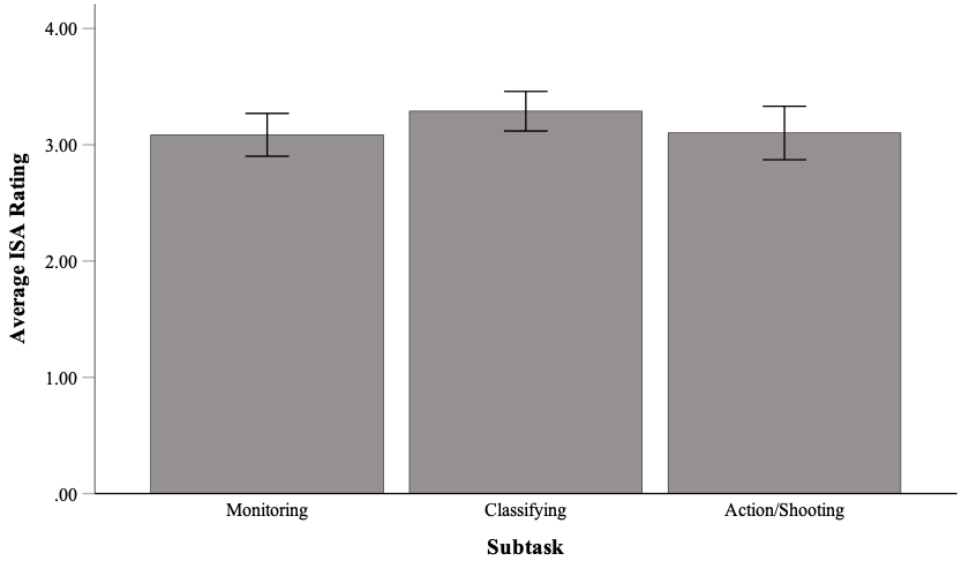
Subtask by Difficulty Interaction. Error Bars Indicate 95% Confidence Interval



A paired sample *t*-test showed that the classifying subtask was rated as a higher workload subtask than the action subtask when the task difficulty was high, $t(31) = 3.01$, $p = .005$, $d = 0.53$. The classifying subtask was also rated to generate a higher workload monitoring when the task difficult was low, $t(31) = 2.96$, $p = .006$, $d = 0.52$. The main effect of task difficulty was significant, $F(1, 31) = 166$, $MSE = 0.82$, $p < .001$, Cohen's $f = 1.57$.

Figure 26

ISA ratings across subtasks. Error Bars Indicate 95% Confidence Interval.



The main effect of subtask on workload ratings was significant $F(1.5, 48.4) = 4.05$, $MSE = 0.52$, $p = .033$, Cohen's $f = 0.12$ (see Figure 26). A paired sample t-test showed that the classifying subtask was rated as a higher workload subtask than monitoring, $t(31) = 3.91$, $p < .001$, $d = 0.69$. The difference in workload rating between the classifying and the action subtasks was not significant, $t(31) = 2.04$, $p = .050$, $d = 0.36$. In addition, difference in rating between the monitoring and the action subtask was not significant, $t(31) = 0.18$, $p = .86$, $d = 0.03$.

Subjective Response (Post-trial)

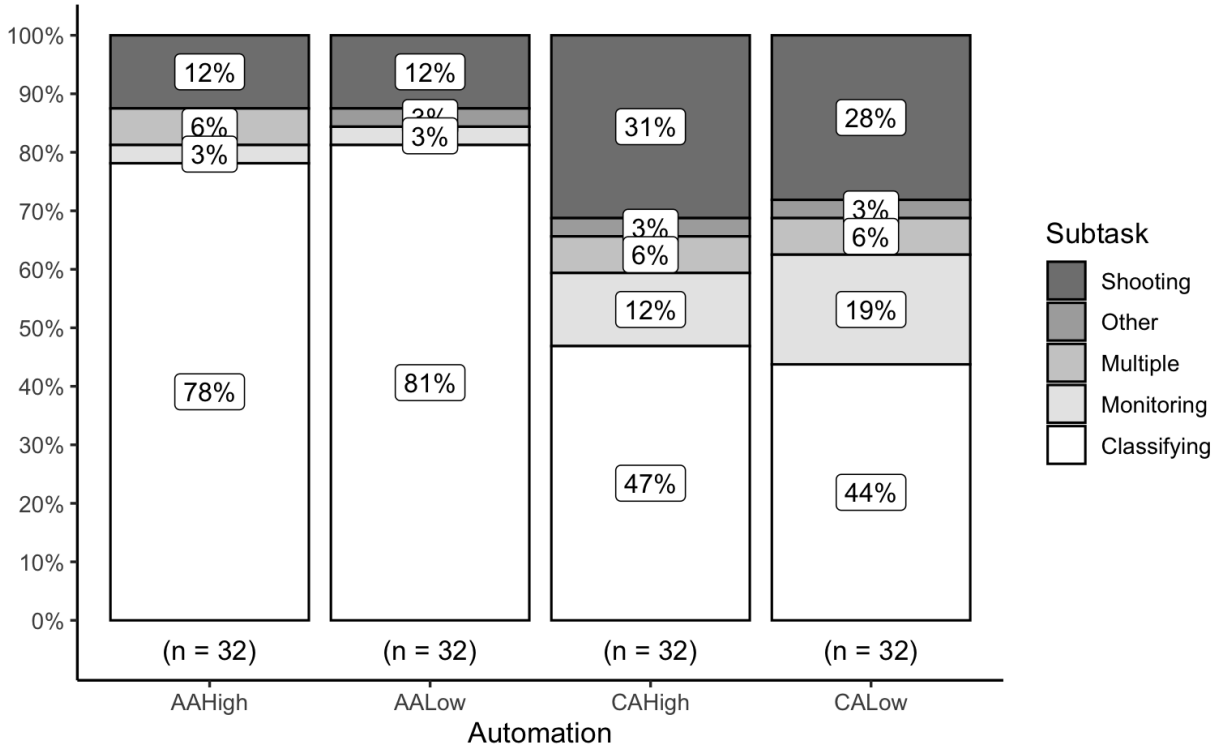
Subjective responses were collected after each trial, four times per participant. Participants were asked to choose what they felt was the most demanding subtask in a trial. A total of 128 responses were collected and 118 of the cases were single responses of either monitoring, classifying, or shooting. There were three cases where participants gave nonspecific responses and were categorized as "other" responses.

The distribution of responses indicated that the classifying subtask was perceived as the most difficult subtask across all conditions. However, there appeared to be a change in

perception where the classifying subtask was rated as less demanding in the Action Automation condition than in the Classification Automation condition (see Figure 27). It was also interesting to see that the transition comes with an increase in the perceived demand of the action and monitoring subtasks. The action subtask seemed to be perceived as more demanding in Classification Automation than in Action Automation. In addition, the monitoring subtask was also considered as more demanding in Classification Automation than in Action Automation.

Figure 27

Frequency of participants' post trial selection of most difficult subtask by automation and task difficulty conditions. (e.g., CALow is Classification automation-low task difficulty, AAHigh is Action automation-high task difficulty condition).



Fisher's exact test was performed to assess the relationship between automation conditions and Subtasks (p = .026). There was a significant association between automation conditions and subtasks. In other words, people's perception of subtask difficulty varied depending on the automation-task difficulty conditions. Monitoring and shooting (action) subtasks were perceived as demanding more frequently in the Classification Automation

condition than they were in Action Automation condition. In contrast, classifying subtask was less frequently rated as the most difficult in Action Automation than it was in Classification Automation. This makes sense because when the classifying subtask was automated, participants seemed to spend more time monitoring and acting on targets. In other words, Classification Automaton freed up some of the time for participants to engage more with other subtasks. As a result, participant spent more time engaging with the action and monitoring subtasks in Classification Automation condition than in the Action Automation condition.

Collision Performance

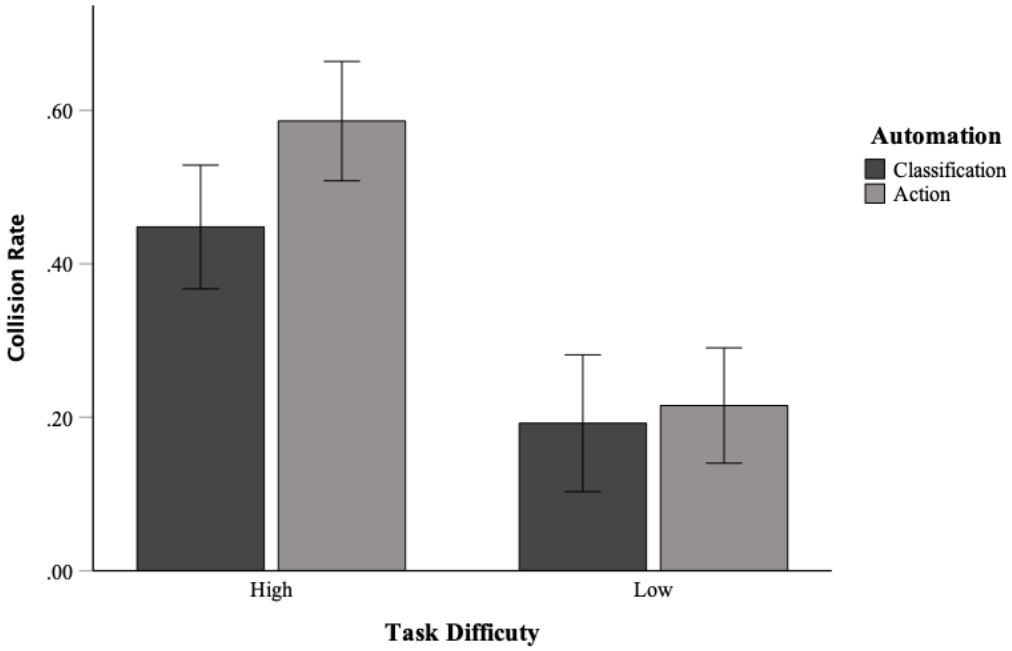
Similar to Experiment 1, two types of collision performance measurements were used. The first was the collision rate which measures the rate of collision per condition. The collision rate measure was used to compare overall workload between the two automation conditions. This measure was calculated by number of collisions incurred, minus number of allowed collisions, divided by number of total possible collisions. The second measurement was the effect of subtasks on collision events used as a performance-based measure for subtask workload. It was expected that when a subtask has a high demand, participants failed to monitor and act on collision events. As a result, subtasks that participants were engaged with during moments of collision were recorded and counted.

Collision Rate

A 2 x 2 repeated measures with two levels of task difficulty, and two types of automation was used to investigate Collision rate performance. The analysis showed that participants conceded less collision in the Classification Automation condition than in the Action Automation condition when the task difficulty was high (see Figure 28).

Figure 28

Collision Rate performance. Error bars represent 95% CI



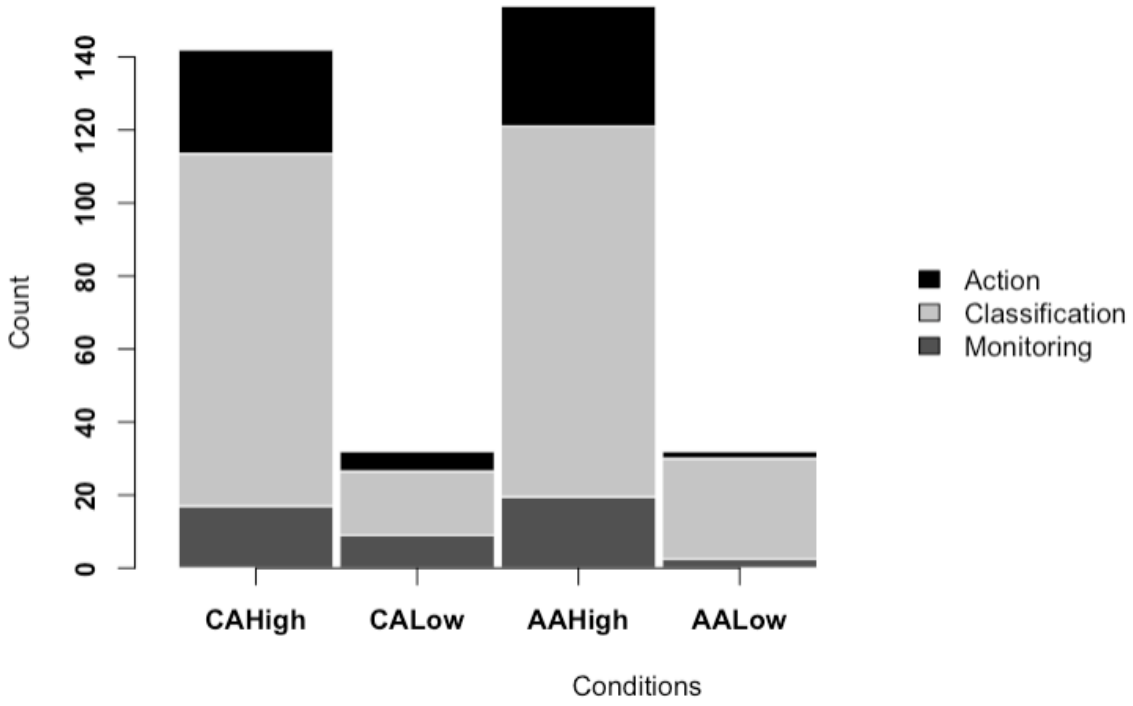
There interaction effect of automation and task difficulty was not significant, $F(1, 30) = 3.76$, $MSE = 0.027$, $p = .062$, Cohen's $f = 0.11$. There was a significant main effect of automation, $F(1, 30) = 4.28$, $MSE = 0.047$, $p = .047$, Cohen's $f = 0.13$. Participants conceded less collisions in the Classification Automation condition than in the Action Automation condition. There was also a significant main effect of task difficulty, $F(1, 30) = 77.01$, $MSE = 0.039$, $p < .001$, Cohen's $f = 1.04$.

A follow up paired sample *t*-test showed that participants performed significantly better in Classification Automation than in Action Automation when the task difficulty was high, $t(31) = 3.03$, $p = .005$, $d = 0.54$. The automation effect was not significant when the task difficulty was low, $t(31) = 0.44$, $p = .661$, $d = 0.08$. However, this is a speculative analysis since the interaction effect of automation and task difficulty was not significant.

Subtasks and Collisions

Consistent with Experiment 1, subtasks participants were performing during collision events were recorded and analyzed. There were 142 collision events recorded in the Classification Automation-High task difficulty condition (CAHigh), 32 in the Classification Automation-Low task difficulty condition (CALow), 154 in the Action Automation-High task difficulty condition (AAHigh), and 32 in the Action Automation-Low task difficulty condition (AALow).

Figure 29
Proportion of Subtasks that participants were performing during collision events in each condition



A chi-square test of independence was performed to assess the relationship between the subtasks and the automation conditions. The was no significant relationship between the two variables $X^2(6, 360) = 12.01$, $p = .062$. In other words, the proportion of subtasks participants were performing during collision events appeared to remain the same across all conditions (see Figure 29). The classifying subtask was consistently the more dominant subtask participants

performed during collision events. The action subtask was the second most demanding subtask followed by monitoring.

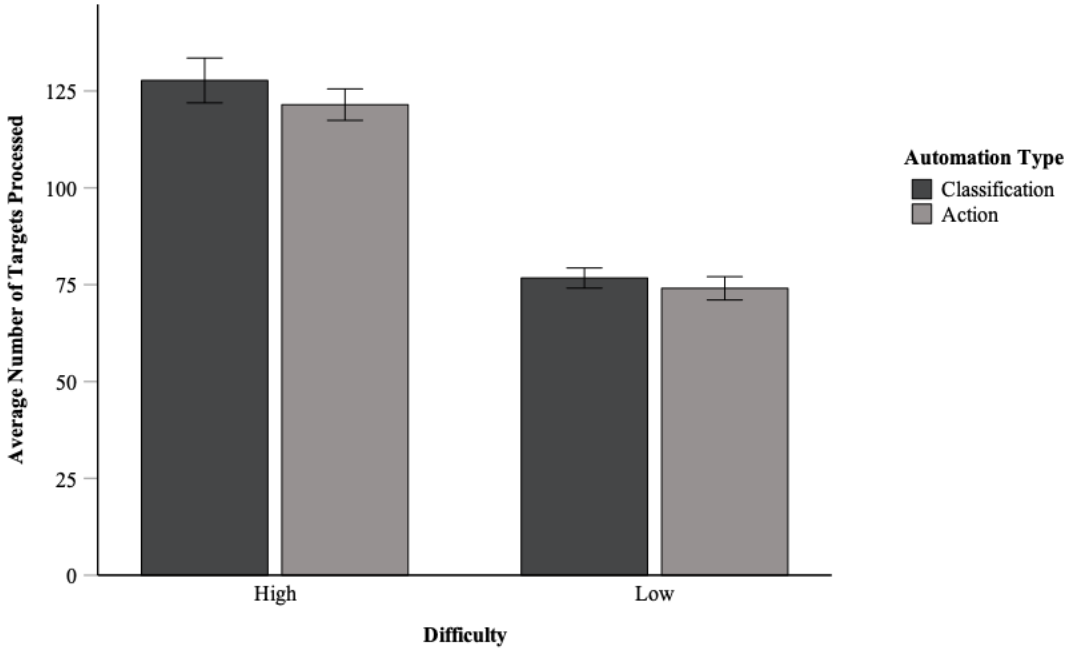
Number of Targets Processed and Shot

Number of Targets per Trial

As described in the design section, differences in the number of targets participants process could highlight some of the effects of automation. A 2 x 2 repeated measures ANOVA with two level of task difficulty and two types of automation was conducted to compare differences in number of targets processed per condition (see Figure 30).

Figure 30

Average number of targets per trial. Error bars represent 95% CI



There was no interaction effect of Automation and Task Difficulty, $F(1, 31) = 2.09$, $MSE = 49.35$, $p = .16$, Cohen's $f = 0.06$. There was a significant main effect of Automation type on average number of targets processed, $F(1, 31) = 14.1$, $MSE = 45.02$, $p < .001$, Cohen's $f = 0.33$. The main effect of task difficulty on number of targets processed was also significant, $F(1, 31) = 593$, $MSE = 130.64$, $p < .001$, Cohen's $f = 0.90$. The main effect of task difficulty was expected

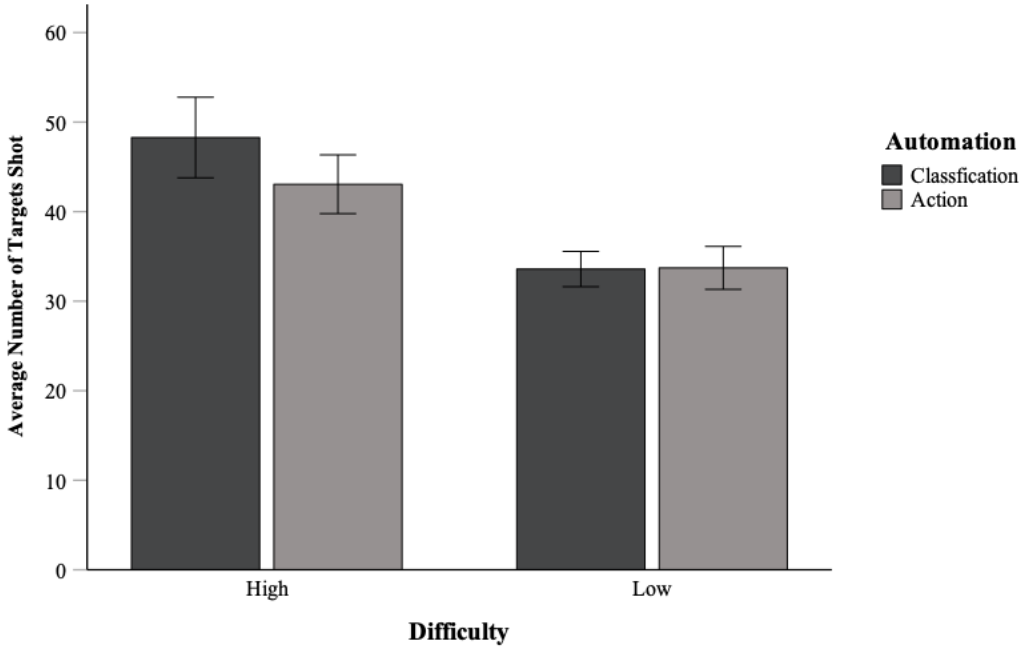
since there were more targets in the high task difficulty condition. As a result, the main effect of Automation on number of targets processed was the primary interest in this analysis.

Number of Targets Shot

A 2 x 2 repeated measures ANOVA with two level of task difficulty, and two types of automation was conducted to compare differences in number of targets shot per condition. Participants shot more targets in the Classification Automation than in the Action Automation condition when the task difficulty was high (see Figure 31).

Figure 31

Average number of targets shot pr trial. Error bars represent 95% CI



There was a significant interaction effect between task difficulty and automation, $F(1, 31) = 6.22$, $MSE = 36.72$, $p = .018$, Cohen's $f = 0.17$. Further analysis of the interaction showed that participants shot significantly more targets in Classification Automation than in Action Automation, $t(31) = 2.74$, $p = .01$, $d = 0.49$, when the task difficulty was high. When the task difficulty was low, difference between the two automations was not significant, $t(31) = 0.11$, $p =$

.914, $d = 0.02$. The difference between high and low task difficulty conditions was significant in both automation conditions as expected.

Discussion

The objective of this experiment was to compare the effect of automating high and low workload subtasks on human performance. Classifying and action (shooting) subtasks were designated and automated as high and low workload subtasks. Participants performed in both automation conditions under two levels of task difficulty. The primary inquiry was whether the automation contributed to an enhancement in overall task performance. Findings revealed that automating the high workload (classification) subtask resulted in a more substantial performance improvement compared to automating low workload (action) subtask. Participants demonstrated significantly better performance in the Classification Automation (CA) condition than in the Action Automation (AA) condition, particularly when the task difficulty was high.

The next inquiry was whether the automation led to an improvement in overall workload. Indeed, collision rate scores indicated that Classification Automation reduces workload compared to Action Automation. Participants conceded less collisions in the Classification Automation condition than in the Action Automation condition when the task difficulty was high. This outcome indicated that Classification Automation allowed participants to allocate more mental resources to monitor and respond to collision events. This makes sense because it was also observed that participants monitored more in Classification automation than in Action Automation. This pattern was also evident in participants assessment of task difficulty of the three subtasks.

To get a deeper understanding of the sources of the overall performance differences, subtask performance and subtask workload ratings were examined. Given the enhanced

performance in the Classification Automation condition, the first question addressed was whether the automation contributed to improved classification subtask performance. The results showed that there was no significant difference in the proportion classified and classification accuracy between the two automation conditions. The only difference observed was that participants classified a greater proportion of targets in low task difficulty conditions than in high task difficulty conditions. This indicated that the overall performance differences between automation conditions were not because participants classify more targets in one condition than the other.

Surprisingly, classification accuracy did not show a significant difference between automation conditions. The only difference observed in classification accuracy was because of task difficulty, with participants showing a superior performance in low task difficulty conditions compared to high difficulty conditions. However, it is important to acknowledge that Classification Automation's accuracy is dependent on sensor reliability which had a maximum reliability of 90%. The automation was designed to make the most accurate decision based on classifications from sensors (AWACS), similar to human operators. In essence, if the human operator has sufficient time to think and decide, they could potentially perform as well as the automation. However, given the task demand in this experiment, it was unlikely that humans matched the automation's performance. If that was the case, there would not be an overall performance difference in CA than in AA. Classification automation did improve performance, but the effect was not reflected in classification accuracy and proportion classified.

Having established that the overall performance difference did not stem from the classification performance alone, a subsequent analysis was conducted to understand whether the performance differences originated from the shooting component of the task. The questions

posed here were whether participants processed and shot more targets in the Classification Automation compared to the Action Automation condition. Results indicated that participants processed more targets in Classification Automation condition than in the Action automation condition, particularly under high task difficulty. Furthermore, participants shot more targets in Classification Automation condition than in the Action Automation when the task difficulty was high. This suggested that Classification Automation enabled participants to process and act on more targets, resulting in an improved net score performance. The impact of automation was not directly observed on the automated subtask itself, but rather in its ability to free up participants' mental capacity to engage with other components of a task. It seemed that the addition of classification automation improved the action part of the task than the classifying subtask itself. This is consistent with the theory that resource allocation within a task could be dependent on the demands of the components (Norman & Bobrow, 1975).

Similar to the analysis of subtask performance, an examination of subtask workload and its contribution to the overall task workload was conducted. Subtask workload was measured by ISA and post-trial subjective ratings. The main question addressed here was whether the automation led to a decrease in subtask workload. Both the ISA and post-trial subjective responses indicated that Classification Automation did, indeed, alleviate the classification subtask workload. When participants were asked to rate the difficulty of each subtask, their ratings varied depending on automation conditions. The proportion of difficulty ratings for the classification subtask was higher in the Action Automation condition than in Classification Automation. This showed that the perceived workload of the classification subtask was reduced due to Classification Automation. In addition, the workload rating of the monitoring and the action subtasks increased in Classification Automation compared to the Action Automation. The

reason could be that participants engaged more with the monitoring and action subtasks when the classifying subtask was automated. The increased engagement with the monitoring and the action subtask might have resulted in higher workload rating. Despite classification remaining the predominantly demanding subtask in all conditions, its impact was diminished with the application of Classification Automation. This demonstrated that Classification Automation did reduce the perceived workload of the classifying subtask.

This shift in participant's perception of subtask due to automation was also reflected in the ISA analysis. In Experiment 1, although differences between subtasks were not significant, classifying was consistently rated slightly higher in all conditions. However, in this experiment, interaction between automation condition and subtask, and between task difficulty and subtask were observed. The classifying subtask was rated as more demanding in Action Automation than in Classification Automation. In addition, within the Action Automation condition, the classifying subtask was more demanding than the monitoring and action subtasks. These results underscored that automating the classifying subtask did reduce its workload.

The secondary objective of this experiment was to investigate how automation influenced the human interaction with the system. Notably, participants adopted various strategies that might have implications to their task performance. For instance, some participants prioritized classifying enemy targets while neglecting to classify friendly ones. With Classification Automation enabled, participants clicked on "view information," ascertain the target's enemy or friendly status, and classified it only if it was an enemy target. This strategic choice was made because classifying friendly targets did not prompt any subsequent action, as they were permitted to land without further operator intervention. While this strategy could potentially improve overall scores, the accuracy of classifying targets might have suffered because of it. Although

this strategy was predominantly used by one participant, whose data was discarded, similar approaches were observed with other participants in moments when they were rushing to act on targets.

Another strategy observed involved attempts to shoot targets without prior classification. This attempt was prevented because the systems did not allow shooting unclassified targets. Additionally, participants adopted a strategy of classifying targets based on broader categories than exact matches. In other words, participants classified enemy as enemy and friendly as friendly in general. This categorization approach resulted in higher a higher accuracy score in classification by category than by exact match.

Participants' prioritization of the shooting task may have been influenced by motivational factors, as they received feedback on their target elimination performance. This feedback could have prompted participants to allocate more resources to the shooting task. This not only raised concerns about task prioritization but also had the potential to influence subjective ratings of subtasks. Prior research has showed that in demanding tasks, motivation can create a dissociation between performance and subjective measurements, where performance improves while subjective ratings remain high (Yeh & Wickens, 1988). Thus, it was crucial to control for strategies and optimize the automation to mitigate potential issues.

The experiment underscored the nuanced ways in which automation can impact users, highlighting the necessity of testing and optimizing automation to align with desired goals. In Experiment 3, modifications were made to encourage participants to focus equally on the classification subtask and the action part of the task, aiming for a more balanced interaction with the automated system.

Chapter 4

Experiment 3: Optimize and Test the Automation

As previously outlined, the goal of this experiment was to prevent strategies wherein participants might show a tendency to favor some tasks, aiming to maximize their gain in overall performance scores. When participants prioritized a subtask than the other, it is difficult to understand the genuine impact of Classification Automation on subtask performance improvement. Previous research also showed that strategies developed during a task might affect workload rating responses (Hancock & Matthew, 2018). Clever strategies could affect both performance of the subtasks and the corresponding ratings. Hence, an effort was made to balance the importance of the tasks, ensuring participants refrained from favoring one subtask over the other, or employing shortcuts to reach their primary goals.

In the preceding two experiments, the absence of a scoring mechanism for classification performance may have inadvertently prompted participants to emphasize shooting down enemy aircraft, possibly overshadowing their classification performance. To address this, a classification scoring scheme along with corresponding feedback was introduces in Experiment 3. Considering a ceiling effect was observed in low task difficulty conditions, only high task difficulty condition was evaluated in this experiment. The effect of automation was more pronounced when the task difficulty increased.

Method

Participants

The number of subjects needs for the study was determined by an *a priori* power analysis using G*Power for a repeated measures ANOVA with a significance level of .05. A total of 28 subjects was sufficient to detect a medium sized effect of 0.25 with 80% power (Table 6). There

were 33 participants recruited for the study. Participants were recruited from Rice University undergraduate subject pool, and they were compensated with credit towards a course requirement. There were 17 female and 16 male participants with age ranging from 18 to 23 (M =19.85).

Table 6
Power Analysis

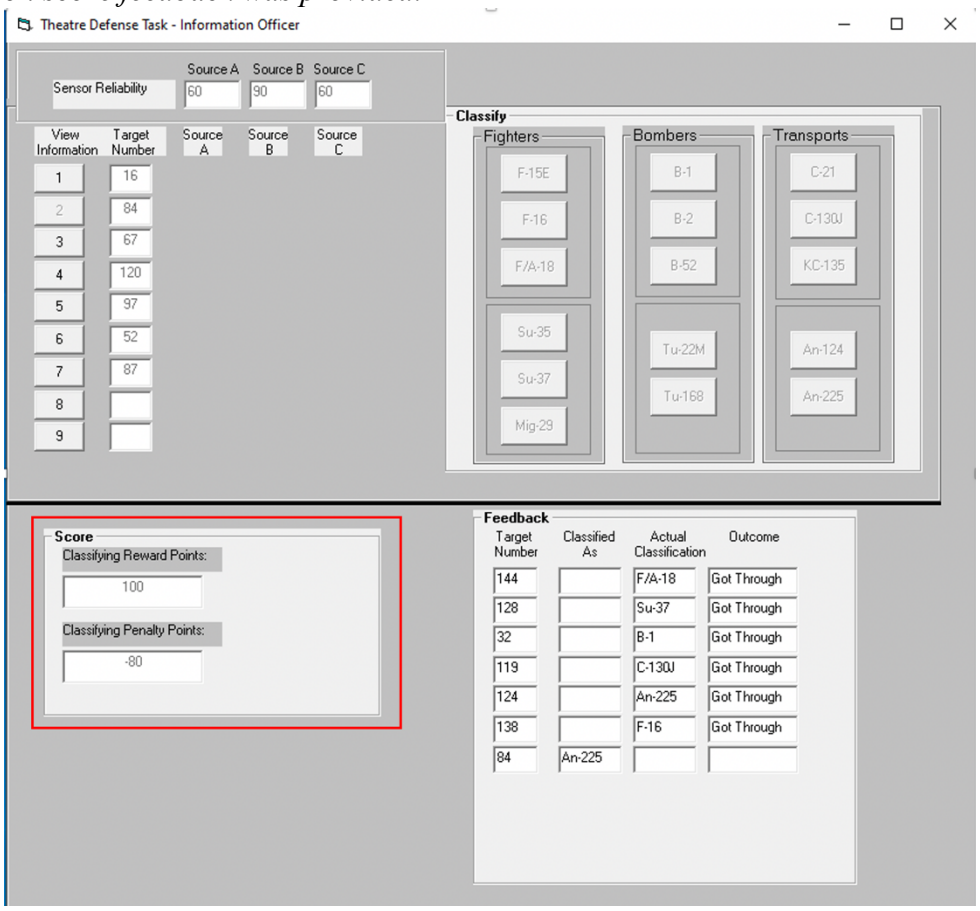
	Power			
Cohen's *f*	.8	.85	.9	.95
.10	163	184	213	259
.25	28	31	36	43
.40	12	13	15	18

Task

The task used in this experiment was the same as the one used in Experiment 2. The only difference was a classification scoring scheme and feedback were added (see figure 32). The Classifying Reward Points showed the sum of reward points earned from classification performance, and the Classifying Penalty Points part showed the sum of negative points accumulated from failing to classify targets.

Figure 32

Modified IO Screen. The score section highlighted in a red box shows the area where classification score feedback was provided.



Design

A 2 x 3 within-subjects design was used with two types of automation and three subtasks. The task difficulty was high in all conditions, and order of Automation presentation was randomized. All dependent variable measures from Experiment 2 were used in this analysis as well. In addition, a new classification performance score was used to compare classification performance.

The additional classification measure was points earned based on classification accuracy. When participants accurately classified a target, they were rewarded with 20 points (e.g., An-124, had to be classified as An-124). When targets were classified by type, participants earned 15

points (e.g., An-124 classified as An-225, both are enemy transports). If participants classified a target only as enemy or friendly, they received 10 points (e.g., C-130J friendly transport classified as B-1 friendly bomber, or An-124 ET classified as Tu-22M EB). When participants misclassified enemy as friendly or vice versa, they were penalized 15 points. In addition, if participants neglected to classify a target, they were penalized 10 points. Classification performance score was used to calculate classification score consistent with the overall task performance net score used in previous experiments.

Classification Score

Classification score was calculated from reward score, penalty score, and the total number of targets processed. Reward score was calculated by summing the positive points gained for classifying targets and subtracting the negative scores incurred from misclassifying targets. The penalty score was the sum of negative points incurred from failing to classify targets.

$$Classification\ Score = \frac{Reward\ score - Penalty\ score}{Total\ number\ of\ targets\ per\ trial} \quad (2)$$

Classification score was used as a dependent variable to measure performance differences between the Classification and Action Automation conditions.

Results

Data from 33 participants was used in these analyses. Various analyses were conducted comparing performances between Classification and Action Automation conditions. Before going into the specific subtask performances and the effect of automation optimization, it was important to see if the overall performance scores were consistent with Experiment 2.

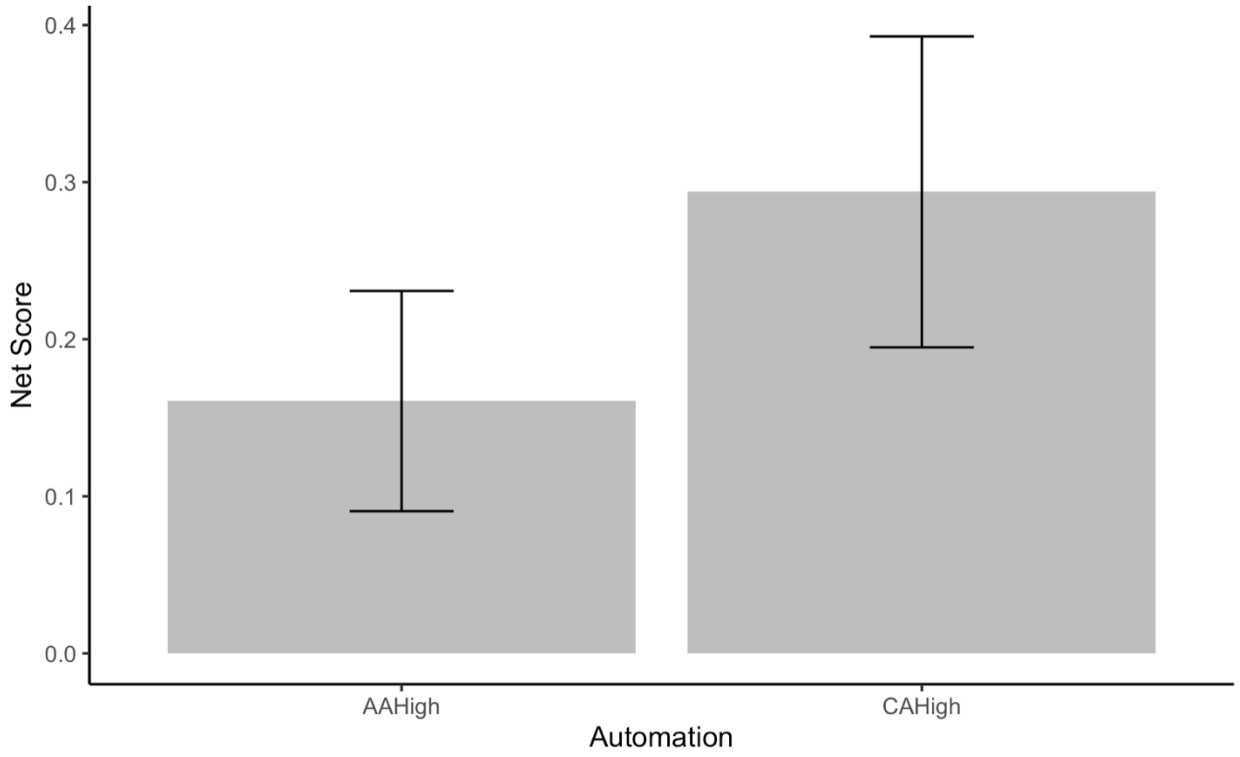
Overall Performance

Net Score

Net Score was calculated using equation 1, and it measured the overall task performance. Consistent with Experiment 2, participants performed better in the Classification Automation condition than in the Action Automation condition (see Figure 33).

Figure 33

Net score performance. Error bars represent 95% CI



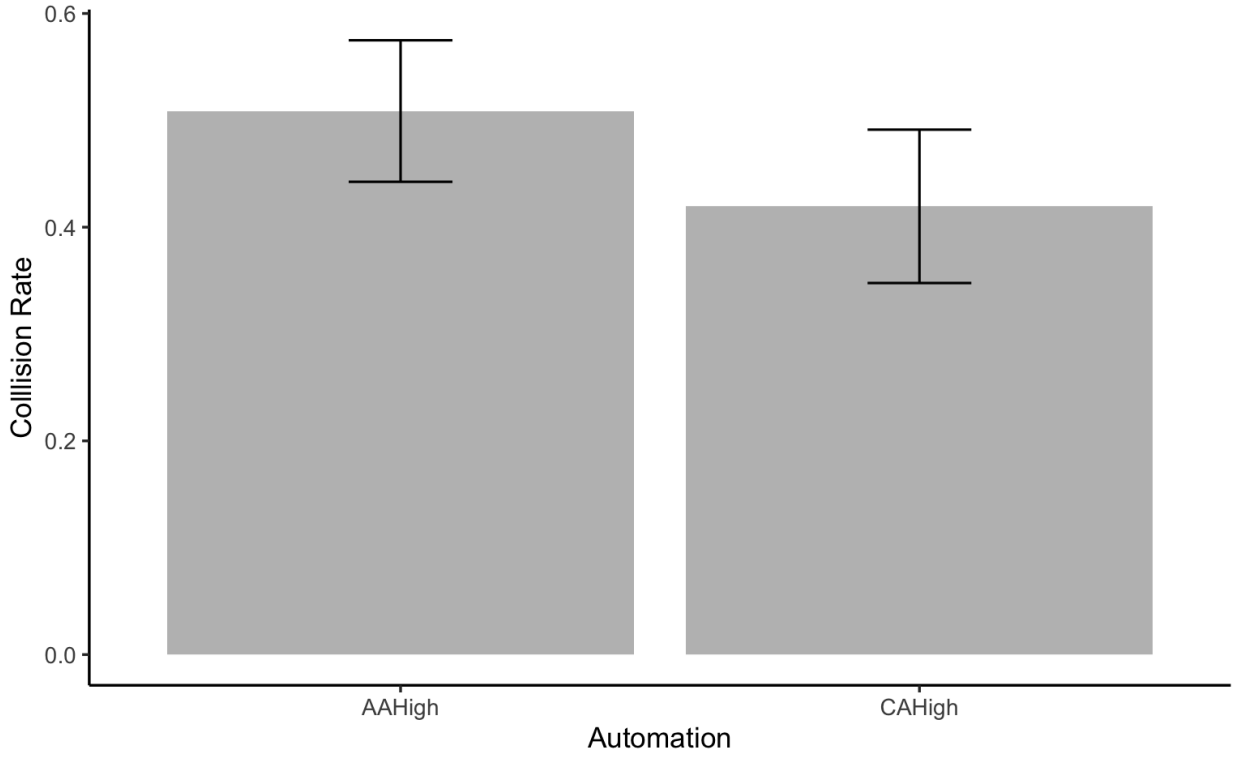
A paired sample *t*-test was showed that participants had a significantly higher net score in the Classification Automation condition than in the Action Automation condition, $t(32) = 2.58$, $p = .015$. $d = 0.45$. This finding affirmed that Classification Automation, or automation of high workload subtask, leads to enhanced overall performance. Thus, this result supports the hypothesis that automating high workload subtask helps improve task performance.

Collision Rate

Collision rate was calculated as the ratio of number of collisions conceded by the number of total possible collisions per trial. It was observed that participants conceded less collisions in Classification Automation than in the Action Automation condition (see Figure 34). Consistent with Experiment 2, the difference in performance showed a higher overall workload in the Action Automation than in the Classification Automation conditions.

Figure 34

Collision Rate Net Score by Automation Condition. Error bars represent 95% CI.



A paired sample *t*-test showed participants conceded significantly less collisions in Classification Automation than in Action Automation, $t(32) = 2.42$, $p = .021$. $d = 0.42$. This result showed that in the Classification Automation condition, participants had adequate mental resources to monitor, identify, and act on collision events. The next crucial step was to delve into the source of this performance improvement and assess the effectiveness of the automation

optimization. The first component to look at was if the Classification Automation improved the classification subtask.

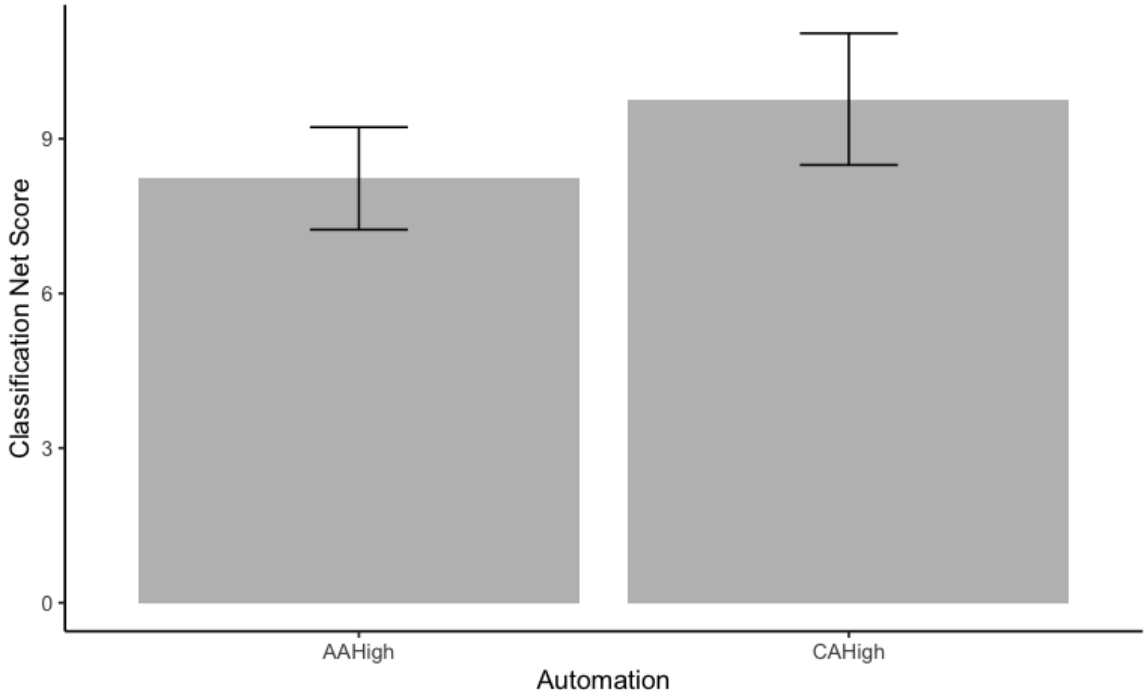
Classification Performance

Classification Score

Classification score was computed using equation 2. Classification score analysis showed that participants performed better in the Classification Automation condition than they did in the Action Automation condition (see Figure 35).

Figure 35

Classification net score. Error bars represent 95% CI



A paired sample *t*-test showed that participants had a significantly higher Classification Score in Classification Automation than in Action Automation, $t(32) = 2.36$, $p = .024$. $d = 0.41$. This score depended on how participants classify targets. Consequently, the higher classification score in Classification Automation indicated that participants were either more accurate in their classifications, or able to successfully classify a larger number of targets, or perhaps both.

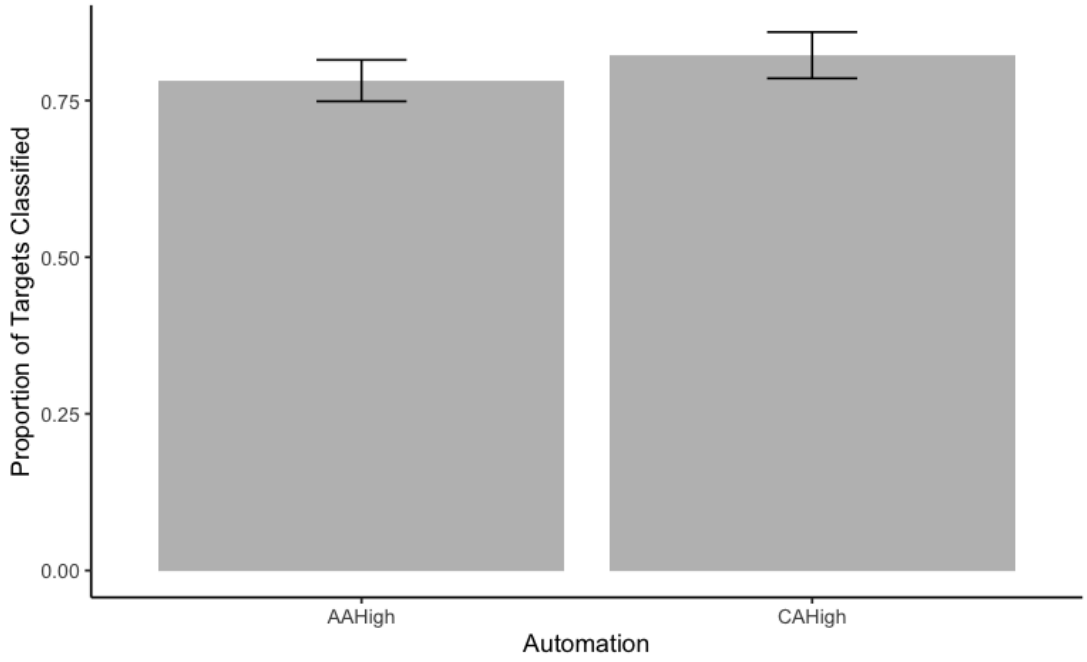
Importantly, this result underscores that optimization of the automation effectively encouraged operators to prioritize their classification subtask to the same extent as the shooting subtask.

Proportion of Targets Classified

As mentioned in the previous section, the Classification Score may be attributed to participants classifying a greater number of targets. The proportions of targets classified was used as a metric to assess whether participants classified a larger number of targets in one automation condition compared to the other. Notably, participants classified a higher proportion of targets in the Classification Automation condition than they did in the Action Automation condition (see Figure 36).

Figure 36

Proportion of targets classified. Error bars represent 95% CI



A paired sample *t*-test showed that participants classified significantly more targets in the Classification Automation condition than they did in the Action Automation condition, $t(32) = 2.49$, $p = .018$. $d = 0.43$. The outcome of this measure was not significantly different between automation conditions in Experiment 2. This further supports the notion that the intervention

introduced in experiment 3 encouraged participants to allocate more attention to their classification performance.

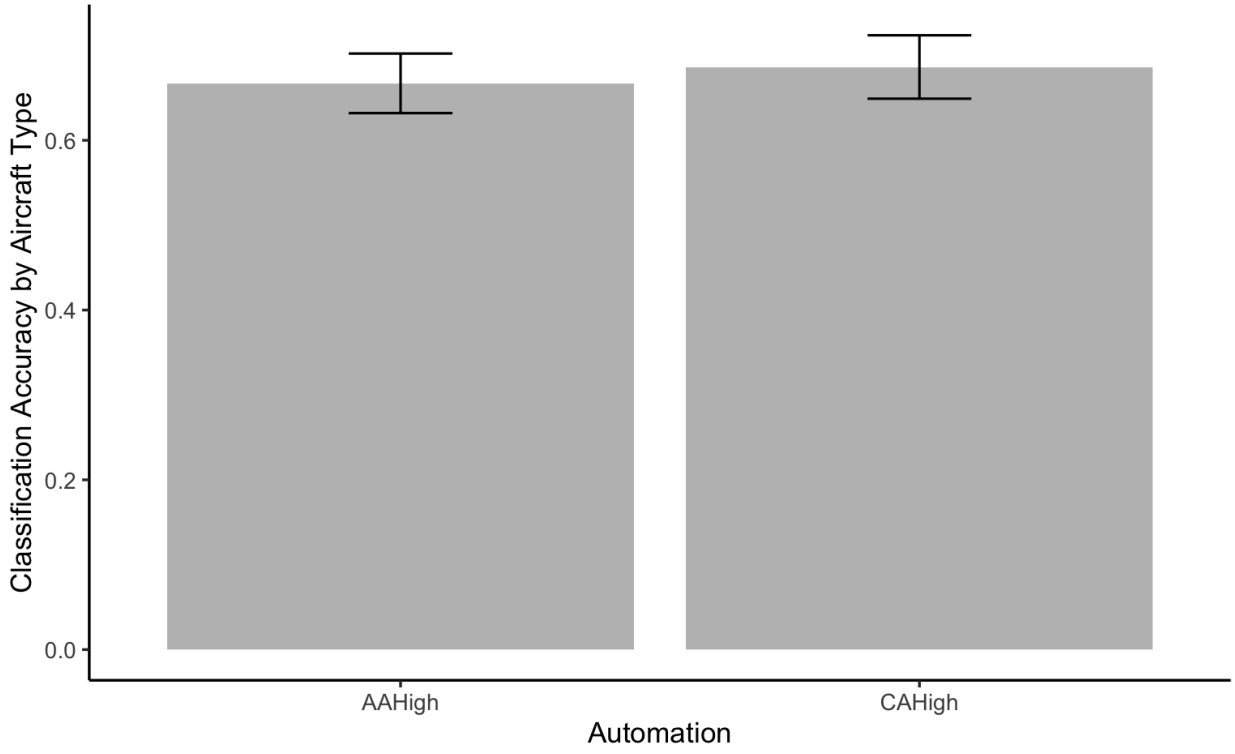
Another potential contributor to the difference in Classification Score could be variations in classification accuracy performance. As a result, two classification accuracy analyses were performed, similar to the approach used in experiment 2.

Classifying by Aircraft

The analysis showed that there wasn't a significant difference in classification accuracy between the two automation conditions (see Figure 37).

Figure 37

Classification Accuracy by Aircraft Type. Error bars represent 95% CI.



A paired sample *t*-test showed that accuracy did not differ significantly between the Classification Automation and the Action Automation conditions, $t(32) = 0.86$, $p = .40$. $d = 0.15$.

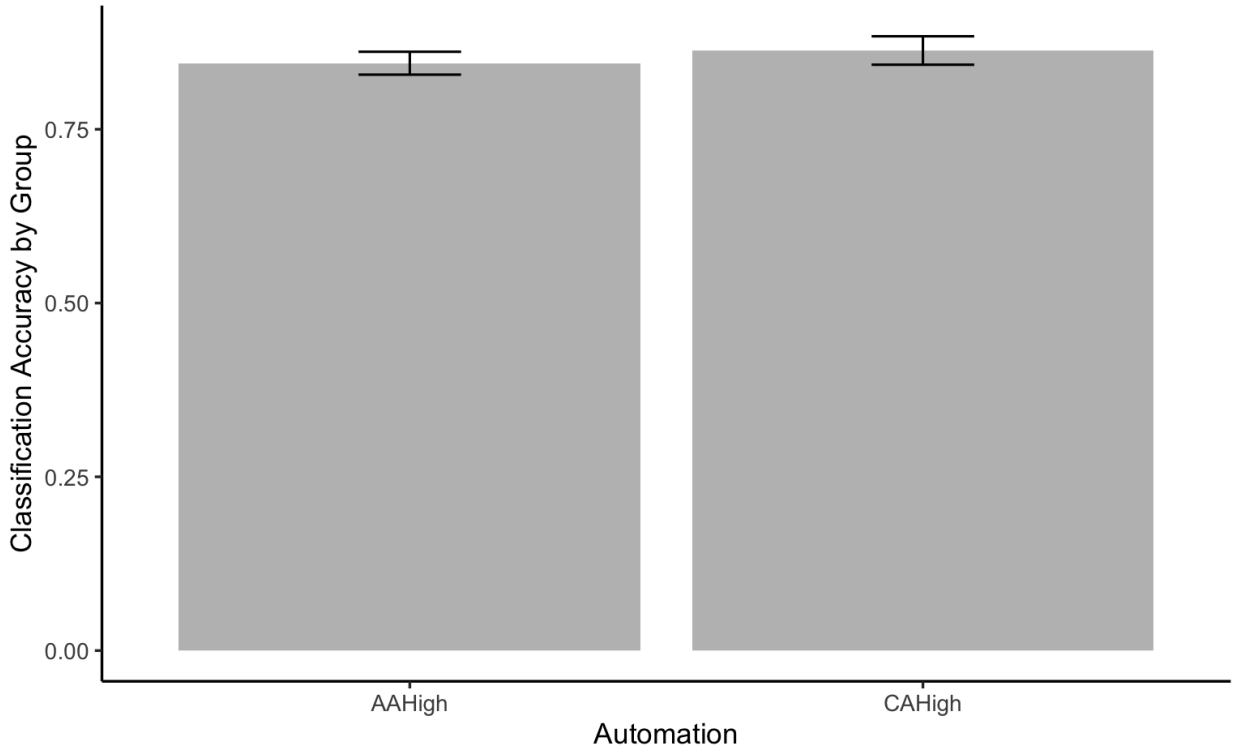
Similar to Experiment 2, the automation effect was not observed in the classification accuracy by aircraft type measure.

Classifying by Category

The classifying by category performance difference between the two automation conditions was not significant (see Figure 38).

Figure 38

Classification accuracy by group (within enemy or friendly). Error bars represent 95% CI



A paired sample *t*-test showed that accuracy did not differ significantly between Classification Automation and Action Automation, $t(32) = 1.49$, $p = 0.15$. $d = 0.26$. Much like Experiment 2, the effect of automation was not evident in the classification accuracy by category measure. The classification accuracy results remained consistent between the second and third experiments, indicating that the intervention did not lead to observable changes in these accuracy measures.

The classification performance results showed that the automation optimization contributed to performance improvement in certain aspects of the classifying performance. This improvement was evident in the outcomes of the classification score and proportions classified. However, accuracy did not show a significant improvement. How about the shooting performance? Consistent with Experiment 2, was there a performance difference between automation conditions?

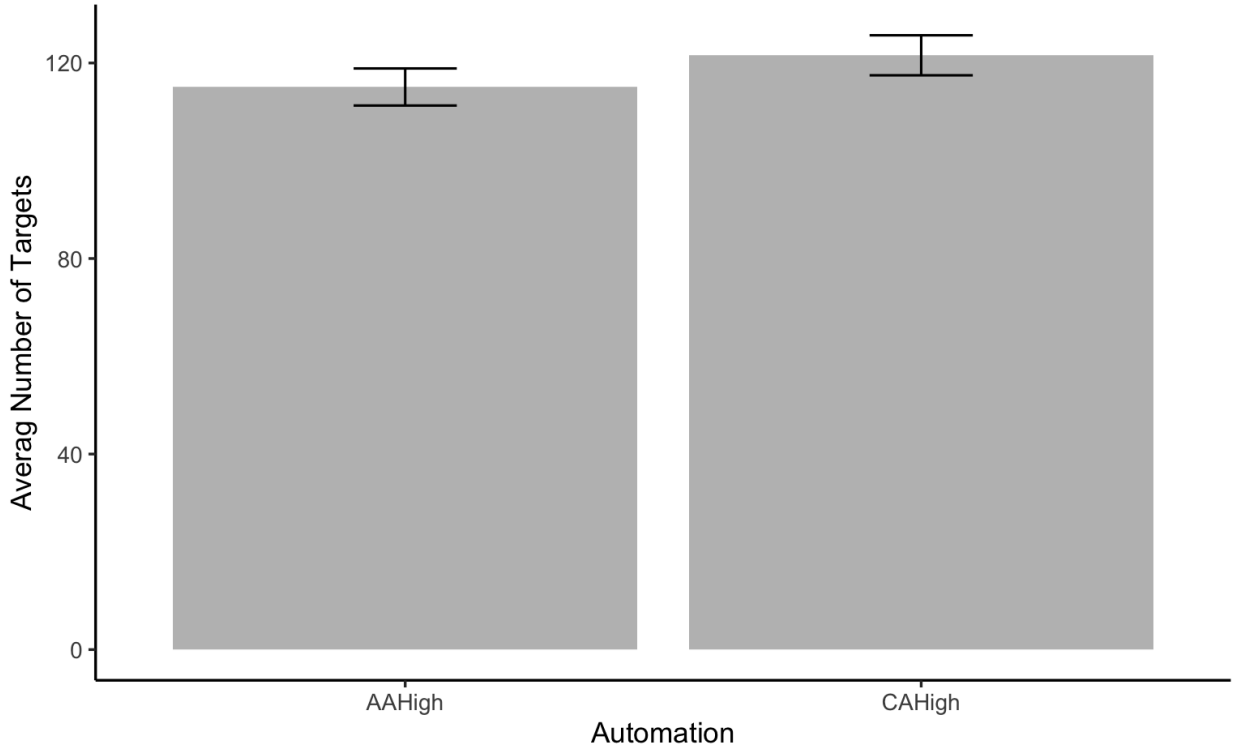
Number of Targets Processed and Shot

Number of Targets per Trial

The total number of targets per trial was the same measure used in Experiment 2. Performance differences between number of targets processed was an indicator that participants acted on targets quicker. Analysis of number of targets per trial showed that participants processed more targets in the Classification Automation condition than in the Action Automation condition (see Figure 39).

Figure 39

Average number of targets per condition. Error bars represent 95% CI



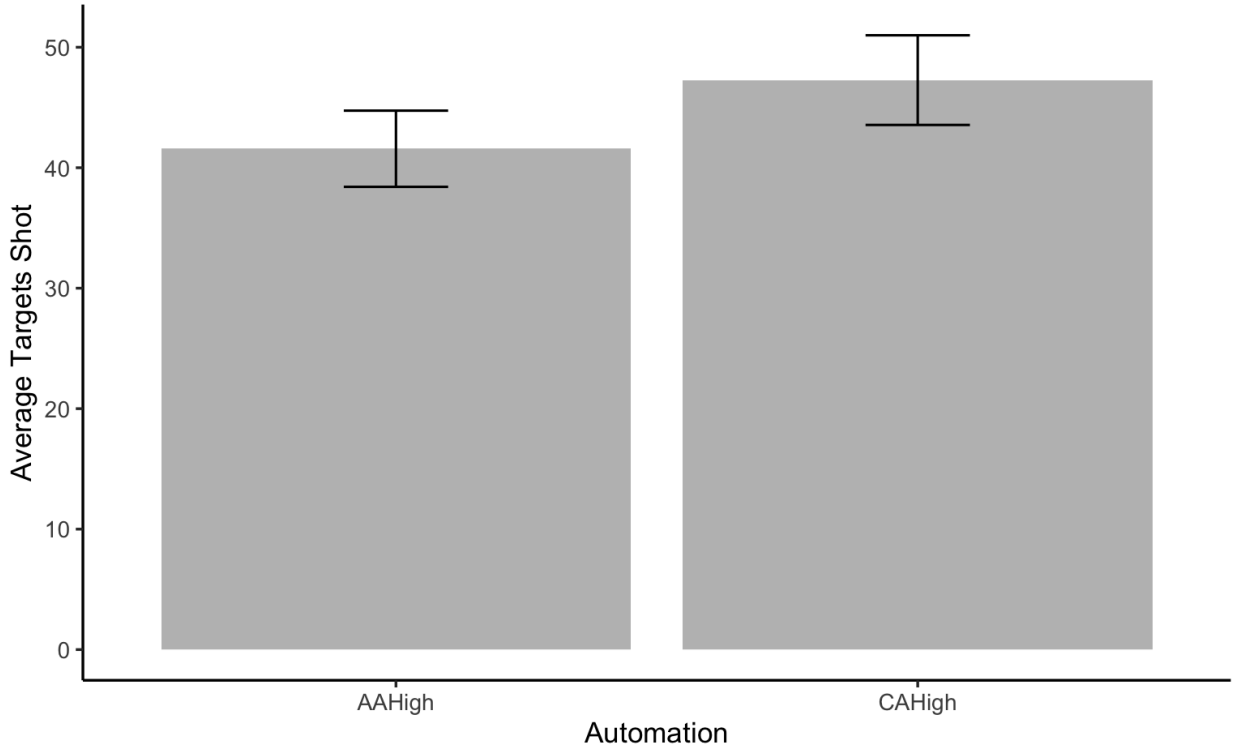
A paired sample *t*-test showed participants processed more targets per trial in Classification Automation than in Action Automation, $t(32) = 4.65$, $p < .001$. $d = 0.81$. This result showed that participants had more capacity to act on and process targets quickly when the classifying subtask was automated.

Number of Targets Shot

In this analysis the total number of targets shot per trial were compared between Classification Automation and Action Automation Conditions. Results showed that participants were able to act on or shoot more targets in the Classification Automation condition than in the Action Automation condition (see Figure 40).

Figure 40

Average Number of Targets Shot per Condition. Error bars represent 95% CI.



A paired sample *t*-test showed participants processed more targets per trial in Classification Automation condition than in the Action Automation condition, $t(32) = 4.66$, $p < .001$. $d = 0.81$. Participants had more time or capacity to act on more targets in Classification Automation than in Action Automation.

The analysis of shooting performance suggested that the automation employed in this experiment not only improved the shooting subtask but also enhanced aspects of the classifying subtask. The optimization of the automation appears to have had a positive impact on both overall task performance and individual subtask performances.

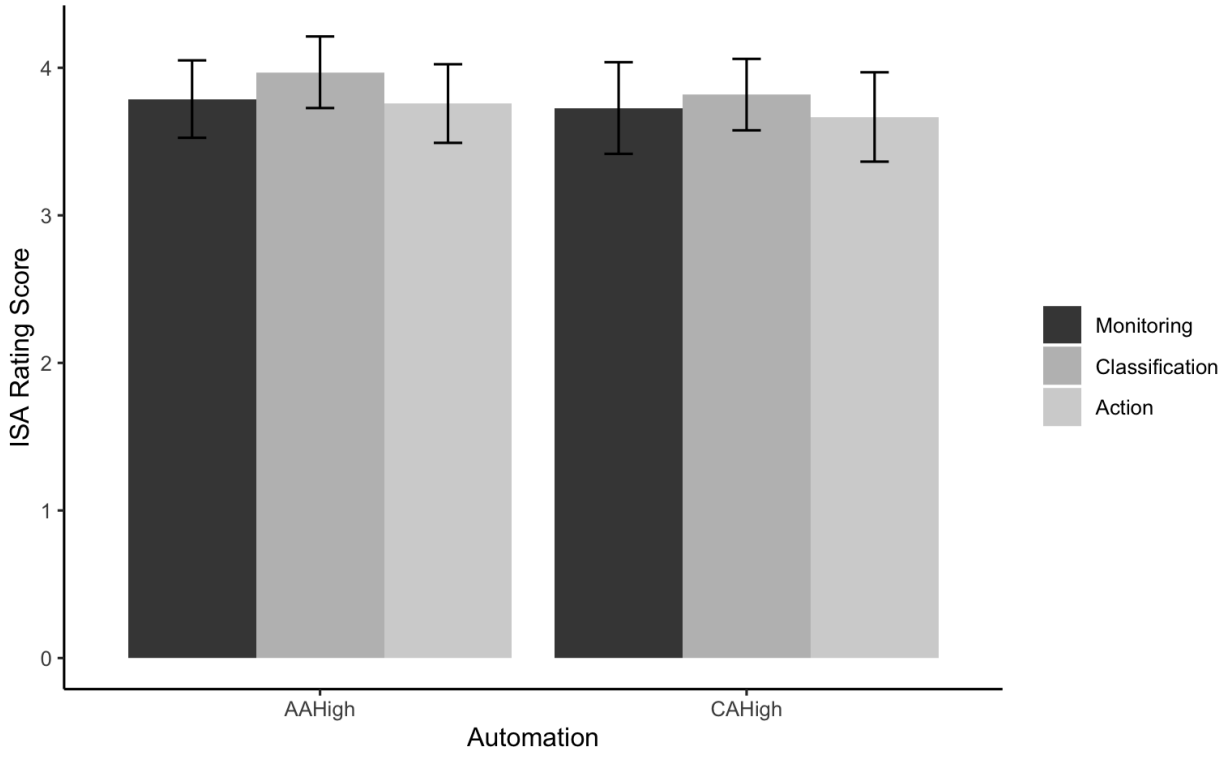
Subjective Workload Measures

Instantaneous Subjective Assessment (ISA)

A 2 X 3 within-subjects ANOVA with two types of Automation and three subtasks was conducted to see if ISA ratings differed between subtasks. The workload rating of subtasks did not show a significant difference (see Figure 41).

Figure 41

Average ISA rating scores per automation condition. AAHigh is Action Automation-high task difficulty condition and CAHigh is Classification Automation. Error bars represent 95% CI



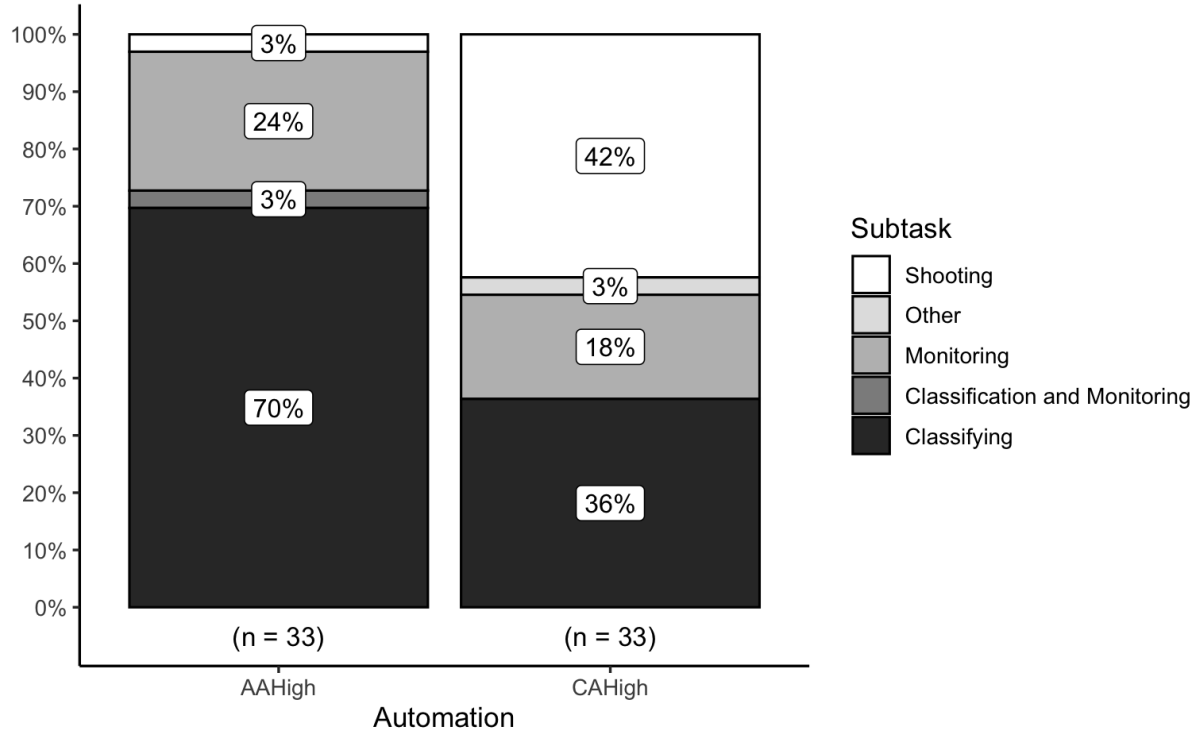
There was no interaction effect of automation type by subtask, $F(2, 64) = 0.20$, $MSE = 0.18$, $p = .83$, Cohen's $f = 0.01$. There was no main effect of Automation, $F(1, 32) = 1.17$, $MSE = 0.43$, $p = .29$, Cohen's $f = 0.04$. And there was no main effect of subtask, $F(2, 64) = 1.73$, $MSE = 0.34$, $p = .19$, Cohen's $f = 0.05$.

Subjective Response (Post-trial)

Subjective responses were collected after each trial and participants were asked which subtask, they felt was more demanding. One participant chose two tasks as equally difficult, and another participant said something outside the three subtasks (see Figure 42). A total of 66 responses were collected and analyzed.

Figure 42

Proportions of subtasks rated as most demanding per condition.



Fisher's exact test was performed to assess the relationship between automation conditions and Subtasks ($p < .001$). Classifying subtask was rated as the most demanding in both automation conditions, but its proportion decreased in the Classification Automation condition (see Figure 42). The action subtask was rated as more difficult when the automation was Classification Automation. The change in the rating of the monitoring subtask was minor.

Collision Performance

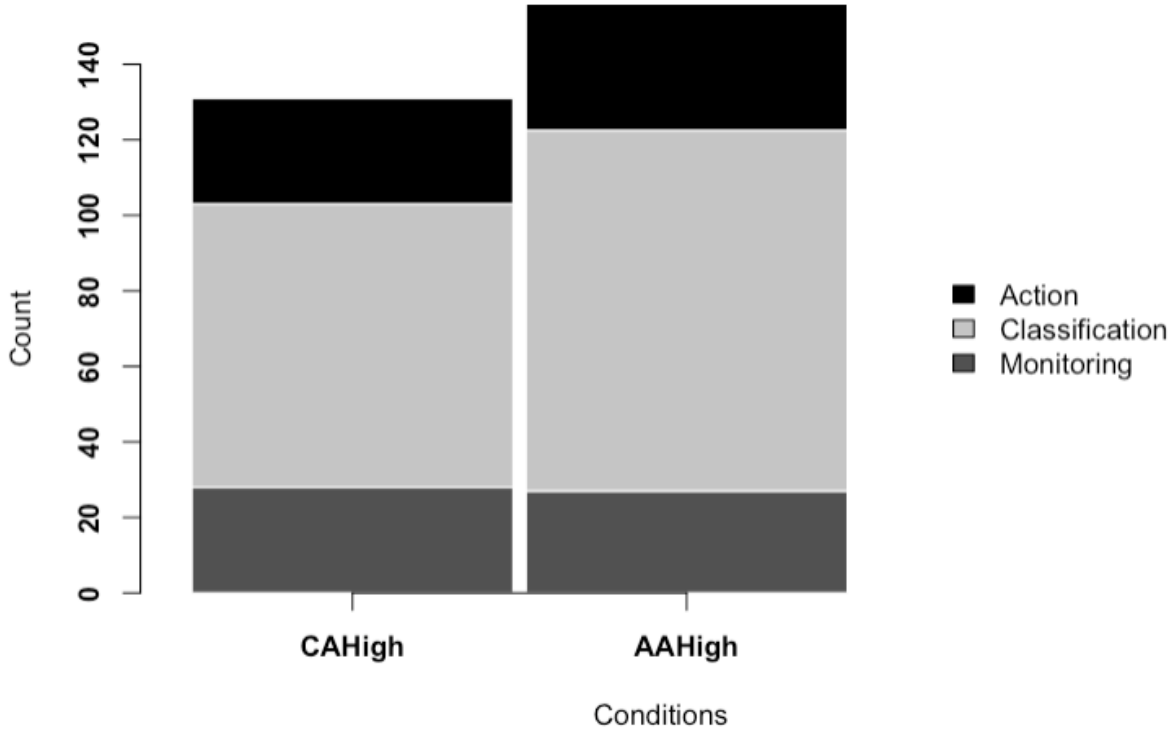
Collision performances were measured to evaluate performance differences between automation conditions, and to understand the effect of subtask difficulty on collision. As shown in the earlier collision rate was an indicator of the difference in overall workload between automation conditions. Subtasks participants were performing during moments of collision were used as a measure of subtask workload.

Subtasks and Collisions

Subtasks during collision events were recorded in the same way as in the previous two experiments. There were 156 and 131 collision events in the Action and Classification automation conditions respectively. Results showed that the proportion of tasks that contribute to collision events did not show a significant difference across automation conditions (see Figure 43).

Figure 43

Proportion of subtasks contributing to collision events per automation condition



A chi-square test of independence was performed to determine whether there was a relationship between the collision events and the subtasks. The proportion of subtask that contribute to collision were not dependent on automation, $X^2(2, 287) = 0.80$, $p = .67$. The result shows that classifying was the subtask participants were more frequently engaged with during collision events regardless of the automation condition. The Classification Automation reduced the rate of collision events but when collisions happen, participants were usually engaged with the classifying subtask than the other subtasks.

Discussion

The goal of this experiment was to optimize the automation and prevent participants from developing strategies where they prioritized the shooting part of the task. In Experiment 2, some participants appeared to focus more on maximizing their overall score that taking time to accurately classify more targets. To encourage participants to equally focus on the classifying task, a scoring scheme and feedback was added in this experiment. Participants performed the task in Classification and Action automation under high task difficulty condition.

Consistent with Experiment 2, participants performed better in Classification Automation compared to Action Automation, as evidenced by the overall performance net score, collision rate, the number of targets processed, and the number of targets shot. Notably, participants improvement in the classifying subtask, as indicated by the proportion of targets classified, and the classification net score. Unlike Experiment 2, where there was no significant difference in the proportion of targets classified, participants in this experiment classified more targets in the Classification Automation than in the Action Automation conditions. This shift in behavior aligns with the observation that none of the participants adopted a strategy of leaving some targets unclassified. Overall, the automation not only improved overall performance but also

enhanced components of the task, showing a positive impact on participant engagement and strategies.

The subjective rating of subtask showed a consistent result with Experiment 2 in that the classifying subtask was rated as most demanding in Action Automation. In addition, the action subtask was rated as the most demanding in Classification Automation. These indicated that participants felt a lower perceived workload when supported with automation in both conditions. However, most participants said they found Classification automation as most helpful, and they have not used the Action Automation feature. Hence the difference in the perceived workload might have come from the presence or absence of the Classification Automation.

ISA ratings did not show a significant difference between automation conditions and between subtasks. In Experiment 2, The classifying subtask was perceived as more difficult in Action Automation than in Classification Automation when measured across automations. In addition, within the Action Automation condition, the classifying subtask was perceived as more difficult than both the monitoring and the action subtasks. The absence of an effect in this experiment could be an indication that with the addition of a scoring scheme for the classification part, participants did not perceive the classifying subtask as less demanding even with the application of the Classification Automation.

In summary, the three experiments collectively affirmed the feasibility of applying a workload-based task selection for automation. Addressing the research questions, the findings indicated that it is indeed possible to automate subtasks based on the measurement of their respective workload. Moreover, the experiments demonstrated that a workload-based automation can enhance performance, particularly when applied to the more challenging components of a task. Automating high workload subtasks led to improved task performance, reduced overall

workload, improved subtask performance, and reduced subtask workload with appropriate optimization. The study also highlighted that automation might influence performance in unforeseen ways, underscoring the importance of optimizing automation to align with specific objectives

Chapter 5

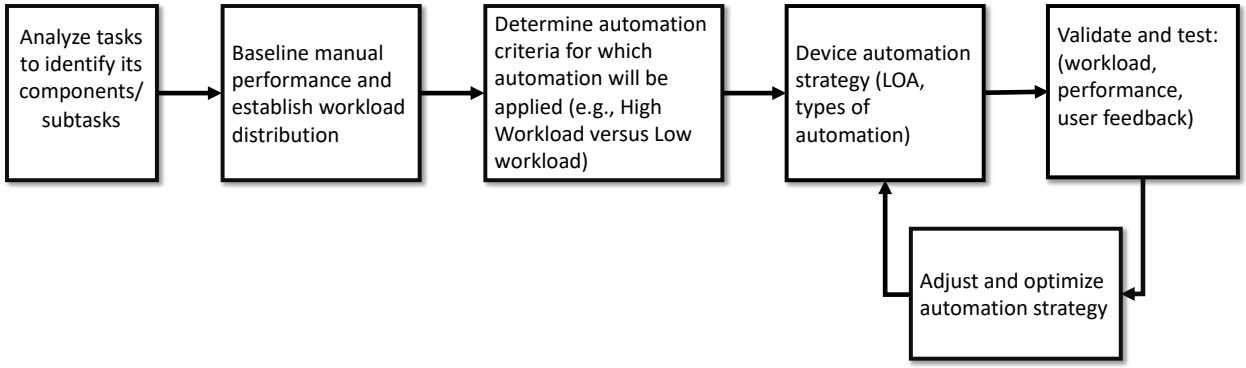
Theoretical implication

Automations are implemented with the aim to improve efficiency, reduce cost, and reduce workload. From the human factors perspective, effective automation should not only enhance performance but also maintain an optimal balance in workload, situation awareness, and trust at an optimum level. However, prior research showed that the introduction of automation often fails to decrease workload and can even result in increased workload. This is often due to the added responsibility of monitoring and supervising automated systems. Task selection is an important factor in design of automation. However, in the context of workload in automation, many studies focus on assessing workload to enable adaptive automation and dynamic function allocation. In these instances, workload is evaluated to decide when to introduce automation, and often neglect to explain which specific subtasks should be automated. This experiment focused on developing a framework for task selection by measuring the workload distribution of a task, with the aim of guiding automation decisions.

The first step of the framework used in this experiment was to analyze the Theatre Defense Task and identify its components (see Figure 44). After identifying the three main subtasks, the next step involved measuring the overall task workload and the workload of individual subtasks while participants performed the task with no automation. Once the workload distribution was known, it was shown that automating the high workload subtask improves performance and reduces workload. Once a task selection is made, automation strategies could be devised, implemented, and iteratively tested on humans until a desired level of performance is reached.

Figure 44

Framework for Workload-Based Task Selection



The ISA ratings, post-trial subjective ratings, and in-task performance measures were used to evaluate workload of subtasks. The ISA ratings did not reveal a significant difference between subtasks. However, outcomes of post-trial subjective responses and the secondary task performance showed consistent outcomes, allowing the identification of high and low workload subtasks. While there was a strong association between post trial subjective responses and performance measures, their outcome differed from the ISA ratings. This dissociation between measures is in line with prior research that showed different workload measures do not aways give a similar output. The sensitivity of post-trial subjective responses could be explained by prior findings that subjective ratings are more reliable to measure relative workload of tasks.

It could be assumed that the dissociation maybe because ISA was not sensitive enough to discriminate workload of subtasks. However, ISA ratings were sensitive to task demand changes across high and low workload conditions. In this case the result from ISA was aligned with outcomes of performance measures. In addition, ISA measures did show differences for subtasks when comparisons were made between high and low workload automation conditions. The research showed that automation of high workload tasks improves performance when the task difficulty is high. This is consistent with prior research that showed that effect of automation manifests itself in high task difficulty conditions.

In high workload automation, the subtask performance improvement was more evident on the subtasks that were not automated. This finding is aligned with resource allocation theories that state that resources are allocated to tasks based on the task demand. When the high workload subtask is automated, participants seem to utilize their mental resources to handle other components of the task. The high workload subtask automation did not show a significant effect on the accuracy performance of the automated subtask. Similar findings were observed in prior research where participants showed a significant improvement in a target detection task with support of automation, but their accuracy performance did not show a significant difference. It is important to note that performance improvements in the number of targets classified and classification score were observed after the introduction of a feedback. This finding is in line with the theory that resource allocation is affected by factors such as motivation.

Practical Implications

Utilizing Workload-based task selection can enhance human performance and contribute to reduction of both overall workload and the demands associated with individual subtasks. Automation of high workload subtasks yielded improved overall performance, reduced overall workload measured by performance measure, improved low workload subtask performance, improved part of the high workload subtask, and showed a reduction of subtask workload measured by ISA ratings and post-trial subjective responses.

While the task used in this research is a simulated version of a target elimination task, the insights gained from this study have potential to be applied to diverse domains. First, the study demonstrates that it is feasible to choose potential automation subtasks based on measurement of their workload. This could be practically employed in wide range of domains such as driving, aviation, and others that involve automating aspects of a task or system. Certainly, practitioners

should cautiously assess how this automation impacts human interaction with systems. However, this principle is applicable to all types of automations because improvement in overall performance often don't explain shifts in strategies and behavior. Hence, the current study showed that it is important to evaluate automation effects both at the subtask and overall task performance levels.

Limitations

There were several limitations in this research. First, the subject pool of this research was restricted to the Rice undergraduates. A broader range of participants could have yielded additional insights and perspectives. Second, the study did not use psychophysiological measurements in addition to the subjective and performance-based metrics. More specifically, the use of an eye tracking tool might have been useful in identifying the relationship between subtasks and collision events. Identifying subtasks participants were engaged in during moments of collision posed challenges during the research.

The third limitation was that the reliability of the automation may have been a bottleneck. The lack of significant effect in classification accuracy performance could have come from the reliability of the automation. The automation in this experiment was as reliable as a human can potentially be. If the human had sufficient time to process information, as in the low task difficulty conditions, it could perform as well as the automation. This was because the automation relied on the same information as the human and used similar processes to make decisions.

Future Directions

There are many possibilities for future research in assessing the efficacy of workload-based task selection for automation. In this study it was observed that there could be associations

and dissociation between workload measures. Future studies can utilize physiological measures to clarify why dissociations occurred in this context.

The current research used Blended Decision Making as an automation strategy. However, the effect of workload-based automation could vary depending on the types or levels of automaton used. Hence, it would be interesting to explore potential interaction between types of automation and the effects of high and low workload subtask automations. In addition, future research can be done by varying the reliability of automation and its effect on different types of subtask automation. For instance, participants have asked if it was possible to fully automate the action subtask, wherein classified enemy targets would be automatically destroyed. It would be interesting to see if increase in levels of automation change the effects of high and low workload subtask automations.

References

Afergan, D., Peck, E. M., Solovey, E. T., Jenkins, A., Hincks, S. W., Brown, E. T., ... & Jacob, R. J. (2014, April). Dynamic Difficulty Using Brain Metrics of Workload. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3797-3806). https://doi.org/10.1145/2556288.2557230

Bailey, N. R. (2004). The Effects of Operator Trust, Complacency Potential, and Task Complexity on Monitoring a Highly Reliable Automated System. *Old Dominion University. Available from ProQuest Dissertations & Theses Global*. (305101376).

Bailey, B. P., & Iqbal, S. T. (2008). Understanding Changes in Mental Workload During Execution of Goal-Directed Tasks and its Application for Interruption Management. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *14*(4), 1-28. https://doi.org/10.1145/1314683.1314689

Bolstad, C. A., & Endsley, M. R. (1999). Shared Mental Models and Shared Displays: An Empirical Evaluation of Team Performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 43*(3), 213–217. https://doi.org/10.1177/154193129904300318

Calhoun, G. L., Ward, V. B. R., & Ruff, H. A. (2011). Performance-based Adaptive Automation for Supervisory Control. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 55*(1), 2059–2063. https://doi.org/10.1177/1071181311551429

Deng, Y., Shirley, J., Zhang, W., Kim, N. Y., & Kaber, D. (2019). Influence of Dynamic Automation Function Allocations on Operator Situation Awareness and Workload in Unmanned Aerial Vehicle Control. *International Conference on Applied Human Factors*

and Ergonomics (pp. 337-348). Springer, Cham. https://doi.org/10.1007/978-3-030-20040-4_31

Endsley, M. R. (1987). The Application of Human Factors to the Development of Expert Systems for Advanced Cockpits. *Proceedings of the Human Factors Society Annual Meeting, 31*(12), 1388–1392. https://doi.org/10.1177/154193128703101219

Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned from Human–Automation Research. *Human Factors, 59*(1), 5–27. https://doi.org/10.1177/0018720816681350

Endsley, M. R., & Kaber, D. B. (1999). Level of Automation Effects on Performance, Situation Awareness and Workload in a Dynamic Control Task. *Ergonomics*, *42*(3), 462-492. https://doi.org/10.1080/001401399185595

Eysenck, M. W., & Keane, M. T. (2015). *Cognitive psychology: A student's handbook, 7th ed.* (pp. xviii, 838). Psychology Press. https://doi.org/10.4324/9781315778006

Fitts, P. M. (Ed.). (1951). Human Engineering for an Effective Air-Navigation and Traffic-Control system. *National Research Council, Div. of.*

Hart, S.G. and Staveland, L.E. (1988) *Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research*. In Hancock, P.A. and Meshkati, N. (Eds.) Human Mental Workload. Amsterdam. North Holland Press.

Kaber, D. B., & Endsley, M. R. (1997). The Combined Effect of Level of Automation and Adaptive Automation on Human Performance with Complex, Dynamic Control Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 41*(1), pp. 205-209. https://doi.org/10.1177/107118139704100147

Kaber, D. B., & Riley, J. M. (1999). Adaptive automation of a dynamic control task based on secondary task workload measurement. *International journal of cognitive ergonomics, 3*(3), 169-187. https://doi.org/10.1207/s15327566ijce0303_1

Leggatt, A. (2005). Validation of the ISA (Instantaneous Self-Assessment) subjective workload tool. In *Proceedings of the International Conference on Contemporary Ergonomics* (pp. 74-78).

Matthews, G., & Reinerman-Jones, L. (2017). *Workload assessment: How to diagnose workload issues and enhance performance*. Human Factors and Ergonomics Society.

Muñoz-de-Escalona, E., Cañas, J. J., Leva, C., & Longo, L. (2020, December). Task demand transition peak point effects on mental workload measures divergence. *In International Symposium on Human Mental Workload: Models and Applications* (pp. 207-226). Springer, Cham.

Okamura, K., & Yamada, S. (2020). Empirical Evaluations of Framework for Adaptive Trust Calibration in Human-AI Cooperation. *IEEE Access*, *8*, 220335-220351. https://doi: 10.1109/ACCESS.2020.3042556.

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation: An integrated Meta-Analysis. *Human factors*, *56*(3), 476-488. https://doi.org/10.1177/0018720813501549

Parasuraman, R., & Mouloua, M. (Eds.). (1997). Automation and human performance: Theory and applications. *Routledge*

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, *30*(3), 286-297.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making*, *2*(2), 140–160. https://doi.org/10.1518/155534308X284417

Parasuraman, R., Cosenzo, K. A., & De Visser, E. (2009). Adaptive Automation for Human Supervision of Multiple Uninhabited Vehicles: Effects on Change Detection, Situation Awareness, and Mental Workload. *Military Psychology*, *21*(2), 270-297.

Sheridan, T. B., & Verplank, W. L. (1978). Human and computer control of undersea teleoperators. *Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab*.

Spector, P. E., & Jex, S. M. (1998). Development of four self-report measures of job stressors and strain: Interpersonal Conflict at Work Scale, Organizational Constraints Scale, Quantitative Workload Inventory, and Physical Symptoms Inventory. *Journal of Occupational Health Psychology*, *3*(4), 356–367. https://doi.org/10.1037/1076-8998.3.4.356

Stapel, J., Mullakkal-Babu, F. A., & Happee, R. (2019). Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *60*, 590–605. https://doi.org/10.1016/j.trf.2018.11.006

Stevens, N.J., Salmon, P.M., Walker, G.H., & Stanton, N.A. (2018). Human Factors in Land Use Planning and Urban Design: Methods, Practical Guidance, and Applications (1st ed.). *CRC Press*. https://doi-org.ezproxy.rice.edu/10.1201/9781315587363

Wickens, C. D. (2008). Multiple Resources and Mental Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394

Wright, M. C., & Kaber, D. B. (2005). Effects of Automation of Information-Processing Functions on Teamwork. *Human Factors*, *47*(1), 50-66. https://doi.org/10.1518/0018720053653776

Yan, Z., Kantola, R., & Zhang, P. (2011, November). Theoretical Issues in the Study of Trust in Human-Computer Interaction. *In 2011 IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications* (pp. 853-856). https://doi: 10.1109/TrustCom.2011.114.

Yeh, Y.-Y., & Wickens, C. D. (1988). Dissociation of Performance and Subjective Measures of Workload. Human Factors, 30(1), 111–120. https://doi.org/10.1177/001872088803000110