Influence of Display Designs on Spaceflight Supervisory Control for Operations and Training

by

Savannah L. Buchner

B.S., University of California Davis, 2019

M.S., University of Colorado Boulder, 2022

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Ann and H.J. Smead Department of Aerospace Engineering Sciences

2025

Committee Members:

Dr. Allison P. Hayman, Chair

Dr. Torin K. Clark

Dr. Hanspeter Schaub

Dr. Joseph B. Lyons

Dr. Stephen K. Robinson

Buchner, Savannah L. (Ph.D., Aerospace Engineering Sciences)

Influence of Display Designs on Spaceflight Supervisory Control for Operations and Training

Thesis directed by Dr. Allison P. Hayman

Spaceflight involves humans working or teaming with robotic or semi-autonomous systems to perform complex tasks, such as satellite rendezvous, docking, or Earth imaging. Operators often interact with these systems as remote supervisors, by monitoring them, providing high-level objectives, or intervening as needed. These teaming tasks can present many challenges for the operator, and current displays and training practices may not be sufficient for these future complex operations. Motivated by this, a series of experiments was designed to investigate virtual reality (VR) for satellite operations and training, as well as to investigate how displays are used to make decisions when teaming with autonomous systems for operations.

In Aim 1, a systematized literature review was conducted on how to display menus and text in VR, as there was a lack of dedicated VR display design principles that are applicable for operational use. From these display design guidelines, a series of displays were developed, and 3 VR experiments were performed. In Aim 2, VR and screen-based 3D Visualizations were compared to traditional displays during a spacecraft monitoring task. In Aim 3, this research was extended to investigate supervisory control, where the operator had the ability to intervene and take action, using the same display categories. Both Aims 2 and 3 address the gap in the literature surrounding using VR for monitoring and supervisory operations. These experiments found that 3D visualizations (either on a screen or in VR) provided benefits over traditional displays for the monitoring or supervising of spacecraft operations, but that VR did not provide additional benefits. Specifically, visualizations were found to improve situation awareness for monitoring tasks and improve performance and subjective utility for supervisory control tasks. In Aim 4, these display modalities were studied for use as a training paradigm. VR was found to be a promising training modality for spacecraft operations, as it improved level 2 situation awareness (comprehension) and usability in operations and had higher perceived utility. This addressed the lack of literature on VR training and supervisory control operations. The fifth, and final aim, of this work was an assessment of how and what display components are used on a screen-based display when assessing correct decision making and trust. Gaze behaviors could explain decision making, notably operators spent the majority of their time on the left side of the screen, but these cannot fully explain trust.

The main contributions of this work includes the synthesis of VR literature on menus and text for operational contexts and the use of novel displays for supervisory control operations and training. VR's application to operational contexts such as remote monitoring operations, remote supervisory operations, and training for supervisory control is particularly understudied. In all of these applications, an operator's understanding of key information is crucial for success, which can be achieved through effective display design for training and operations. One specific operational context that is further understudied in the context of display design is spaceflight operations. As space missions grow more complex and novel display technologies become more accessible, understanding how to design displays to promote appropriate usage, and consequently, facilitate mission success, is critical. For the first time, this work synthesizes display design principles for VR in operational environments such as spaceflight. Critically, this can inform the adoption of VR in these contexts. Furthermore, this work examines operators' usage of current supervisory displays and how this usage affects their decision making and trust, allowing for an understanding of how to improve future display designs. Together, these contributions advance our understanding of how operators use displays to complete their objectives and subsequently how to better design future displays for spaceflight supervisory control.

Acknowledgements

There are many people who have made this work possible, and I am grateful for all of their support.

First, I would like to thank my advisor, Prof. Allie Hayman, for her unwavering support and patience throughout this PhD. I knew I could always count on leaving a meeting with her feeling more confident, and that she would have some sort of positive outlook on whatever problem I was up against. I'm fortunate to have worked with someone who constantly pushed me to be a better scientist and researcher, but also encouraged me to have fun along the way. I'm thankful for everything I've learned from Allie and am proud to be a Laikanauts graduate.

I'd like to thank my committee, Prof. Torin Clark, Prof. Hanspeter Schaub, Dr. Joe Lyons, and Prof. Steve Robinson, for their support and willingness to provide feedback, assist with experimental design, and provide mentorship throughout this process. Torin was a constant calm presence in many meetings and throughout this process, providing constant head nods during presentations, and providing lots of advice for stats and skiing. Prof. Robinson was one of my first research advisors as an undergrad at UC Davis, and I am thankful for his support as I continued to CU for grad school.

This work would not have been possible without the people who helped develop, debug, and commiserate over Unity code with me, and those who dedicated time and effort to assist with running experiments. I'm also grateful to the participants who willingly took part in the experiments.

The Bioastronautics lab, both past and present, has been the most amazing lab I could ever imagine working in. Their research ideas, thoughtful discussions, and sanity checks on my crazy ideas made this research possible. More importantly, though, they offered kindness and friendship. I consider myself lucky to work with all of you and help contribute to the lab. I'd especially like to thank the middle desk for entertaining me with stories and being the best desk mates.

To my parents, who have always encouraged me to shoot for the stars and to be curious about and explore the world around me.

To all my friends and family - thank you for keeping me sane throughout this process. From backpacking adventures to long hikes and runs, ski weekends, board game nights, and long-distance phone calls. I am incredibly lucky to have such amazing people in my life.

This work has been supported by SpaceWERX STTR grant FX21S-TCSO1-0345 and SpaceWERX STTR grant AWD-23-04-0184.

Contents

Chapter

Appendix

Table

Figures

Figure

Chapter 1: Motivation

Spaceflight, like many exploration and operational paradigms, involves humans working or teaming with robotic or autonomous systems to perform complex tasks, such as satellite rendezvous, docking, and Earth imaging and detection of targets. Currently, the majority of these activities involve humans on Earth remotely working with automated or semi-autonomous systems in space, such as satellites. Operations for remote tasks where the operator and system are not co-located, like with mission control, can be difficult. Due to time delays and communications limitations, operators (i.e., the individuals responsible for working in these settings) cannot directly command or control these assets unless in direct communication; instead, operators act in a remote supervisory control paradigm and send intermittent commands or set goals for the system to accomplish [1, 2].

The allocation of responsibility between humans and the systems is not consistent between use cases. Satellite operations, such as controlling orbital burns, satellite attitudes, or rendezvous, and proximity operations, are often made with a human-in-the-loop providing commands to the satellites, but not directly controlling them. Further from Earth, for deep space exploration and planetary rovers, humans often command actions like orbital burns, camera targets, or rover paths, but cannot directly fly or drive the vehicles [2]. Supervisory control is also used for current human spaceflight missions, such as the docking of resupply vehicles to the ISS.

Outside of commanding satellites, supervisory control is often used when working with satellites to gather or interpret data. Autonomous agents have been placed on Earth observational satellites to assist in determining imaging targets, such as military and ship movement, wildfires, or harmful algae blooms, and automatically process data in real time before sending it back to Earth [3, 4]. However, humans are still on the loop when teaming with these systems and act in a supervisory position. Humans may help predetermine areas of interest, review images to ensure the autonomous system is accurate, or schedule follow-up images [4].

These types of operations can lead to challenges for the operator. Remote separation leads to the operator's perceptual processing abilities being decoupled from the environment, decreasing situation awareness (SA) [5]. Additionally, inappropriate trust can lead to misuse or disuse of the autonomous systems [6]. Overall, these factors reduce mission effectiveness and task completion. A key component for these operations is the display, or how the operator is able to receive the required telemetry and data. Current mission control displays can be ineffective for monitoring [2, 7] and increase the operator's workload, especially when trying to process 3D data on 2D screens [8–10]. It has been recommended that improvements to the display and mission control be made for future, complex satellite operations [2]. Beyond this, when teaming with autonomous systems, it is not well understood what display components are most useful to promote accurate decisions, appropriate reliance, and calibrated trust.

Another key component of successful operations is training. Insufficient or ineffective training can harm mission performance, potentially leading to loss of mission, particularly for the complexity of these future planned operations. Current training consists of presentation-based learning, simulations, and on-the-job training [11]. However, high-fidelity simulators are often costly to use and may not easily be modified to cover multiple scenarios, and it has been suggested that new technologies may be able to improve current training practices.

With the advancement of virtual reality (VR) technologies, there has been an increasing interest in VR for operations and training to combat these issues. VR offers the ability to present tasks immersively and has been shown to improve aspects of operations like increased SA and reduced workload for some environments [5, 12–14]. Additionally, it has been a promising training modality for a variety of operational tasks [15–18], but there is still an open question of how VR can translate to supervisory control tasks. Additionally, while VR is proposed as a solution, there is a lack of consensus in the literature about how to optimally design displays for VR, and a gap in established display design principles. This may lead to conflicting results in the literature on potential VR benefits [14, 19].

This motivates the following thesis focus on display designs for remote supervision of spaceflight applications. This includes VR display design principles, VR for monitoring and supervisory satellite operations, and VR for training. Beyond this, it will also study the type of information displayed to promote appropriate decisions, reliance, and trust for teaming.

Chapter 2: Background

2.1 Supervisory Control

Human-autonomy teaming, where a human operator is working remotely with an autonomous system to achieve a shared goal [20–22], is a challenging yet important modality for future exploration or operational environments, including spaceflight. For spaceflight, these environments are characterized by highly trained operators, where there are often consequences to safety and performance due to improper action, particularly when there are uncertainties in the state of the system.

Working with autonomous or robotic systems is often classified by the level of control afforded to the human operator. While not a perfect representation [23], one such way of categorizing the level of control has been developed by Sheridan [24]. The control modality most relevant to this research and spaceflight operations paradigms is supervisory control and can be seen in Fig. 2.1. This is where the operator provides intermittent commands, but the majority of the control is through the computer without human input. While there are many different aspects to these control paradigms, this research will primarily focus on the display block.

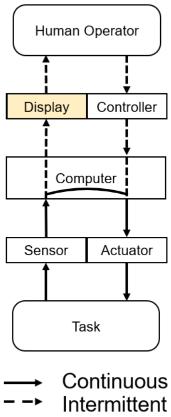


Figure 2.1: A representation of supervisory control operations. The focus of this research is on the display component.

Supervisory control is found in various forms in many exploration, transportation, industrial, military, and medical contexts [1, 23, 25]. For example, in commercial aviation, pilots often spend the majority of the flight monitoring the autopilot system. They occasionally provide inputs, such as setting the destination or altitude, and only take over in emergencies. In addition, supervisory control is often used in situations in which the human operator and autonomous system are not co-located. This may be due to the safety of operators when the autonomous system is working in a dangerous environment, like space, mining, underwater, or military zones [2, 26]. In many spaceflight operations, the human operators are on Earth working with assets in space. In these settings, time delays are also present from the order of seconds to hours, eliminating the ability to directly command the asset. This time delay is one of the initial early drivers of supervisory control, since requiring continuous control with a time delay present produces instability in the system [1].

Supervisory control involves many cognitive demands as the human operator has multiple functions, including planning the task and teaching the computer how to perform it, monitoring the autonomous system to ensure the task is going as planned, intervening to update directions, and then learning for future work [24, 25]. Monitoring is a very important aspect of remote supervision [24], as it enables operators to understand system states, anticipate future issues, and quickly detect and respond to failures [1]. Monitoring is also how an operator spends the majority of their time. During intervention, operators can provide new subgoals or aid in decision making for the autonomous system.

Remote operations are where operators and the systems they work with are separated spatially and potentially temporally (i.e., due to time delay of sending information), like in a spacecraft mission control center, and are the focus of this work. Remote operations can be challenging due to the lack of environmental context and the decoupling of processing abilities from the physical environment [5, 27], creating a host of new challenges for human information display, including compromised **situation awareness** (SA) [5, 28]. SA can be defined as having 3 levels where level 1 is “the perception of critical elements in the environment”, level 2 is “the comprehension of their meaning”, and level 3 is “the projection of their status into the future” [29]. Low SA can make operations a difficult and cognitively challenging task, reducing mission effectiveness and task completion. This problem is compounded as monitoring operations are often already correlated with lowered SA [30].

Additionally, supervisory control is influenced by the need for effective human-autonomy teaming, which is important to ensure mission success, performance, and appropriate use of the system. One critical aspect of human autonomy teaming is **trust**, or “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” [6]. Inappropriate trust in a particular system can lead to overtrust or distrust, and supporting appropriate trust is critical to avoid misuse and disuse of an autonomous system. Inappropriate use of the system could lead to inadequate performance or a mission failure. Trust is especially critical for supervisory control operations, where the autonomous system has leeway to make its own decisions and recommendations.

2.2 Current Practices for Spaceflight Applications

2.2.1 Satellite Operations

The operation of satellites is a monitoring and supervisory operational domain that is often understudied. Beyond the challenges present with supervisory control, future proposed mission objectives may present additional difficulties. For example, on-orbit inspection and servicing tasks are becoming increasingly desirable, but involve multiple satellites in close proximity [31–33]. The relative trajectories of the satellites must be quickly understood by operators in order to avoid collisions, but are governed by flight dynamics that can be complex and non-intuitive. Additionally, the close proximity between satellites during these tasks results in less time to make decisions and send commands. Critical events, such as for collision avoidance maneuvers, may necessitate actions to be made in seconds [34]. The complexity of the system increases with uncertainties in the states of the satellites (i.e., their exact positions, velocities, and orientations) and can be especially significant when repairing a dead or injured satellite.

Current satellite operations take place in a mission control room, an example of which is seen in Fig. 2.2. While different organizations have different protocols and displays, the majority of displays consist of densely packed telemetry data, 2D graphical representations of values over time [2], and many times different subsystems have their own displays and alarms [24]. The data presented comes from a variety of sources, and can include past, present, and/or predicted future data [1, 27]. Additionally, mentally processing 3D data that is represented in 2D can increase the operator's workload [8–10]. **Workload** can be influenced by the task and environment, as well as the operators' skills, behavior, or perception, and is defined to be the cost incurred by an operator to achieve a particular performance level [35]. Workload can include aspects like mental, physical, and temporal demand, as well as be related to effort, frustration, and performance. Additionally, workload can be related to the amount of data that needs to be processed [7], and having to quickly comprehend and react to the data can further increase workload. To improve these displays for next-generation supervisory control rooms, it has been suggested to develop systems to manage the large amounts of data and provide only needed or requested information, which would improve SA and reduce workload [2].

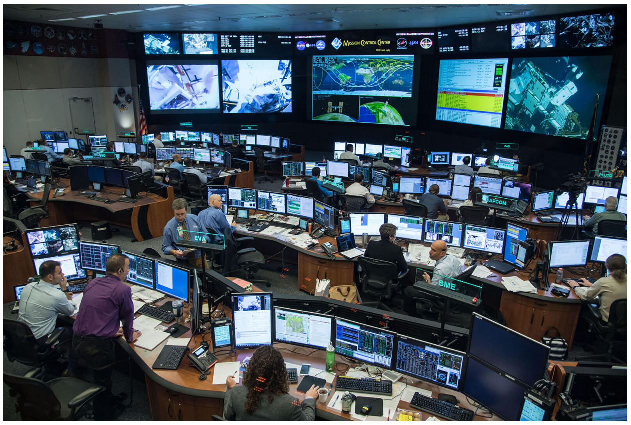


Figure 2.2: ISS mission control center. Each operator has multiple computer screens to monitor. There are multiple operators in the control room and there are additional back rooms with more operators. (Source: NASA)

2.2.2 Training for Satellite Operations

Due to the complexities and risks of current and planned satellite operations, it is important that operators have appropriate training, as failures can lead to the loss of spacecraft. Training is defined as the process by which individuals can gain job-relevant skills, knowledge, or competency [11]. Effective training ensures operators have an accurate sense of system capabilities and adequate practice in putting together and understanding complex information. This is often developed through repeated interaction. Inadequate training can make it more difficult for operators to maintain SA during the task [11], which can harm mission performance.

Current supervisory control training practices often involve classroom education, simulators, and on-the-job training (OJT) [11]. Spaceflight operators in training are often taught along these lines. For example, command controller training at one satellite mission control facility consists of 10 weeks of classroom-based lectures [36]. In contrast, NASA space shuttle mission controllers, in some years, completed over 100 simulations to become certified [37]. Other organizations may take between 3 months and a year to complete training [38–42]. The training times and requirements, particularly for classroom lectures, may be increased for training operators who do not have previous experience or an educational background that is necessary to complete the task.

Figure 2.3: Example of a current air traffic control training simulator (Source: Air Force)

Practice-based training can include simulators and OJT and can be a costly but important component of training. This type of preparation is necessary to build mental models, skills, and judgment to complete a task without errors in both nominal and off-nominal situations [1]. Mental models are “mechanisms whereby humans generate descriptions of the system purpose and form, explanations of the system functioning and observed system states, and predictions of future system states” [43]. Anticipating future states, enabling pattern recognition, and information-seeking have all been attributed to mental models [44]. Additionally, mental models can aid in improving SA, particularly the higher levels [45]. For many operations, especially in novel environments, it is critical for operators to include consideration of imperfect information and the possible impacts of that uncertainty in mental models of their actions. For example, in satellite operations, there may be uncertainties during a satellite burn or uncertainties inherent in a given sensor. These imperfections in knowledge of the current spacecraft state may affect the appropriate actions required and can impact future states. It is apparent that appropriate mental models will include an appreciation of uncertainties in the system, and that this must be taught as part of the training program.

However, even the most comprehensive current training environments have their downsides. Simulator training devices, particularly high-fidelity ones, are expensive to design and may require many personnel to run. In addition, simulators can become obsolete compared to current technology due to the large cost of updates [37]. It has been suggested that incorporating new technology into training as an alternative to simulations can reduce total cost and time [37].

2.2.3 Human Autonomy Teaming with Satellites

Beyond operations, humans may also be working with autonomous systems on-board the satellite to aid in processes like image identification. When working with autonomous systems, operators may verify or confirm the autonomous systems’ recommendations. Other times, operators may follow a recommendation of the system, ignore the system’s recommendation, or not take action if no alert is given. Compliance is an active form of agreement, where an operator positively responds to alarms [46], and reliance is the passive form of compliance, where an operator does not correct or override a system when there is an absence of alerts.

These constructs are influenced by trust, but not determined by it [6]. Verification can be a relatively objective measure or distrust [47–49]. Reliance is related to the trust in automation exceeding an operator's self-confidence, and high reliance can be representative of high trust, high workload, complacency, inadequate training, or high risk situations [50, 51]. Likewise, reliance and compliance are affected differently based on the type of errors exhibited by the system. A false alarm affects both, but a missed detection affects only reliance [46].

2.3 Alternative Display Designs

Displays are a critical component of successful supervisory control operations, as it is how the operator receives the information from the system. As described above, current displays and training paradigms may not be adequate for the additional challenges of future missions, and improvements need to be made. For satellite operations, there is a need for improved understanding and SA, which changes to the display, such as introducing VR or 3D visualizations, may be able to provide. For operational training, there is a need for low-cost, high-fidelity simulations, which VR may also be beneficial for. However, there is a gap in the literature in understanding how these display changes to VR can affect spaceflight operations and operations. Finally, for promoting human-autonomy teaming and appropriate trust, reliance, and decision making, there is a need to better understand what information is currently used in order to include the appropriate information in future displays.

2.3.1 Virtual Reality

VR has been proposed as an alternative to traditional 2D display interfaces in many situations that would benefit from increased immersion in the environment and 3D interactions, such as operations and training. An example of a virtual reality setup is in Fig. 2.4. Such displays increase telepresence, which is the feeling of being present in an environment other than where one is physically [52]. This is due to the immersive nature of VR, where immersion is defined as "the extent to which the computer displays are capable of delivering an inclusive, extensive, surrounding and vivid illusion of reality to the senses of a human participant" [13]. Throughout

Figure 2.4: A demonstration of a virtual reality setup being used for satellite orbit visualizations and training (Source: Space Force)

this work, immersion is provided through a head-mounted display (HMD) VR system. VR may enhance performance through the improvement of perception, increased field of view, and the ability to change viewpoints without the loss of telepresence [5, 12, 53]. The ability to change viewpoints (i.e., teleportation) offers improvements in performance over a fixed viewpoint but comes with the risk of increasing cognitive load [54]. Furthermore, the ability to increase the field of view through natural head motions in VR can help improve collision avoidance and understanding of future vehicle states over the reduced field of view 2D displays.

2.3.1.1 Virtual Reality for Operations

Most of the prior research into using VR and remote operations has focused on direct teleoperation, where the human operator manipulates or controls a robot [5, 14, 54–56]. This includes situations such as operating undersea robots [53], remotely driving a car [57], or remotely manipulating robotic arms [14]. Whitney et al. [14] found that using VR to complete a teleoperation task with a robotic arm led to faster completion time, lower workload, and improved usability compared to traditional monitor and keyboard interfaces. Similarly, Elor et al. concluded that stereoscopic VR displays led to faster completion times, increased usability, and increased perceived presence and performance over desktop displays in an underwater capture task. However, in this study, no differences were found in workload [53].

Consistently, studies have found that VR displays can improve depth perception and collision avoidance, lead to faster task completion, increase the sense of presence, increase usability, and reduce perceived effort compared to the 2D displays [5,12,14,53]. In prior work, immersive displays has been studied in the context of data visualizations and immersive analytics, and have been shown to improve estimations of depth, size, distance, cluster identification, and trajectories [58–62], further promoting the hypothesis that VR may be useful for monitoring and supervisory operations. However, there are mixed results on the effects of VR on task performance. For example, some studies find VR does not change workload [53] while others reported a decrease [14,63]. With regards to SA, some studies have found improvements [57], while others have reported no differences [64] between VR and screen displays. Some of these differences may be attributed to display design choices, as there is a lack of validated design principles for VR operations, and ineffective display design choices may result in reduced performance. In addition, the task type or complexity may also influence the outcome, as VR may not be appropriate for all tasks. Finally, differences may also be attributed to inconsistencies in how these variables are measured. For examples, using subjective [64] or proxy measures of SA [57] may not capture the same aspects of SA [65].

The direct control paradigms that have been studied in depth may not be appropriate for all future operations, such as spaceflight, where time delay and bandwidth limitations inhibit direct control. Some of the benefits seen with direct control, such as improved collision avoidance, may still apply. However, other benefits may not be as applicable due to the differing control authority and cognitive demands on the operator. Manual control tasks tend to have higher workload [66], which VR may be able to reduce [14, 63]. Unlike direct control, monitoring tasks already have a low workload [66], so VR might not reduce it further [19]. However, it is still important to study VR for monitoring tasks, as successful monitoring relies heavily on an operator's SA, which VR may increase.

As such, a gap in prior work is the use of VR for remote supervisory and monitoring operations, particularly with a lack of research on satellite operations. VR has been proposed for use in operations for many monitoring control rooms, such as for spaceflight [67,68], maritime [19,69], and air traffic control [70, 71]. For satellite operations in particular, VR has been suggested as a way to improve environmental context and allow operators to better understand 3D orbits. While there has been research focusing on these monitoring applications, primarily for air traffic control, using augmented reality [72–76], there have been limited experimental studies into VR for supervisory operations.

Lager et al. compared the use of traditional 2D GUI, 3D screen-space GUI, and 3D VR GUI to remotely monitor autonomous surface vehicles and found that users were better able to detect collisions and had improved SA in both the 3D and VR displays compared to the 2D display. The 3D GUI had significantly reduced cognitive load (as measured through a proxy variable of a secondary task) compared to both the VR GUI and 2D GUI. However, participants subjectively felt that if they had many hours of training, the VR display would be best for the task, indicating that VR has the potential for monitoring applications [19]. Other research has studied VR for maritime control room monitoring and has found that VR could replace complex monitoring dashboards [69]. Although an experiment was conducted, no comparisons were made between VR and current maritime displays, representing a gap in VR monitoring research and understanding how VR can facilitate improvements. A different study comparing VR to physical displays for the monitoring of autonomous cars found that VR increased task load and simulator sickness, and decreased usability [77]. However, the authors acknowledge that these differences are likely attributed to the hardware used (i.e., headset weight and resolution) and the fact that their VR display was not designed or optimized for use in VR, indicating that it is important to design with VR in mind. Finally, for air traffic control applications, VR has been demonstrated to reduce the number of errors and aid in identifying dangerous situations compared to the typical 2D view controllers see [71], but no comparison was made to a 3D view, making it unclear if the benefits come from increased immersion or 3D visualizations.

2.3.1.2 Virtual Reality for Training

VR has been studied and used for training in medicine [15, 16], firefighting [17], human spaceflight [78], education [79], among other disciplines [80, 81]. VR has been demonstrated to facilitate skill transfer to the real world [15], and lead to improvements in mental models [82] and perceived learning [83].

VR can provide a customizable, adaptable, and immersive environment to simulate tasks that might otherwise be costly, unsafe, or unintuitive to replicate in the real world [81]. This allows users the opportunity to train for a variety of scenarios and in a variety of environments, both nominal and off-nominal. During VR training, trainees can make mistakes with no real-world risk to safety or expensive systems and build a better understanding of how actions and outcomes are connected. VR has also been shown to induce realistic stress responses, which can lead to operational realism and enhance training [84]. Although using immersive displays for training and learning is generally promising across a wide variety of fields, little work has been done to apply these principles to remote supervisory operations, where the consequences of inappropriately conveyed uncertainty in this environment are profound for the operator's human perception and performance.

2.3.1.3 Virtual Reality Limitations

While VR has many potential benefits, it also has some limitations that may affect how much benefit it can provide for a particular application. Some of these are related to limitations in the hardware, which is actively improving as technology progresses, but still must be considered to ensure current operational use and operator buy-in.

VR displays can lack the resolution to display blocks of text in a readable way to properly understand large chunks of data [85], which may be required for the supervision of autonomous systems [24]. Operator buy-in and susceptibility to cybersickness are influenced by refresh rates, resolution, and other headset properties as well as display design choices, including viewpoint selection, field of view, amount of control over the environment, and headset properties [86, 87]. Finally, VR headsets can be uncomfortable to wear for long periods of time, due to eye strain and pressure points [88].

Many of these technological limitations can be overcome through display design choices. Display design choices can also reduce other common VR limitations. Information overload and misinterpretation of the signal may be exacerbated in a VR-based environment representation [89]. Inappropriate choices for interaction and selection techniques can result in selection errors, longer competition time, fatigue, and decreased usability [90,91], which may be problematic for operational uses. Teleportation can increase an operator's cognitive load [54], and mismatches between the expected and actual viewpoints can decrease SA and cause simple tasks to be challenging [5]. Finally, VR may cause distractions due to its immersive and novel nature, which can cause learners to not focus on understanding the desired lesson [79] or lead to distractions during operations.

However, VR is a developing technology, and there is still a lack of validated display design principles that inform what choices are most effective for a certain application and a need for continued research on the user experience in VR [92]. Some companies have released guidelines, such as Google [93] and Oculus [94]; however, these are often geared towards video game design and may quickly become outdated with technology progression. While these guidelines may help inform operational display design, they are often designed to improve constructs such as immersion, which may be counter to some of the needs of operations. Little guidelines exist on how to display text, which is common for these supervisory control displays, or how to interact with the system.

Without ensuring the display is designed appropriately for the potential application, it is also hard to understand and research whether VR is beneficial for that application or if the benefits are just not realized due to ineffective design choices. Thus, subjective assessments of VR, such as display usability or utility, are important; if users find VR uncomfortable or unusable, operators may avoid its use for both operations and training, even if it can lead to better outcomes.

2.3.2 Human-Autonomy Teaming and Decision Making

Working with complex autonomous systems also requires novel ways to display information. This is needed to ensure that operators are able to make appropriate decisions and have calibrated trust when working with or reviewing the autonomous system's suggestions, as the display and type of information provided can influence trust, reliance, and decision making [6]. Open areas of research for displays often focus on transparency and explainability. Transparency refers to "the descriptive quality of an interface about its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, plans, and reasoning process" [95]. Transparency has often focused on providing users with appropriate information explaining the system's reasoning before taking action [96–98] and may offer better usability [99], acceptance [100], more appropriate trust [101], and reduced workload [102]. Stowers et al. found that increasing the amount of information increased the number of correct decisions, but decreased usability and lengthened response time [100, 103]. Additionally, the level of detail in the display was found to produce a speed-accuracy trade-off, where more detail increases accuracy, but decreases response speed [104]. One concern for increasing the transparency is that displaying more information may increase display clutter and overwhelm the operator's processing ability [105,106], and it has been suggested that the amount of information must be adjusted by the available decision time [107].

While much of the existing literature focuses on the amount of transparency, little focuses on the type of information and what information operators use to make decisions. When designing these future displays, it is important to include the information that is most useful to the operator in enabling them to make correct decisions in the amount of time they have available. By determining what information operators are processing when making decisions, better displays can be designed to promote the appropriate information. A challenge, though, is understanding what type of information to include and what contributes most to appropriate decisions, representing a gap in the literature.

While there are different ways to understand what information is being used to process decisions, one unobtrusive way to do so involves gaze or eye-tracking metrics. Eye movements have been used to infer mental processes during decision-making in areas such as behavioral economics [108], psychological sciences [109], and piloting tasks [110] and have been suggested to be a powerful way to assess cognitive processes [111,112]. This may extend towards human-autonomy teaming and being able to understand what areas of a display are being used to make appropriate decisions. Most of the previous work involving human-autonomy teaming and gaze has been for trust and has not considered the areas or the meaning of the areas the operator is looking at. Subjective trust ratings have been correlated to total fixation durations, total fixation count, and number of transitions, as well as metrics like rate of transition and scan path per section. It was found that lower trust (and reliability) lead to longer, and more fixations than with higher trust [113]. Likewise, for automated driving, negative correlations were found between fixation frequency and trust [114, 115], and low gaze dispersion has been associated with higher trust and less monitoring [116].

Beyond gaze, other behavioral metrics may be related to an operator's decision-making. As discussed previously, reliance and compliance are metrics that can be calculated for a period of time to understand appropriate usage. Compliance is calculated by the rate of agreement per block of time, and reliance is the rate of not overriding automatic control in the absence of alerts per block of time [117]. While related to the use of the system, these do not consider the correctness of an operator's decision, but may be related to an operator's trust in the system and are important to consider to ensure the system is used appropriately. Previous research is inconclusive about how the display affects reliance. In a study for driving simulations, display design and realism were found to influence trust, but not reliance [118]; however, other studies found that an increase in realism increased trust and led to a greater reliance [119]. Overall, there has been little work understanding how specific interface features affect reliance and system usage [6].

Chapter 3: Investigative Rationale and Specific Aims Summary

The following gaps, as identified through the literature review, motivate this thesis on display designs for training and operations in supervisory control paradigms.

Gap 1: There is a lack of dedicated design principles for VR operations in literature. In order to design appropriate VR displays, especially for aerospace applications, it is important to identify relevant human factor principles that can be applied in VR.

Gap 2: Few studies have explored the use of VR for monitoring and supervising autonomous systems. To demonstrate the efficacy of VR in these control modalities, the benefits of VR that come from the increased immersion or 3D visualizations need to be assessed.

Gap 3: Most studies into training with VR have focused on simulating environments for manual control tasks. There is a lack of research in VR for training of supervisory control tasks.

Gap 4: It is known that trust and decision-making depend on the content, details, and format of the display. However, there is a lack of research into understanding what specific contents of a display are being used to make the appropriate decision, and how this relates to trust, particularly for remote supervision.

The proposed thesis will investigate display designs for complex spaceflight supervisory control operations and training. It will investigate VR as a display modality for training and operations in remote monitoring and supervisory paradigms using a satellite servicing and repair mission scenario. In addition, this work will further study display design for satellite teaming, with a focus on display elements that influence trust during a satellite image classification task. The following aims comprise this thesis:

Aim 1: Establish a Coherent Set of VR Display Design Principles Derived from Literature

Summary of work: A systematized literature review is conducted on VR-relevant display design principles, with a focus on menus/interaction and text. These synthesized guidelines are used in the development of the VR displays used throughout the rest of the aims. This Aim addresses Gap 1.

Aim 2: Investigate the Effects of Visualization and Immersion in Displays for Remote Monitoring

Operations

Summary of work: Three display designs with various degrees of immersion and visualizations are compared in monitoring satellite operations. The specific constructs of SA, workload, usability, and subjective utility will be assessed between display designs. As monitoring is an important aspect of supervisory control, this aim will help assess the use of VR during operations. In this aim, we hypothesize that visualizations will improve SA, lower workload, and improve usability and subjective utility over displays without visualizations. In addition, we further hypothesize that immersive VR displays will provide additional benefits over visualizations. This aim addresses Gap 2.

Aim 3: Investigate the use of VR for Remote Supervisory Control Operations

Summary of work: This aim extends upon the work of Aim 2, but considers situations in which operators have some degree of control authority. It considers the same three display types for a similar task, but operators can now make intermittent commands. This aim aids in understanding the impact of VR on operations with increased control authority. We hypothesize that visualizations will improve SA, increase usability and utility, and improve performance. Additionally, we hypothesize that immersion will further improve SA, increase usability and utility, and improve performance. Based on the results from Aim 2, we hypothesize no differences in workload due to the typical low workload of supervisory tasks and anticipated appropriate display design implementation. This aim also addresses Gap 2.

Aim 4: Investigate VR Training to Improve Operations for Remote Supervision

Summary of work: This aim explores the degree to which training in VR for remote supervisory tasks improves the operator's understanding of uncertainty compared to traditional training modalities, and is done in conjunction with Aim 3. The benefits of training in a different modality than operations are conducted in are assessed. Same modality (i.e., training in a traditional display and operating in a traditional display) and cross-modality training (i.e., training in VR and operating in a traditional display) are compared based on subsequent performance in a traditional display. We hypothesize that training with 3D visualizations will lead to better SA and performance in the operational trials, and have a higher utility rating. Furthermore, we hypothesize that VR will lead to further improvements in these metrics than 3D visualizations alone. We do not believe there will be a difference in workload scores or usability based on training conditions. This aim addresses Gap 3.

Aim 5: Investigating Display Design Elements that Lead to Calibrated Trust and Appropriate Decision Making During a Satellite Human-Autonomy Teaming Task

Summary of work: This aim studies how operators use a display during operations, specifically focusing on the human-autonomy teaming component of operations. Instead of controlling the satellites, the operator works with an autonomous system to identify if images taken by a satellite contain objects of interest. Operators have the ability to blindly trust the system, or to review the telemetry the satellite obtained in making the decision. We track human behavior and identify what information they viewed when making their decision. This investigates what types of information people use to make their decisions, as well as whether these decisions are related to their trust. This aim addresses Gap 4.

The findings from Aim 1 will influence the VR display design used in Aims 2-4. Additionally, aims 2-4 will use a similar task and environment, with the main difference between the degree of control authority. Aim 5 will shift the focus from operating satellites and VR to working with satellites to achieve a common goal on a screen-based visualization.

Chapter 4: Aim 1: VR Display Guidelines

4.1 Introduction

The objective of this aim is to develop a set of VR literature-derived display design principles that can be used for designing systems for operational settings. This review focuses on head-mounted VR devices, as opposed to other immersive systems like augmented reality (AR) or CAVE Automatic Virtual Environment (CAVE, a projector-based VR system). Additionally, it will focus on two elements that may be unique or different for operational use cases, as opposed to VR use cases like video games that are commonly studied. This includes menus, or how a user can access information, and the use of text, or how a user can read and interpret essential data. Operational use of VR often requires high accuracy and lots of text, which may lead to the optimal choices being different than VR video games or other use cases. Oftentimes, video games and similar applications promote immersion, aesthetics, and usability, as opposed to accuracy and speed.

This aim focuses on menus/interactions and text since the initial literature review identified these as requiring unique considerations for operational VR displays that are different than existing human factor principles, operational guidelines, and VR video game design principles. Two systematized literature reviews are conducted. The reviews are modeled after Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines [120]. Three databases are searched, including Web of Science, Compendex, and Inspec. This provided a variety of sources and identified studies across different disciplines. This approach was applied to 2 different aspects of operational display design: Menu and Interactions and text. The results of this aim are used in the development of the VR interfaces for Aims 2-4.

4.2 Menu and Interactions

The first area of interest is the use of menus and how users interact with them. Menus are critical to facilitate changing the state of the interface through the selection of one or more options. However, there are different ways of presenting and selecting the menu options, which this review

Table 4.1: Search terms for menu review. Note * indicates terms with multiple variations

Topic	Search Term	Location
	Menu*	Title
Technology	VR OR “Virtual Reality”	Title/Abstract/Keywords
Study Type	participant* OR Subject* OR user* OR study	Title/Abstract/Keywords
General	compar* OR evaluat* OR effect* OR explor*	Title/Abstract/Keywords

is interested in understanding.

4.2.1 Methods

A literature review is conducted using the search terms in Tab. 4.1. The initial search resulted in 179 articles. The modified PRISMA procedure is documented in Fig. 4.1. The metadata is screened to remove papers that are not in English, duplicates, published prior to 2010, and not peer reviewed. During the title and abstract screening phase, results are excluded for no VR application, no menu development, or no menu testing. Only the author screened the titles and abstracts of all papers. The inclusion criteria for the full text review is having a comparison of menus. Menu comparison testing is important to be able to assess the menu performance and understand the benefits of certain menu types in VR. Interaction types (i.e., how to navigate the menu) are not directly searched for in the review processes, but are identified in the resulting papers. Following the full text review, 7 papers are included, with 1 additional paper identified through scanning references of all full-text review articles to identify additional articles not originally identified. Table 4.2 lists all the articles and their attributes.

4.2.2 Results

Across the literature, various combinations of menus within the design space are used. The design space consists of menu layouts (i.e., how the items are organized), interaction (i.e., how to select items), and anchoring schemes (i.e., where the menu is located in space). A summary of the design space is in Fig. 4.2.

The commonly used metrics for comparison are speed, number of errors, and user preference. In some cases, comparisons also include fatigue, usability, and immersion. Often, comparisons

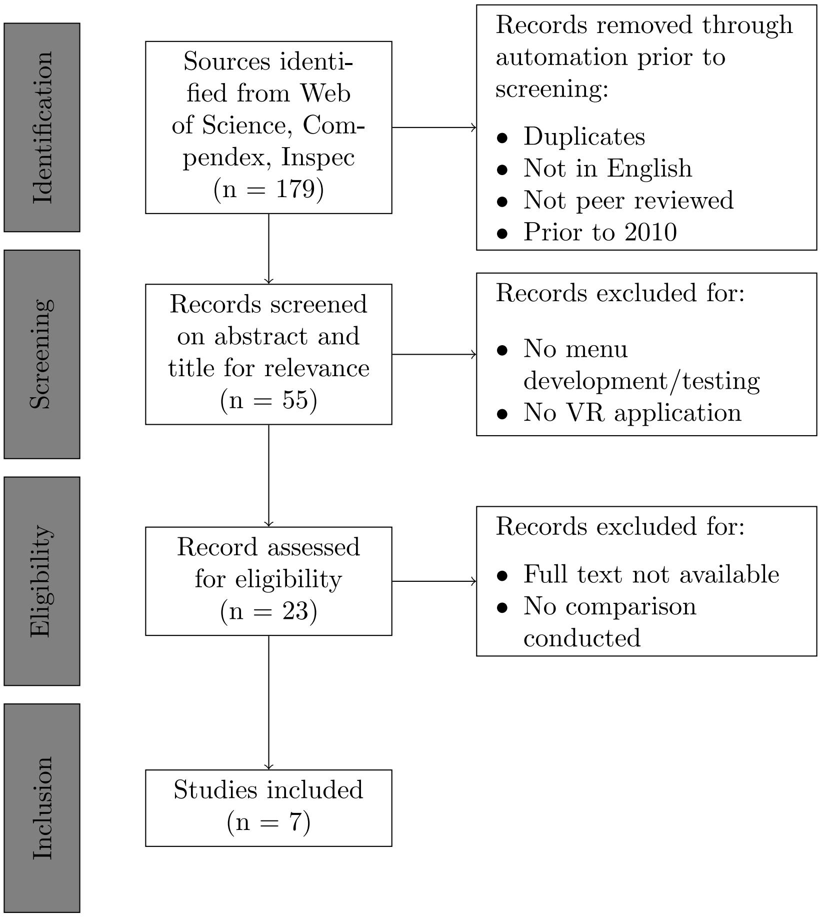

Identification
Sources identified from Web of Science, Compendex, Inspec (n = 179)
Records removed through automation prior to screening:
• Duplicates
• Not in English
• Not peer reviewed
• Prior to 2010
Screening
Records screened on abstract and title for relevance (n = 55)
Records excluded for:
• No menu development/testing
• No VR application
Eligibility
Record assessed for eligibility (n = 23)
Records excluded for:
• Full text not available
• No comparison conducted
Inclusion
Studies included (n = 7)

Figure 4.1: The modified PRISMA flowchart for the menu literature review

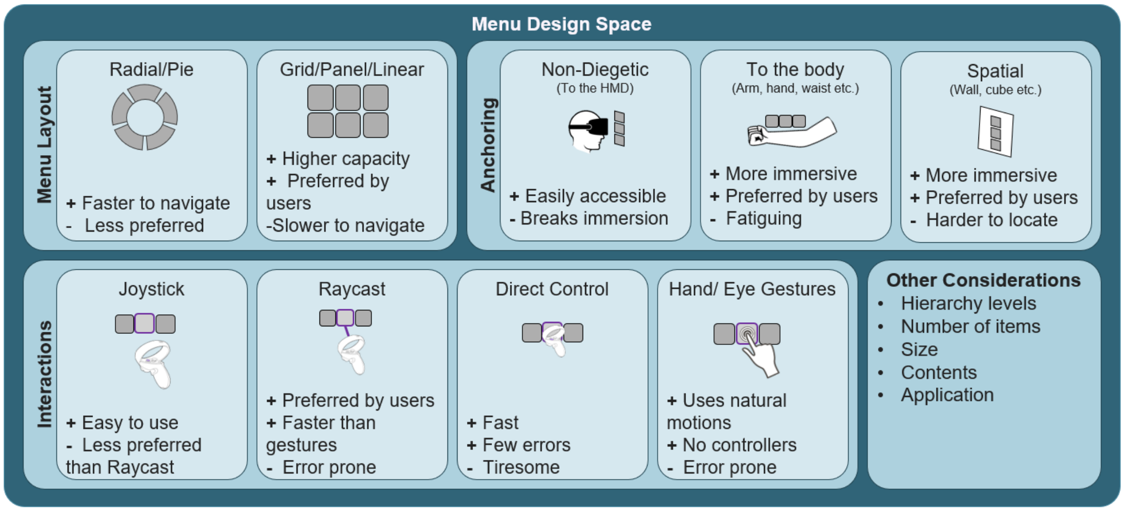


Figure 4.2: The simplified menu design space and summary of the pros and cons.

Table 4.2: Studies included in the final analysis on menus

Author (year)	Menu Layout	Interactions	Anchoring	Relevant Metrics Studied	Sample Size
Andersson & Hu (2023) [92]	Radial, Top down	Raycast, Controls	HMD	Time, Usability, Cybersickness	20
Lediaeva & LaViola (2020) [121]	Panel, Radial	Raycast, Eye, Head	Arm, Hands, Waist, Wall	Time, Error, Usability	24
Monteiro et al. (2019) [122]	Radial, Panel	Controls, Raycast	Hands, wall	Time, Error, Usability	51
Mundt et al. (2020) [123]	Radial	Raycast, Gesture, Controls	Hand	Time, Error, Usability	24
Pandey et al. (2024) [124]	Grid, Radial	Raycast, Hand	Hand	Time, Error, Usability	5
Santos et al. (2017) [125]	Radial, Linear	Raycast, Controls	Wall, HMD	Time, Error, Usability	8
Wentzel et al. (2024) [90]	Panel, Radial	Raycasting, Direct, Marking	Hand, World	Time, Error, Usability	18

include different factors within the design space, so results are inconsistent between studies.

4.2.2.1 Menu Layouts

Five of the papers included in the literature review discuss comparisons on menu layouts [92, 121, 122, 124, 125], or how the items within the menu can be organized. The menu layouts are compared on aspects of speed, accuracy, and usability. The common layouts studied and used are either in a circular form (i.e., radial, pie) or in a panel form (i.e., rectangular grid, linear vertical, or linear horizontal panel). Beyond these, other designs have also been considered that are more specialized, such as the hexa-ring [124] or TULIP [126], but these are not commonly used in real-world applications due to their novelty. These, and other specialized menus, appear often in literature to showcase their development, but the research does not contain comparisons to common menu types to understand how they might improve the design space. Thus, these specialized menus are not focused on in this analysis.

Radial menus are found to be faster to navigate than panel menus in VR [125]. The speed of radial menus has been attributed to Fitts' law, which states that the amount of time to move a pointer to a target depends on the distance to the target divided by the size of the target [127]. For radial menus, there is a smaller average distance to menu items, and hence it should be faster to navigate. However, this is not universally true, as Lediaeva and LaViola find there is no difference in time to complete based on menu shape alone, and instead find a difference with the interaction of anchoring and shape [121].

With regards to error rates, it is commonly found that there are no differences in error rates among different menu layouts [121, 125]. In one situation, radial menus increase the number of unnecessary steps required to complete a task, but only for a wall-based anchoring system. No differences are found for hand-based anchoring [122]. (for more on anchoring, see section 4.2.2.3).

Users mainly prefer panel menus over radial menus based on subjective survey results [92, 122, 124]. In one instance, no subjective differences are reported [125]. However, this may also be due to the coupling of other factors. Monteiro et al. only find a difference when considering the interaction with anchoring location: the preference is for panel menus anchored to the wall, versus radial menus on the hand [122]. No differences are found in preference between panel menus anchored to the hand, or radial menus anchored to the wall. Additionally, Andersson et al. have participants use different interaction techniques for both the panel and radial menus, potentially influencing user preferences [92].

Finally, radial menus are found to be harder to navigate when more items are included, as the selection area or angle for each item becomes smaller [124]. Panel methods can hold more items without resorting to hierarchical structures; however, the size of the menu becomes larger, obscuring visual space in VR.

4.2.2.2 Interactions and Selection

Four papers included have comparisons on interaction types [90,121–123]. Menu interactions and selection relate to how the user provides input to the menu, both in terms of navigating to the item of interest and selecting it. Interaction types include raycasting, direct controls, joysticks, and hand/gesture tracking. Raycasting, one of the most common selection techniques [90], is where a ray is projected from the hand/controller to the menu. Beyond raycasting, the controller can be used to make a selection using the joystick, trackpad, or buttons, similar to many 2D games or applications. These will collectively be referred to as joystick controls. Alternatively, direct controls involve physically moving the controller to touch the icon in the VR scene in order to make a selection. Raycasting, joystick controls, and direct controls are all controller-based modalities. Beyond controller-based modalities, selection can also be done using naturalistic movements such as hand tracking/gestures, eye tracking, or head motion. As with layouts, interactions are compared based on speed, accuracy, and usability.

Mixed results are found in the literature for which interaction types are fastest. In some studies, raycasting is found to be faster than gesture-based interactions [121] and joystick-based methods [122]. In other situations, raycasting and joysticks are found to be the same speed [123], but direct control is faster than joysticks [123]. Mundt et al. reported that some of these differences may have been due to the fact that for raycasting and direct controls, participants are able to prime themselves and put the controller in the expected position for the next task, which is not possible with the joysticks.

In general, direct control is found to be the least error prone [123]. Raycasting was found to be less error prone than joysticks [123], but more errors than head-based gestures [121]. The number of errors often depends on the target size. Raycasting, hand tracking, and eye tracking can be sensitive to movement or jitter, which may result in more errors [121]. Changes in technology may make these more intuitive and accurate in the future [121].

In situations where usability is recorded, raycasting is consistently found to be preferred among users [90, 123].

Other considerations for interactions include accessibility, as some interaction types may require more arm movement (i.e., hand tracking or direct control). This can result in users having to prop their arms for stability or lead to fatigue, as reported by some participants using these methods [123].

4.2.2.3 Anchoring

The menu placement, or anchoring, influences the user's ability to access it, how integrated it is to the environment, and the amount of occlusion of the surrounding visualization. Anchoring comparisons are conducted in three of the papers [121, 122, 125]. Menus can be non-diegetic, where they are not integrated into the environment, such as those attached to the HMD, and are always in the same spot and accessible. They can also be contextual, either attached to a body part (i.e., hand, arm, waist) or spatial (i.e., attached to a wall). Contextual menus promote increased immersion with the environment [122]. The anchoring location is primarily compared in terms of usability and preference.

No differences in user preference are found between non-diegetic and spatial menus (attached to a wall) [125]. However, among different contextual menus, wall menu has higher reported usability than a hand-based menu [122] and arm-based menu [121].

It is found that arm menus require moving the head down, which some people find uncomfortable [121]. Spatial menus are found to require less movement than arm and hand-based menus [121].

4.2.2.4 Other Considerations

In addition to the above design space, there are other considerations when selecting a menu. This includes the number of items in the menu, hierarchy levels, and menu size. These considerations may interact with the other display considerations.

The more items in the menu, the more complicated it may be to navigate and interact. At a single level, panel menus can hold more items while maintaining a reasonably sized area to select. However, this also makes the overall menu bigger and may block more of the visual scene [90]. Instead, to compensate for the increasing number of items, hierarchy can be used, where there are nested layouts.

The menu size, or how large a space the menu takes up, is an important consideration for immersion and occluding the display. The bigger the menu, the more likely it is to interfere with the visual scene, reducing immersion and the ability to process and understand the scenario, which is important for operations [90].

4.2.3 Discussion

The optimal menu depends on the exact application and needs, as there are both pros and cons to every method. In general, across the different studies, all the menu types scored well in all metrics, indicating that they can be useful and still maintain good performance, even without clarity on which type may be optimal for a given application [122, 123].

Different considerations may be needed for the application when considering the design of the menu. Depending on the situation, different weights may be placed on speed, accuracy, immersion, or user preference, resulting in different optimal menu designs. For example, in situations where fatigue, speed and accuracy are paramount over immersion, the design may be driven towards considerations like radial menus, controller inputs, or raycasting, and anchoring to the head. Factors like frustration and usability are important for people to use the system, but preference between two highly usable systems may not be as critical.

Finally, the duration of the operation must be considered. If operations long in duration, ensuring that the system is not tiresome to use is important, especially if there needs to be constant interactions with the menus. This may make the non-diegetic, attached to the HMD, menus desirable as they can be recalled as needed and are not fatiguing. It also may make the interaction technique decision important to limit arm fatigue, such as using controller inputs.

Some papers compare components of a menu design without holding the other confounds constant, such as using different interaction techniques for each menu [122]. Due to the interactions seen within the design space, this may influence the findings.

4.3 Text

The second area of interest is the use of text and how to display it. This focuses more on the location and context of text, not on text size, as guidelines already exist for this [94]. In operational use, text cannot always be completely replaced by visualization or other graphics. This review is interested in understanding the best way to display text that is found to be critical for a given application.

4.3.1 Methods

The initial search results in 470 articles. The PRISMA-like procedure is documented in Fig. 4.3 and Tab. 4.3. During the title and abstract screening phase, results are excluded for no VR application and no mention of the display of text. Only one reviewer screened the titles and abstracts of all papers. Following the full text review, 6 papers are included.

4.3.2 Results

A total of 6 articles satisfied the criteria and are included. Table 4.4 lists all the articles.

Table 4.3: Search terms for text review. Note * indicates terms with multiple variations

Topic	Search Term	Location
Technology	VR OR “Virtual Reality”	Title/Abstract/Keywords
Text	Text OR HUD OR “‘Heads up display” OR diegetic	Title/Abstract/Keywords
Exclusion	NOT (AR OR Augmented Reality OR Automotive)	Title/Abstract/Keywords

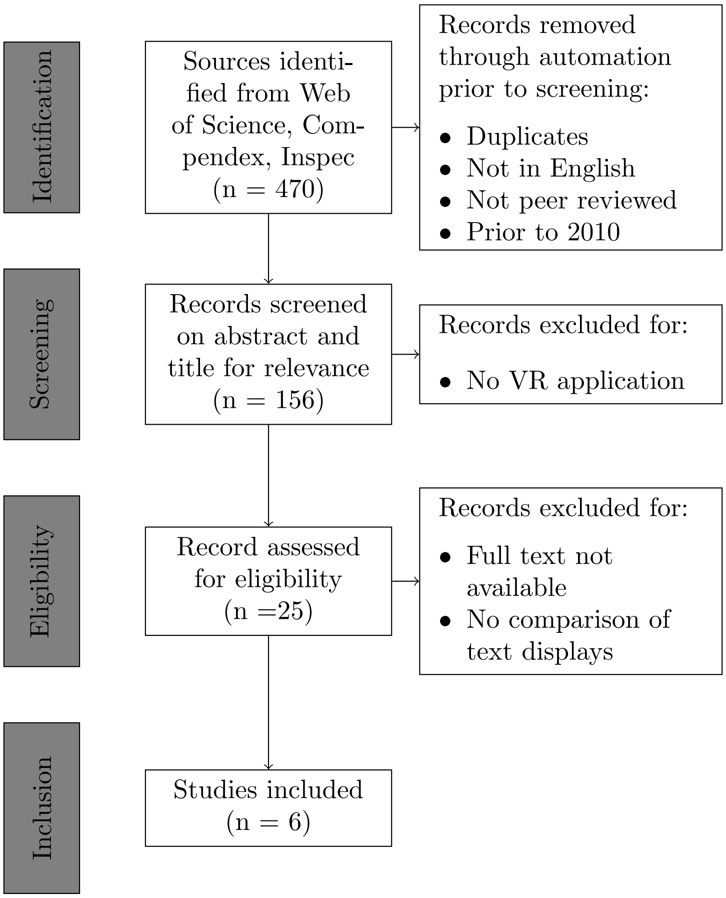

Identification
Sources identified from Web of Science, Compendex, Inspec (n = 470)
Records removed through automation prior to screening:
• Duplicates
• Not in English
• Not peer reviewed
• Prior to 2010
Screening
Records screened on abstract and title for relevance (n = 156)
Records excluded for:
• No VR application
Eligibility
Record assessed for eligibility (n =25)
Records excluded for:
• Full text not available
• No comparison of text displays
Inclusion
Studies included (n = 6)

Figure 4.3: The modified PRISMA flowchart for the text literature review

Table 4.4: Studies included in the final analysis on text

Author (year)	Interfaces	Application	Relevant Metrics Studied	Sample Size
Dickinson (2021) [128]	Spatial menu, Diegetic (toolbox)	CSI training	Presence, Workload	58
Koehle et al. (2021) [129]	Non-diegetic, Watch, Physical	Video Game Health Status	Presence, Utility	37
Marre et al. (2021) [130]	Diegetic, Non-diegetic	First person shooter video game	Performance, Presence, Enjoyment	41
Nava et al. (2024) [131]	2D interface, Diegetic (Billboard)	Construction training	Performance, Immersion, Usability	14
Queck et al. (2023) [132]	HUD, Watch, In situ	Sports physiological data	Performance, Presence	29
Salomoni et al. (2017) [133]	Watch, World text	Video game	Presence, Simulator sickness	10

Most of the articles compare text-based interfaces in a video game context, including both diegetic and non-diegetic interfaces. Diegetic interface describes a situation where the controls or information appear as a part of the simulated environment, rather than on a separate menu system or screen overlay. Examples of diegetic interfaces include putting information on a watch or a clipboard. A common type of non-diegetic interface is a heads-up display (HUD), which provides information as an overlay but may hide elements in the main scene. These interfaces are commonly compared on measures of immersion, presence, obtrusiveness, usability, accuracy, and reaction times.

The main benefit of diegetic interfaces is the idea that they can improve immersion [129], which may help promote many of the benefits of VR displays. The improvement to immersion is likely due to the fact that they are unobtrusive [129]. HUDs, on the other hand, are intrusive [129] and can cause occlusion problems, hiding the scene behind them [133]. While they improve immersion, presence is not always improved by diegetic interfaces, counter to popular belief. Comparisons found equal presence between diegetic and HUD [128, 129, 131], potentially due to an increase in workload [128]. One study finds improvements to presence with diegetic interfaces, but notes the nature of the interface is critical [130]. Additionally, users often prefer diegetic interfaces [132, 133], but on occasion, non-diegetic interfaces improve usability [131].

HUD promotes higher accuracy and reaction times. Across studies, HUDs are found to be more accurate [129] and decrease the number of missed notifications [132]. HUDs are also found to decrease reaction time to an alert [132], are quickly available [129], and are more efficient [129]. While diegetic interfaces are also been found to be accurate [129], it is noted that they often require users to actively check and that they may be impractical to check when there are many actions required [129, 133].

Beyond this, there are a few other considerations to note. Diegetic interfaces are found to decrease cybersickness [133]. Cybersickness includes symptoms such as nausea, eye strain, disorientation, and fatigue, and is not ideal for operational uses.

Finally, the location of the HUD or how the diegetic information is integrated plays an essential role. For diegetic interfaces, it is suggested that in situ placement may be a way to overcome occlusion from HUD or elaborate movements (such as required by a watch) [133]. This in situ placement should remain close to the action and still remain accessible to improve performance [130]. For HUD, text displayed in the center and bottom, as opposed to top right leads to less workload. The top and peripheral are deemed more comfortable and unobtrusive, while the middle and bottom are more noticeable [134].

4.3.3 Discussion

Text is often critical in many operational environments to give guidance on exact values (i.e., speed, distance, times), abstract information, or warnings. Some types of operational displays, such as those for spacecraft monitoring [2] or power plant operations, are often predominantly text-based. Other operational displays, such as those inside a car, have more limited text, and instead, users rely more on information gained from the surrounding view [135]. In many of these operations, though, there is critical text or graphics that are required to be easily accessible.

Many of the benefits of VR in operational use are believed to come from the immersive environment. Having too much text, or a HUD, is counter to an immersive environment, and instead, this may be promoted by using visuals or diegetic textual interfaces. It is suggested, if possible, that the text should be placed near the corresponding item and near the action scene, so it is accessible. This is consistent with the principles of spatial contiguity [136] and the proximity compatibility principle [137].

However, diegetic interfaces are not accurate or quick to retrieve information from, which can lead to missed notifications and mistakes due to not seeing the necessary information. It is suggested that HUD, despite its limitations, may still play an important role in operational tasks. Similar to a fighter pilot or aircraft HUD/HMD [138, 139], in VR HUD should contain the text or graphics that are important for people to know at all times and be easily accessible [130]. Effective use of a HUD means considering the appropriate information to include; there are concerns with too much text/excessive information increasing occlusion, and being harder to process, overwhelming, and hard to read [140].

A potential appropriate use of text for operations may include putting mission-critical text and alerts in a HUD, similar to fighter pilots. Other text, that may not always need to be accessible, can be placed diegetically near the item it is referenced to help promote immersion. Future research should be conducted to understand the maximum amount of text in a HUD before too much immersion is lost, and optimal HUD placement. Additionally, research should be done to compare different operationally relevant diegetic text interfaces to understand which applications are more beneficial to each.

4.4 Discussion

This aim focuses on menus/interactions and text since the initial literature review identified these as requiring unique considerations for VR displays. Beyond these, there are still many other components of a design that need to be taken into account for an operational display. However, VR guidelines already exist for many of these, such as the size and font of text, controller mapping, and accessibility [94,141]. In addition, other traditional human factor principles are still often applicable, such as minimizing information access cost, and the proximity compatibility principle [142, 143]. Beyond this, many operational paradigms or organizations have their own internal guidelines that include information about alerts, colors, graphics, and auditory (i.e., FAA regulations for aviation), which can often be extended to VR. The main difference is ensuring that they still achieve the same purpose in VR, so items like how to interact with them, or their location and sizing, may need to change. These recommendations and the developed design space can be applied towards a variety of contexts, as long as the selected design aligns with the mission goals. By optimizing the displays for use in VR, through appropriate design choices, the benefits of VR may be realized. This may reduce the risk of VR performing badly due to inappropriate display designs [77].

There are some limitations and confounds to this research. The majority of papers do not focus on operational applications and outcomes. While this review considered the implications for operations (such as reducing error rates or increasing speed), these results are not guaranteed to transfer to a new environment. Any decision should still be tested within the environmental context it will be used in to ensure it meets the needs and requirements of the operations. As more organizations are interested in VR for operations and training, future research should consider focusing on these applications.

Additionally, as VR technology changes, these recommendations might change. The date range of the articles is set to post-2010, but even in the past 15 years, technology has rapidly improved. Technology is most likely to affect the interaction and selection techniques that rely more on hardware and software to accurately track controllers, hands, or eyes. As these improve further, it may decrease errors and selection times, changing the ideal interaction technique. Other considerations, like fatigue, may not be influenced by technology. The display of text is additionally heavily impacted by resolution. A limitation of VR is that large blocks of text may be challenging to read due to low resolution. Although the literature currently suggests that large blocks of text are also not beneficial for immersion or occlusion reasons, as resolution improves, there is a chance that text sizes and line widths can decrease while maintaining the same readability, enabling more text to fit in a HUD without causing problems.

4.5 Summary and Contributions

In this aim, a systematized literature review is conducted to understand the design space and implications for VR menus and text displays in an operational context. In total 13 papers are identified, 7 papers for menus and 6 for text are analyzed to understand the impact of design choices on operational use. While most of the prior work is on video games or training, this aim looks at their findings and considers the implications in operational settings. However, there is a continued need for research in display designs with a focus on operational applications to understand how the decisions impact overall findings. The overall contribution to the literature is the review of the design space for VR operations, with a focus on menus and text.

Chapter 5: Aim 2: Remote Monitoring Operations

5.1 Introduction

The objective and contribution of this aim is to compare the effects of 3D visualization and VR on a remote operator's understanding of uncertainties using a specific application: monitoring of a satellite during operations. Monitoring tasks are chosen, as it is how the operator spends the majority of their time during remote supervision, and has critical implications for being able to take appropriate actions. This aim considers three displays with varying degrees of visualization and immersion, and their effects on SA, workload, usability, and subjective understanding of uncertainties. We hypothesize that 3D visualizations will improve SA, lower workload, and improve usability and subjective utility over displays without 3D visualizations. We further hypothesize that immersive displays, such as VR, will provide additional benefits over 3D visualizations. This research has been published in Frontiers in Virtual Reality [144].

5.2 Methods

In this research, three displays of increasing levels of visualization and immersion are designed and implemented to simulate the remote monitoring of spacecraft operations. The simulated remote monitoring task is a rendezvous mission scenario in which a servicer approaches a target vehicle, performs corrective burns, and changes its orientation to inspect the target. The three display designs are compared through a human subject evaluation.

5.2.1 Scenario Design

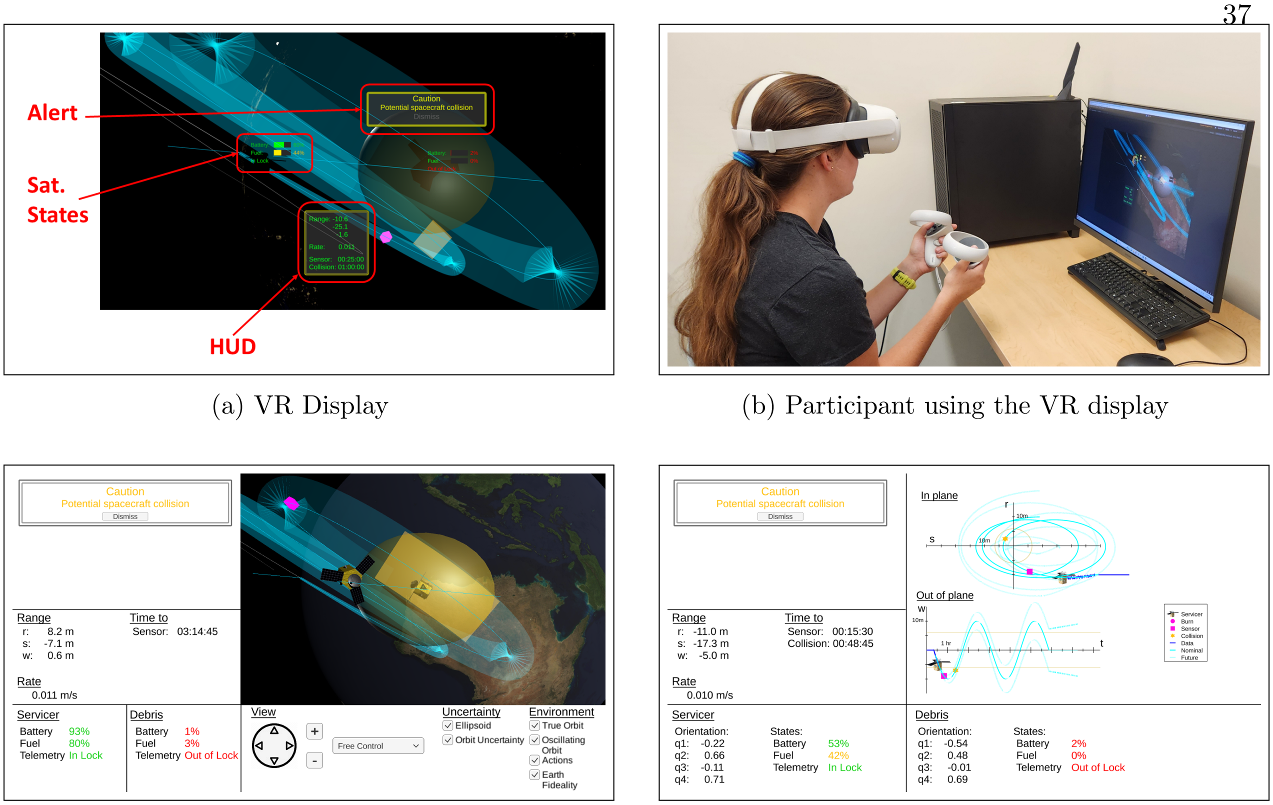
Participants are tasked with monitoring the proximity operations portion of a satellite rendezvous mission. During the task, the participant cannot intervene or provide commands to the satellites. The underlying trajectories used for the simulation are developed using Basilisk, a high-fidelity, flight-proven, physics-based satellite simulation tool [145]. The scenario consists of two satellites in orbit around Earth: a non-operational, tumbling, debris satellite and an active servicer satellite, supervised by a remote operator, sent to inspect the debris satellite. The debris satellite has no communications, fuel, or battery, thus, there is uncertainty in its location.

The scenario is broken into three phases. Pre-burn, the servicer satellite is approaching the debris satellite on a parallel orbit. There are checks to ensure that the thruster plume from the burn will not impinge on the debris satellite. The servicer burns and enters an orbit to inspect the satellite. This orbit is no longer parallel to the debris satellite, and instead, the servicer satellite spirals about the debris satellite with some out-of-plane motion. In the post-burn, pre-sensor update phase, there is uncertainty in the servicer satellite's current location and future location due to uncertainty in the magnitude and direction of the delta-v imparted by the thruster burn. The combination of this and the debris satellite state uncertainty leads to a potential for collision. As the scenario continues after the thruster burn, the knowledge of the servicer satellite of its position relative to the debris satellite improves, simulating the gathering of data from sensor updates. The gathered data results in a reduction of both uncertainties, which leads to a change in collision risk. Although in a real rendezvous scenario, the satellite operations would continue, in this trial, post sensor update, the participants' scenario is terminated after a randomly assigned length of time. The duration of the rendezvous simulation as experienced by a participant is compressed, with 15 seconds of simulation time displaying per 1 second of the participant's real-world time.

5.2.2 Display Design

Three different displays were designed for this experiment to investigate the impact of 3D visualization of data and immersion of display, as seen in Fig. 5.1. The VR display was designed first and then modified to make the other two displays. All displays were designed with a consistent focus on using relevant display design principles to ensure readability (i.e., legibility, contrast, minimizing information access cost) and interpretability (i.e., avoiding absolute judgment limits [143]) so that the results are not skewed due to fundamental differences in how they were developed.

The VR design philosophy was based on a combination of Heads Up Displays (HUD) [146–148], traditional aerospace displays guidelines (MIL-SPEC and FAA regulations) [147,148], and best



(a) VR Display

(b) Participant using the VR display

Caution
Potential spacecraft collision
Dismiss

Range
r: 8.2 m
s: -7.1 m
w: 0.6 m

Time to
Sensor: 03:14:45

Rate
0.011 m/s

Servicer
Battery 93%
Fuel 80%
Telemetry In Lock

Debris
Battery 1%
Fuel 3%
Telemetry Out of Lock

Uncertainty
☑ Ellipsoid
☑ Orbit Uncertainty

Environment
☑ True Orbit
☑ Oscillating Orbit
☑ Actions
☑ Earth Fideality

(c) Screen Visualization Display

Caution
Potential spacecraft collision
Dismiss

Range
r: -11.0 m
s: -17.3 m
w: -5.0 m

Time to
Sensor: 00:15:30
Collision: 00:48:45

Rate
0.010 m/s

Servicer
Orientation:
q1: -0.22
q2: 0.66
q3: -0.11
q4: 0.71
States:
Battery 53%
Fuel 42%
Telemetry In Lock

Debris
Orientation:
q1: -0.54
q2: 0.48
q3: -0.01
q4: 0.69
States:
Battery 2%
Fuel 0%
Telemetry Out of Lock

(d) Baseline Display

Figure 5.1: The three different display designs and an example participant (person shown is part of the research team) using the VR display. The VR display is annotated in red to show the location of the HUD, the satellite states text, and the caution and warning alerts.

practices for VR [54,149] and visualizations [142,143]. As seen in Fig. 5.1a the VR consists of several parts. The underlying immersive visualization was built as an extension of Vizard, a spacecraft simulation visualization software application that provides the satellite models, relative orbit lines, location relative to Earth, and appropriate Earth-Sun lighting [150]. Overlaying the relative orbit lines and satellites are transparent display objects designed to illustrate the uncertainties and locations of upcoming actions, including burns and sensor updates. The uncertainty of the servicer satellite's future position is represented by a blue tapered extrusion along the curve of the projected orbit with increasing diameter representing increasing uncertainty. The ellipsoid surrounding the debris satellite denotes the uncertainty of its position and can be used to monitor the likelihood of a collision. Any overlap of these two uncertainty visualizations indicates a potential collision and is highlighted in the same color as the ellipsoid. The ellipsoid color is changed to indicate the level of concern to the participant. Yellow represents a caution, where there is a chance of collision but also the participant will still receive more information through a sensor update. Red represents a warning, where there is a chance of collision, but no chance of receiving new information from a sensor update. These colors are based on the standard alert colors for aircraft displays [147, 148].

The participant can change their viewpoint as desired through panning, zooming, and teleportation using the left VR hand controller. In addition, participants could change their view through natural movements like turning their head or body. The ability to change viewpoints or switch between preset viewpoints has been shown to improve performance and increase operator understanding of the environment, as different perspectives may be beneficial for different aspects of a task [54, 149].

In addition to visualization, critical information is displayed in text-based form (annotated in Fig. 5.1a), as a result of the findings from Aim 1. A HUD panel displays information critical to the mission, including the satellite's relative ranges, rates, and time until the next action. The HUD is always in the same location relative to the VR headset, ensuring the HUD remains visible even if the participant turns their head. The HUD is located in the periphery of the participant's vision, allowing them to access the information through eye movements, minimizing the information access effort [151] and blockage of the visualization. Location and text size were based on VR recommendations [94], [93], and adjusted based on user evaluations for readability and accessibility during pilot testing. The satellite states, which are critical to a specific element of the mission, like battery, fuel, and telemetry status, are presented in text-based form and move along with their associated satellite. This difference in text display choices is attributed to the desire to minimize the amount of text in the HUD. While HUDs are important to display critical information that an participant should always be aware of, they also reduce the immersion experiences in a VR display [146, 152]. Attaching text to elements in the scenario minimizes disruptions to immersion, and thus is a way to display text that may be important, but not required to always be visible [130, 153].

Finally, cautions and warnings are displayed through an alert at the top of the screen. These alerts are triggered automatically based on certain events in the scenario, such as a warning about collision potential. The cautions and warnings are highlighted and presented in a salient location; designed to be easily noticed by the participant. Cautions and warnings and are color-coded yellow and red, respectively. Like traditional spacecraft and aviation displays, the participant can dismiss these alerts [147, 148]. The dismissal of alerts is completed with the VR controller and allows the participant to regain areas of their visual field. Based on the findings from Aim 1, the participants can interact with the display using a radial menu system to toggle on and off different aspects of the visualizations and displayed information, which was controlled through the right VR hand controller. This allows the participant to customize their view in a way that allows them to hide information that is not currently relevant and to more clearly understand the information that is relevant. During development, the VR display underwent evaluations where student volunteers were asked to perform a series of tasks. Then they commented on and used a 6-point Likert scale to evaluate the display's readability, controllability, and interpretability. This resulted in multiple iterations, until all scores were positive, to increasingly improve the text readability, display location, intuitiveness of the controls, and ensured that aspects of the visualization were interpretable.

The screen visualization display maintains aspects of visualization but does not have the immersion (i.e., presented on a 2D screen vs. in VR) that VR allows for and is seen in Fig. 5.1c. It uses the same underlying visualization as the VR display, where the participant can still pan, zoom, and interact with the visualization components in a 3-dimensional manner, however, it is now on a 2D computer screen and thus not immersive. The HUD information, satellite states, and alerts that were previously in text form in VR are now displayed outside the visualization on the screen, creating a consistent scan pattern for participants, and grouped with similar constructs.

Finally, the "baseline" display (Fig. 5.1d) contains no visualization or immersion. All telemetry and system states are presented on a 2D display in graphical and textual form without 3-dimensional modeling or display components. This display is representative of traditional satellite monitoring displays used in current operations that primarily contain text-based information, but also includes graphs of telemetry, as consistent with current operations. It has the same text as the screen visualization layout, however, instead of the 3D visualization view, 2D graphs of the relative in-plane and out-of-plane orbits between the satellites are presented. Participants are unable to customize their viewpoint or interact with the baseline display.

The same information is available to the participant in all three displays, though the presentation of the information differs. Customization of the display by the participant is supported in both the VR and screen visualization displays, however, there is no ability to control the satellites in any display. This remote supervision task addresses an important aspect of remote supervision by requiring the participants to continuously monitor the satellites without the ability to intervene.

5.2.3 Experimental Protocol

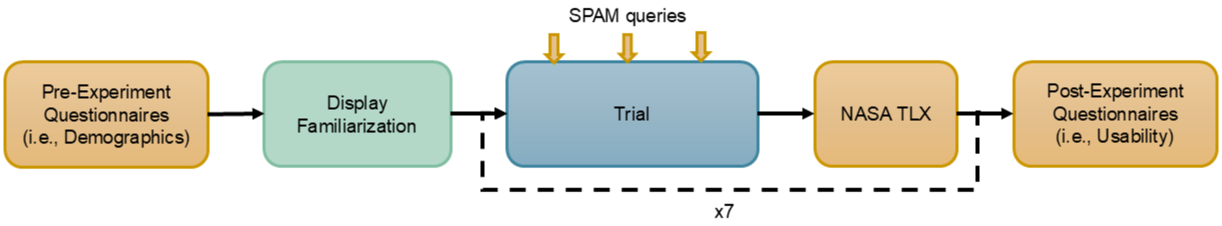

SPAM queries
Pre-Experiment Questionnaires (i.e., Demographics)
Display Familiarization
Trial
NASA TLX
Post-Experiment Questionnaires (i.e., Usability)
x7

Figure 5.2: Experimental design flowchart. The orange boxes indicate data collection through surveys or queries.

The study was approved by the University of Colorado at Boulder Institutional Review Board (Protocol #23-0100). Informed consent was obtained from all participants. Thirty five participants from around the University of Colorado Boulder campus were enrolled. Two participants did not demonstrate an understanding of the task and did not finish data collection. Thus, 33 participants completed the experiment (15 Female, 18 Male; ages 18-57, median age 25 years). All participants were aware of the high-level project goals from the informed consent, but naive to the alternative display conditions or exact manipulations of the scenario. Participants were screened for vision correctable to 20/20, no colorblindness, and a score of less than 90% on the Motion Sickness Susceptibility Questionnaire [154] as a means to identify individuals who would be highly susceptible to simulator sickness prior to data collection.

Participants were randomly assigned to one of the three display conditions: VR, screen visualization (Scr. Viz), and Baseline (11 participants per condition; 5 F, 6 M). Those in the VR condition wore a Meta Quest headset, while those in the other conditions used a computer and 2D monitor. Participants in VR had the option to sit in a spinning chair or stand and walk around. All opted for the chair, but often used head movements and body rotations, in addition to panning, to change their view. Participants in both screen conditions sat in a chair in front of the computer.

A flowchart of the experiment design can be seen in Fig. 5.2. All participants completed a demographic questionnaire on their background, including familiarity with orbital mechanics, familiarity with spacecraft operations, and prior VR experience. Participants were then trained using a PowerPoint presentation. The presentation covered any background orbital and operational knowledge needed to be able to complete the experiment. It also provided context for the scenario they would experience, and values specific to the satellites they would be monitoring (e.g., amount of fuel needed to complete a burn). They were also trained on the specifics of the display modality they were assigned to. After the training PowerPoint, participants were quizzed to ensure an understanding of the scenario and tasks they would perform. They then completed two training trials. The first trial provided an opportunity to become comfortable with the system controls and the location of items within the display. For this trial, there was no monitoring objective. When participants felt comfortable with the display, they were asked a series of questions to ensure they could find critical information and understand the visualization. The second training trial followed the format of a real trial. Participants had to achieve accuracy on the tasks and had to feel comfortable before moving on. These training trials were done to minimize the effect of participants being unfamiliar with the controls or task, rather than due to the display itself.

After training, participants completed seven trials of the experiment. The order in which the participants saw the trials in was randomized; however, all participants experienced the same seven trials. Each trial used the same underlying orbits, but the uncertainty surrounding the debris, the uncertainty resulting from the servicer's burn, and the location of the sensor update varied which varied the likelihood of collision of the two satellites. Additionally, the servicer satellite's initial fuel and battery value varied. The length of each trial varied, but all trials were approximately 8 minutes long.

SA was measured throughout the course of the trial through two different mechanisms. Level 1 SA, or perception, was measured through SA callouts [155, 156]: Participants were instructed to report the servicer's battery and fuel values in 10% increments (e.g., 90%, 80%) and the time to any action (burn, sensor, collision) in 15 min increments (e.g., 15 min to burn). These values did not always change linearly; for example, the battery value would increase or decrease based on the orbital position of the satellite relative to the sun, and the fuel would change based on burns. Callouts made within 2 seconds of the actual event occurring were judged successful, while late callouts are considered missed. An experimenter marked callouts as they occurred; the callouts were then verified post-experiment from audio recordings. The number of total possible callouts varied per trial (between 15 to 21); however, the total percent correct of callouts made over all the trials was used in the data analysis to normalize the values across trials.

To understand level 2 and level 3 SA, the Situation Presence Assessment Method (SPAM) was used [157]. This is a real time SA assessment method that is meant to mimic a control room and has been used often in other operational setting experiments, like air traffic control [158–163]. At three points throughout the trial, a beep was played. The location of the query was randomly selected within each of the 3 phases of the scenario. Participants were instructed to say 'ready' when they felt that they had a low enough workload to be asked queries. At this point, an experimenter would proceed to ask the participant two queries, one for SA level 2 and one for SA level 3. This mimics a second operator in a control room asking for information. The queries were randomly chosen from a list of potential queries. The list was generated through a process similar to goal directed task analysis [28]. Example questions include: "Is the servicer satellite currently in the sun?" (level 2), or "Will there be enough fuel to complete a burn at the time of the next burn?" (level 3). If a SA callout event occurred during a SPAM assessment, participants were instructed to not announce the SA callouts, and this was not counted against them when scoring level 1 SA. For each SA level, the total percent of SPAM queries answered correctly was used in the data analysis.

After each trial, participants assessed their workload through the NASA Task Load Index (TLX) [35]. They rated seven dimensions of their workload on a 21-point scale. This includes mental, physical, temporal, performance, effort, and frustration. At the end of all seven trials, participants then completed the comparisons between subscales. This allowed a weighted TLX workload score to be calculated, which included the subscale rating and relative importance of that subscale, resulting in a workload score between 0 and 100. Additionally, after the end of each trial participants verbally rated their nausea on a scale of none, slight, moderate, and severe to help assess cybersickness. No participant reported symptoms of nausea.

After all trials were complete, participants also completed the System Usability Scale (SUS) [164], which is a 10 question survey in which participants respond on a 5-point scale. These are combined to give a resulting score from 0 to 100. In addition, participants answered questions relating to their perceived understanding of the servicer uncertainty, debris uncertainty, collision likelihood, ease of finding information, and awareness of critical events. The full text of this survey can be found in the supplementary materials.

5.2.4 Statistical Analysis

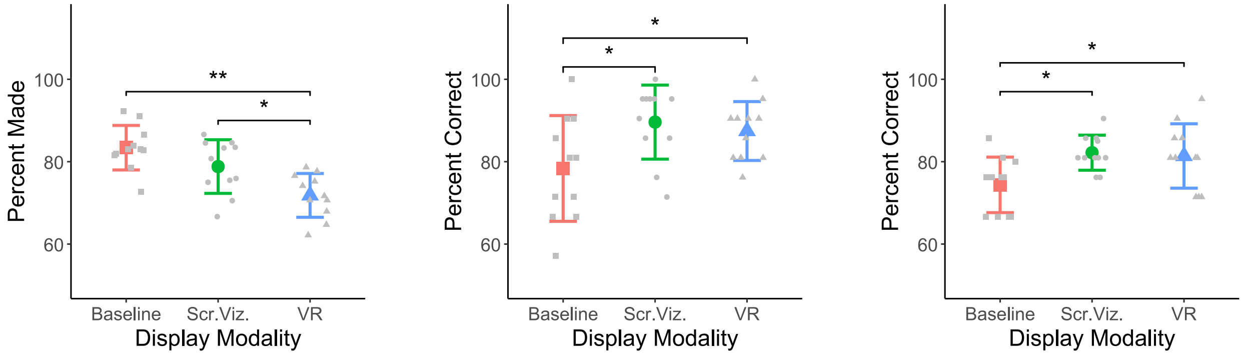
The three displays were compared across the 3 SA levels, workload, usability, and subjective utility. For SA level 1 (perception) the participant averaged percent of correct callouts made was used. For SA levels 2 (comprehension) and 3 (projection), the participant averaged percent correct of SPAM queries for that SA level was used. Unlike traditional SPAM analysis where the response time is used as a measure and the percent correct is treated as the same across conditions [165], participant averaged percent correct was used as different conditions had different accuracies. For workload, the weighted TLX score was used, and for usability, the System Usability Scale score was used. For the utility questionnaire, each question was analyzed independently.

The study collected 231 trials over 33 participants. One trial for two separate participant s were removed as these participant s experienced technical difficulties during those specific trials. All other trials for those participant s were retained since the technical difficulties did not affect the other trials. For all 3 SA levels, usability, and utility there were 33 total data points as each participant had a single averaged measure. For workload, as trials were kept separated, 229 data points were used.

Prior to statistical analysis, SA and workload were inspected for potential confounding factors of trial order to capture undesirable learning effects, and for the scenario parameters experienced, as each participant experienced the scenarios in a different order. No effect of learning or trial experienced was identified, based on the slope of the particular metric over trial order on a per participant basis. In addition, the data was visually evaluated for potential confounds based on the participant's background, including orbital experience, satellite operations experience, gender, and VR familiarity. The participant's orbital mechanics experience was relevant for all three levels of the SA data, but no participant background was relevant for workload, usability, or utility. For all statistical tests the assumptions were met, unless otherwise noted. A criterion of $\alpha = 0.05$, after appropriate correction factors, was used for significance for all tests.

For all 3 SA levels, a linear mixed-effects model was used. The display modality was treated as a fixed effect, and the orbital mechanics experience (coded as none to low, or moderate to high) was treated as a random effect. The model was fit using the *lme4* package in R via penalized maximum likelihood estimation [166]. After fitting the model, the residuals were checked to ensure that they obeyed normality and independence. The significance of display modality was assessed using an F test with a type III ANOVA with a Satterthwaite approximation for degrees of freedom and was implemented using *lmerTest* package in R [167]. Post-hoc tests were done between all pairwise comparisons using estimated marginal means (*emmeans* package in R [168]) with a Tukey p-value correction and Kenward-Roger degrees of freedom correction. The effect size was calculated using the *effectsize* package in R [169].

For workload and usability, no participant background was relevant. For workload, each trial was included as a separate data point and the participant was treated as a random effect nested within display modality. The same analysis pipeline was followed as for SA. For usability, no random effects were included so a linear model was fit between the system usability score and



(a) (b) (c)

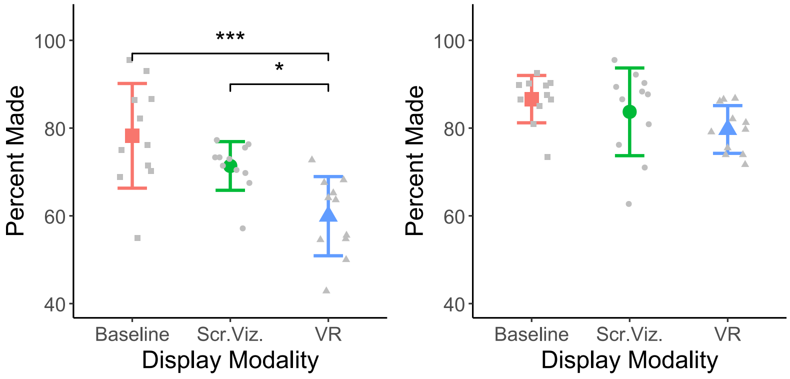
Figure 5.3: The level 1 (a), 2 (b), and 3 (c) SA results. Level 1 SA shows the participant average percent of callouts successfully made over the condition. Level 2 and 3 SA plots the participant averaged percent of SPAM queries of that level answered correctly. All figures show the data mean, standard deviation error bars, and significance is noted between the conditions.

display modality. The residuals were then assessed for normality and independence. A type III ANOVA was used to compare the display modalities.

The subjective utility questions were each on a 5-point Likert scale. Thus, ANOVAs could not be used, and instead, each question was analyzed using a Kruskal-Wallis H-test.

5.3 Results

Significant differences are seen between the display modalities for the 3 SA measures shown in Fig. 5.3. There is a significant difference in the level 1 SA as measured by the participant average percent callouts made. The ANOVA comparing the linear mixed effect models found significance between conditions (F(2, 29.06) = 11.62, $p < 0.005$, $\eta^2 = 0.44$). Follow up pairwise comparisons with Tukey adjusted p-values found that the differences are between the VR and baseline display (t(29.1) = 4.76, p = 0.001, d= 2.04), and VR and screen visualization (t(29.1) = 2.95, p = 0.017, d=1.26), but no differences between the baseline and screen visualization (t(29.1) = 2.95, p = 0.17, d=0.78). Further analysis into level 1 SA compares the differences in display modalities across the different types of callouts, as seen in Fig. 5.4. There are differences in modalities across the percent of satellite state callouts made. This includes the fuel and battery values, and is unique in the VR display as these values follow the satellite's position and change location with time (F(2, 29.06) =

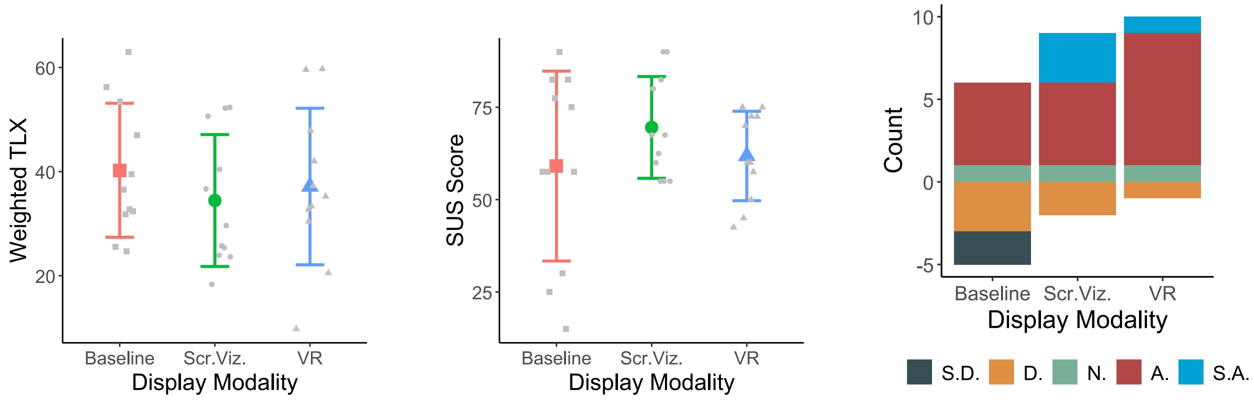


(a) (b)

Figure 5.4: The level 1 SA sub analysis: satellite states (left), time to the next events (right). Both show the participant average percent of callouts successfully made over the condition within a category. The satellite state locations move in VR, the time to next event has a static location. All figures show the data mean, standard deviation error bars, and significance is noted between the conditions.

11.24, $p < 0.005$, $\eta^2 = 0.44$). Post-hoc pairwise comparisons with Tukey adjusted p-values found that the differences are between the VR and baseline display ($t(29.1) = 4.66$, $p < 0.005$, d= 2.00), and VR and screen visualization ($t(29.1) = 2.97$, $p = 0.016$, d=1.27). However, there is not a significant difference in the percent of the callouts made correctly regarding the time until the next event ($F(2,29.17) = 2.50$, $p = 0.10$, $\eta^2 = 0.15$). These callouts are stationary in all 3 displays, as this information is in the HUD component of the VR display.

For both level 2 and 3 SA there is a significant difference between display modalities as measured by the participant averaged percent correct of the SPAM queries asked. For level 2 the ANOVA found a difference ($F(2, 29.02) = 5.57$, $p = 0.0089$, $\eta^2 = 0.28$). Post-hoc pairwise comparisons with Tukey adjusted p-values found that the differences are between the baseline display and screen visualization ($t(29) = -3.13$, $p = 0.011$, d=-1.34) and baseline display and VR display ($t(29) = -2.56$, $p = 0.041$, d=-1.10). Level 3 SA found similar results, with the ANOVA finding differences in display modalities ($F(2, 30) = 4.90$, $p = 0.014$, $\eta^2 = 0.25$), and post-hoc comparisons finding the differences between the baseline display and screen visualization (t(29.3) = -2.81, $p = 0.023$, d=-1.21) and baseline display and VR display ($t(29.3) = -2.52$, $p = 0.044$, d=-1.08). The means and standard deviation for the different display modalities across the different



(a) (b) (c)

Figure 5.5: The Workload (left), usability (center), and utility (right) results. Workload shows the participant average weighted TLX score, and usability shows the System Usability Scale score. Each modality's data means and standard deviation error bars are overlayed. The utility plot shows the results of the question that was closest to being statistically significant: "I found this system enabled me to understand the uncertainty associated with the servicer". (S.D. = Strongly Disagree, D. = Disagree, N. = Neither Agree nor Disagree, A. = Agree, S.A. = Strongly Agree)

SA levels can be seen in Fig. 5.3.

No significant differences are found between the other measures collected. The comparison of workload between displays found no significant difference (F(2, 30) = 0.51, p = 0.61, $\eta^2 = 0.03$). For the system usability scale, no significance is found with the ANOVA (F(2, 30) = 0.97, p = 0.39, $\eta^2 = 0.06$). Finally, none of the utility questions yielded significant differences. However, the comparison of participants' subjective understanding of the servicer uncertainty was nearly significant (H(2) = 5.35, p = 0.069, $\eta^2 = 0.11$). Trends in the data yield toward highest perceived utility for VR followed by screen visualization, followed by baseline. The underlying data for these measures is in Fig. 5.5.

In sum, these results are in partial support of the hypothesis that 3D visualizations improve SA. They support the idea that 3D visualizations can improve level 2 and 3 SA, but not workload, usability, and utility. These results are contrary to the hypothesis that immersiveness will provide additional benefits over 3D visualizations.

5.4 Discussion

This study is one of the first to investigate the use of VR for remote monitoring of spacecraft rendezvous operations. The objective measure of SA shows significant differences between display modalities, with 3D visualizations improving Level 2 and Level 3 SA, but with VR harming Level 1 SA. Contrary to the hypothesis, the subjective measures of workload, and usability, did not show statistical differences. Similarly, subjective assessment of utility did not reach statistical significance but trended toward higher evaluations for displays with visualizations and immersion. In sum, these results provide insight into the understudied area of the utility of 3D visualizations and VR for operators in a remote supervisory, rather than direct command, of autonomous systems.

3D Visualizations and VR impact levels 1, 2, and 3 SA differently. Improving SA is critical for improving performance and enabling appropriate decisions, and poor SA has been a contributor to many accidents or errors [65]. All 3 levels of SA are important, and typically build off each other, such that level 2 SA requires level 1 SA, and level 3 SA requires level 2 SA. However, for remote monitoring and supervision of unintuitive orbital systems, operators will need to have an appropriate level 3 SA to understand collision risk and project the consequences of avoidance maneuvers [65], particularly under uncertainty. For satellite operations in particular, this is especially critical as collisions can adversely affect the viability of space operations across all orbital regimes [170].

While the SA level 1 results indicate that VR led to significantly worse performance over the baseline and screen visualization displays, the difference was only derived from information that was not in a fixed location on the VR display. When comparing the SA callouts, further analysis found no differences in the display modalities for items that were always present in the same location. For the VR display, this includes information in the heads up display. However, there were significant differences in the analysis of items that are in a static location in the baseline and screen visualization display but are tied to specific objects and change location over time in the VR display. The dynamic motion of the satellite states in VR made it so these objects are less salient, and more effort is required to find them. Participants are thus unable to have a consistent scan pattern. Scan patterns are often described using the SEEV (Salience, effort, expectancy, and value) model. Effort and salience are important aspects of this model which has been shown to be predictive of level 1 SA [171]. Although it may be desired to have all the data in a glanceable HUD which can improve monitoring performance over other information displays [172], this also can block visualized information, increase clutter, and disrupt immersion [146, 152, 173]. While tying some information to the satellites may reduce level 1 SA, it still is an important design consideration to avoid some of the pitfalls of HUD, such as putting too much information into the HUD obscuring the visualization. These results imply that information most critical to the success of the mission or information that needs to be consistently monitored should be located in a stationary component of a VR display.

The SA level 2 and 3 results find that 3D visualizations lead to an improved performance over the baseline display without 3D visualizations, but found no differences between the screen visualization and the VR displays. These initial results indicate that in this monitoring task VR does not impact performance. This agrees with the results of a prior remote monitoring VR study [19], which used a proxy for SA. These results are also in agreement with a monitoring study that compared only 3D visualizations to 2D visualizations and found that the 3D visualizations increased SA [174]. Other studies found that VR improves SA, although they consider a direct control paradigm of interacting with robotic systems [14]. The amount of control authority an operator has may be a contributing factor to these differences in results. There is a need for future work to consider other control paradigms that fall between direct control and monitoring, such as supervisory control.

The subjective measures of workload show no differences between display modalities. There are inconsistencies in previous literature as to whether VR increases [19] or decreases [14] workload over non-immersive displays. This research finds no differences. This may be due to the task and experimental paradigm itself: remote supervision, especially monitoring, is typically lower workload compared to direct control [66, 175, 176]. Thus, it is not unexpected that most users experienced similar levels of workload, as they had no control authority. Additionally, participants using the baseline display could not customize their display, while users in screen visualization and VR display could. The effort towards customizing the display or finding appropriate camera viewpoints could inflate the workload of visualization-based displays relative to the baseline. Most of the previous tasks that have found workload differences have been for direct control, where the operators are interacting with a system either through VR or a computer display and there is typically a higher workload overall [14, 53, 66, 175]. Due to inconsistencies in the literature and the varied degrees of operator engagement, future work should investigate other degrees of control authority, like supervisory control. While an ideal display would decrease workload over alternative displays, these results may be considered positive in that they did not exacerbate workload, indicating overall good display design.

For usability and utility, no significant differences are found between display modalities. In prior work, operators often subjectively rate VR displays to have a higher usability and prefer to work with them [19, 53], or prefer 3D visualizations over 2D visualizations [177]. In this research, statistical significance is not achieved, but the utility results trend toward significance. As such, these results are consistent with that of the literature where users tended to subjectively prefer the utility of the VR display. A critical difference between these results and those in the literature is that many of these studies used a within participants design where participants had a chance to experience multiple display modalities and thus their responses reflect these comparisons. By not doing a within participants design, this study is unable to capture some of these subjective preferences. Like workload, no differences in usability may be positive, as having a significantly worse display may be more indicative of poor display design or issues due to limitations with VR technology.

A challenge of this research was implementing a VR display that was designed appropriately, and to ensure that results were not influenced by a participant's inability to read the display or interpret and control the visualization effectively. The results of Aim 1, in addition to established principles for aerospace displays (such as MIL-SPEC and FAA regulations) and human factors (i.e., minimizing information access cost, contrast, and avoiding absolute judgment limits) guided the layout of the VR display. The impact of some of the results (such as the use of HUD and tying text to specific elements) can be extended by future designers of VR displays.

There are some limitations to this aim. Using participants with no prior familiarity with traditional displays makes it unclear how current, highly-trained operators would react to visualizations or a new system they are not as familiar with. Previous research for air traffic control found that while 3D visualizations improved SA among all participants, but those with extensive operational experience provided lower subjective ratings to 3D [174]. Future work should assess to see if the same is true for satellite operations, and if so how to best mitigate the issue of switching displays. In addition, the between-subject design used may impact the subjective measures as participants did not have a chance to experience all three displays and thus did not have the ability to compare between the features and limitations of each display.

While symptoms of nausea were monitored for during the experience, other cybersickness symptoms, like eye strain or fatigue, were not recorded. These other symptoms may influence the outcomes of the metrics assessed in the study or may discourage the use of VR during future operations. Additionally, this research considers the monitoring aspect of remote supervisory control of a simplified, faster than real time satellite operation, in which participants had no control authority. This work focuses on the monitoring of a remote system, which is how operators in such systems will spend the majority of their time in supervisory control paradigms, and therefore foregoes the inclusion of the ability to intermittently provide input to the autonomous system. It is critical to understand how display modalities can impact monitoring performance. If a display fails to facilitate effective monitoring, it will be difficult to use in supervisory control. Chapter 6 will further this work to include intermittent control and increased complexity. Finally, this research considers VR and immersion as applied through a head-mounted display. There are many other ways of providing an immersive environment, such as a CAVE system, and using a single display type represents a limitation of this work. Future research can study the impact of different degrees of immersion using different immersive systems on monitoring.

Beyond this, Aim 3 will expand on this experiment to consider remote supervision paradigms to understand how VR impacts these situations. This also more closely represents what operators might encounter during normal operations when remotely supervising autonomous or semi-autonomous agents. Understanding the impacts of display on monitoring is an important first step, as it is how the majority of an operator's time is spent, providing the operator with limited control authority to make interventions while in supervisory mode may allow display differences to be seen in metrics of workload, usability, and utility. This will also fill in the gap of understanding the effect VR has on various degrees of control authority, as remote supervision is understudied.

5.5 Summary and Contributions

This aim compares the effects of 3D visualizations and VR for remote monitoring of spacecraft operations on SA, workload, usability, and utility. Three displays, with varying degrees of 3D visualizations and immersion, were designed and evaluated through human subject testing. The results of this work indicate 3D visualizations may improve display interfaces for monitoring satellites; however, there is little evidence that immersion, such as that provided by VR, yields additional improvements. 3D visualizations improve level 2 and level 3 SA as measured through SPAM queries, which may lead to improvements for anomaly detection or anticipating collisions. VR reduces level 1 SA as measured through callouts, indicating that VR displays may not be beneficial for processing or monitoring text-based data; this reduction was only noticed when considering information that was not always present on the VR display. There are no differences between displays in workload, usability, and utility. While VR has been demonstrated to be a promising modality for direct control tasks, the benefits do not translate to remote monitoring of autonomous agents.

The overall contribution to the literature is the focus on VR for an operational monitoring task, as well as the inclusion of a screen based visualization modality. This aim builds upon the results of Aim 1, and enables Aim 3, which is a similar task but with a focus on supervisory control.

Chapter 6: Aim 3: Remote Supervision Operations

6.1 Introduction

The objective of this aim is to compare displays with different degrees of immersion or 3D visualization on a satellite supervision task where participants will have some, but limited, control authority over their satellites. VR has been shown to be promising for manual control tasks [5, 12, 14, 53], but does not appear to be promising for monitoring as seen in Aim 2 [144]. However, it is unclear how VR will translate for supervisory control, which is between these two extremes. The compared displays includes an immersive VR display, a 3D visualization computer-based display, and a baseline display with 2D graphical representation. These will be compared on measures of SA, performance, workload, usability, and subjective utility. We hypothesize that 3D visualizations and VR will lead to higher SA, performance, usability, and subjective utility, over the baseline display. We further hypothesize that VR will lead to an increase in these measures over the 3D visualization display. Finally, we hypothesize there will be no differences in workload among the displays.

The experiment for this aim is based on the scenario and displays used in Aim 2. Some key differences do exist. First, development was done to enable limited operator input and decisions, allowing real-time command of the spacecraft and subsequent orbits and uncertainties. Additionally, the uncertainties were modified to be more realistic, incorporating lighting, burn, and positional uncertainties that were constantly changing. Finally, displays were modified to allow for user input, an increased number of alerts, and minor design changes were made based on user feedback. The experiment and displays for this aim are also used in Aim 4.

6.2 Methods

In this research, three displays are designed to simulate a remote supervision of satellite operations. The simulated scenario is a spacecraft inspection task, in which an operator assists a servicer satellite to perform corrective burns and inspect a client satellite. These displays are compared through a human subject evaluation.

6.2.1 Scenario Design

Participants are tasked with monitoring and supervising the proximity portion of a satellite rendezvous mission, similar to that developed for Aim 2. The underlying trajectories used for the simulation are also developed using Basilisk, a high fidelity, flight-proven, physics-based satellite simulation tool [145]. To enable evaluation in a laboratory environment, the simulation is sped up by a factor of 15. The scenario consists of two satellites in orbit around Earth: a non-operational, tumbling client satellite and an active servicer satellite, supervised by a remote operator, sent to inspect the client satellite. The goal of the participant is to successfully complete the mission by servicing the client or abort, only if necessary, to avoid a collision. The client satellite has no communications, fuel, or battery, thus there is uncertainty in its location and attitude. In addition, there is uncertainty in the velocity change imparted by the thruster burn. These uncertainties are combined and visualized as a spheroidal “keep out zone”, which the servicer should avoid entering or there may be a collision. The keep out zone grows and shrinks based on environmental factors and participant actions. A proximity sensor is simulated to have better performance when the satellites are in the sunlight, causing the keep out zone to shrink; when they are in the shadow it grows due to worse sensor performance.

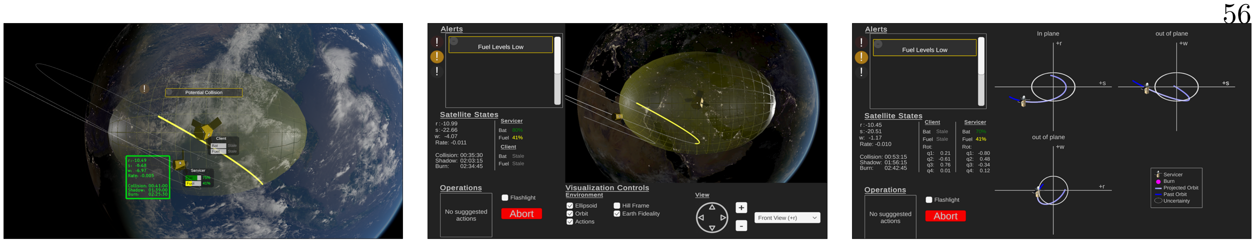
The servicer satellite begins on a drift orbit passing by the client satellite. During the simulation, the servicer fires its thrusters to change its relative orbit to circle the client satellite. The participant can select from one of three burn locations, each separated by 15 minutes of flight time. The ideal burn location may be influenced by the lighting conditions and battery levels. Different burn locations may result in successfully servicing the client satellite or being required to abort. The participant can also turn on a light, which improves the sensor’s performance, and thus reduces the uncertainty; however, this also drains the battery at a faster rate than nominal. The battery nominally decreases slowly in the darkness and increases in the sunlight. If the servicer enters the keep out zone, this is considered a collision. If a collision is unavoidable, the participant can elect to abort the mission, provided the servicer satellite has enough fuel, battery, and time. As such, using the onboard light too much when the battery is not sufficiently charged may result in an abort being impossible due to a low battery.

Different scenarios are created by manipulating the initial orbit, date, lighting conditions, fuel, and battery levels. This influences the outcomes of which burn location is ideal, how long the light should be on, and if the participant is forced to abort. Each trial lasted up to 7 minutes. Aborting or colliding causes the trial to end earlier. Eight experimental scenarios and two familiarization trials were developed. The experiential scenarios were of varying difficulty levels. The difficulty was determined on an easy, medium, and hard basis, based on the number of different actions that could lead to success (i.e., any burn location or amount of light use will lead to mission success is characterized as easy, versus only one specific burn location and specific amounts of light use to succeed is characterized as hard). In 2 of the 8 experimental trials, the most likely outcome was an abort.

6.2.2 Display Designs

As in Aim 2, three displays are designed for this experiment: a VR display, a 3D screen visualization (Scr. Viz.), and a two-dimensional Baseline display as seen in Fig. 6.1. These displays are based on those developed in Aim 2 [144] but with modifications to allow user input and design modifications based on feedback from Aim 2 participants. The displays are built using Unity 2022.3.21f. All three displays present the same information to the participant, but that information is conveyed with different degrees of immersion and 3D visualizations.

The VR display is seen in Fig. 6.1a. The underlying visualizations, including satellite models, relative orbital motion, accurate Earth models, and appropriate sun-based lighting, are based on the Vizard spacecraft simulation visualization software application [150] and is the same as Aim 2. Overlaid are visualizations of the keep out zone, centered around the client satellite, and colored based on collision risk. Red represents a warning, with less than 15 simulation minutes to a potential collision, yellow represents a potential collision at any point within the next two relative orbits,



(a)

(b)

(c)

Figure 6.1: The three different display designs. (a) VR display (b) Screen Visualization display (c) Baseline display

gray represents no collision risk, and blue indicates the keep out zone encompassed only attitude uncertainty was at the minimum size. The portion of the servicer orbit line, where it is projected to enter the keep out zone, is changed to the corresponding condition color. Finally, a representative light is visible when the participant turns on the light.

Similar to the previous Aim 2 VR display, critical information and user input options are displayed in a text-based form. A HUD displays the range, rate, and time to burn, collision, and lighting changes. This is always located in the same peripheral location of the headset, allowing participants to access information through eye movements, minimizing information access effort [151] and minimizing blocking the visualization. Satellite states, such as fuel and battery, are presented both as text and with a gauge, and move with their associated satellite. While this was determined to reduce level one SA for the monitoring task in Aim 2, this was still determined to be the optimal way to display this information, while minimizing text in the HUD and preserving immersion as found in Aim 1 [129, 130, 146, 152, 153]. Text-based descriptions of cautions and warnings are displayed at the top of the screen and are designed to be easily noticed by an participant. These are triggered automatically for certain events, including low battery, fuel, and potential for a collision. As an improvement to the VR display, alerts can now be minimized to remove them from the primary field of view or maximized to review them again [147, 148]. Finally, the user input is provided through a panel with various options presented above the HUD when available. User input panels disappear after an option is selected or the time window for user input to the panel closes.

Similar to the display from Aim 2, participants can control their visualization and operations through a radial menu system and their hand controller. The right controller is used to control the menu and user inputs. This allowed them to turn on and off components of the visualization or text. Additionally, they can change their viewpoint or perform operations like viewing alerts, turning on the light, or aborting. The left-hand controller controls navigation, such as panning, zooming, and teleportation. Additionally, participants can walk around or rotate their bodies to further navigate the scene. Participants had access to a 10 by 10 foot area of floor; the tracking dynamics were set such that the participant could cover the distance to the satellite while staying in this area. Additionally, preset buttons allow easy access to turning on/off the light or aborting. An abort requires confirmation to ensure it was not selected in error. The display and controls underwent human factors testing prior to the experiment to ensure the text was readable and the controllers were acceptable.

The Scr. Viz. display maintains the 3D visualization described for the VR display, but is not immersive like VR. This is seen in Fig. 6.1b. The participant can still interact with the system by panning, zooming, and customizing their visualization, but this time on a 2D computer screen using a mouse. All text, including the HUD, satellite states, alerts, and user input are displayed on the screen outside the visualization, allowing for a consistent scan pattern.

Finally, the Baseline display is designed to be consistent with the text-heavy and graphical displays currently used for most traditional satellite operations. It has only 2D visualizations where the 3D visualization is replaced by graphical telemetry. The range is displayed as plots of the different orbital planes and past and future motion. The same text-based interface in the Scr. Viz. display also surrounded the telemetry plots. For both the Scr. Viz. and the Baseline displays, navigation, and interactions were done with a mouse.

6.2.3 Experimental Design

The study was approved by the University of Colorado Institutional Review Board (#24-0250). Informed consent was obtained from all participants. 45 participants from around the University of Colorado Boulder campus were enrolled and completed data collection (18 female, 27

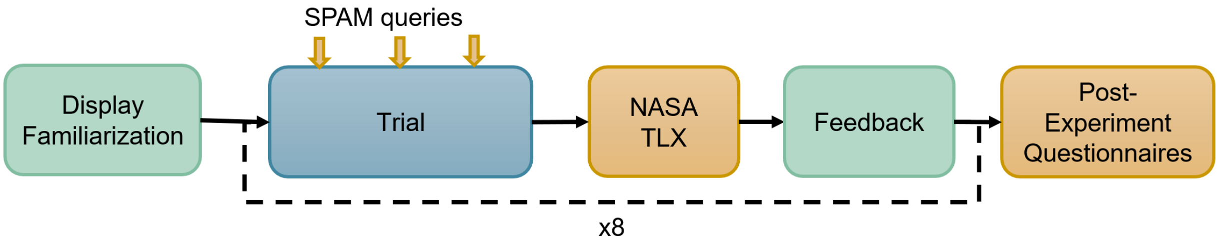


Figure 6.2: Experimental design flowchart. The orange boxes indicate data collection through surveys or queries.

male; ages 18-38, median age 23 years). 47 participants were enrolled, but 2 voluntarily did not complete both visits of the study and thus were removed. Participants were aware of the high-level project goals from the informed consent but naive to the exact manipulations or alternative display designs. Participants were screened to ensure their vision was correctable to 20/20, they were not colorblind, and they scored less than 90% on the Motion Sickness Susceptibility Questionnaire [154]. This was used to identify individuals who may be highly susceptible to simulator sickness before data collection. Additionally, participants in VR were monitored for cybersickness symptoms, particularly nausea, throughout the experiment; there were no reported cases of nausea in VR.

The experiment took place over two visits to allow for the assessment of both operations and training. In the first visit, participants were randomly assigned to one of three display conditions: Baseline, Scr. Viz., and VR, as described above. The second visit was the evaluation day, where participants were asked to perform satellite operations using the Baseline display (Fig. 6.1c), as would be consistent with displays for actual operators. Both days followed the procedures as described below and in Fig. 6.2.

During both visits, participants were first familiarized with the task and their assigned display. This was done through a PowerPoint which reviewed the task's motivations and goals, as well as how to use their display. They then did three familiarization trials. The first trial was primarily an opportunity to get used to the controls and display – there was no collision risk present. The next two familiarization trials mimicked a real trial and had identical initial states to give participants an opportunity to execute different actions so they could understand the impact on the mission outcome. This was done to ensure that participants understood how to use the displays and perform the task before the real trials began and to eliminate issues of performance due to lack of familiarity with the display. During both the PowerPoint training and familiarization trials, participants were quizzed to ensure they understood the necessary information.

Participants then completed eight trials. Every participant completed the same eight trials on both visits, but the scenarios were presented in a randomized order. During each trial SA, workload, and performance were assessed. To assess the three levels of SA, SPAM was used [157]. Up to three points throughout the trial, an auditory tone was played. Participants were instructed to provide verbal confirmation when they had the ability to answer the questions. At this point, or after 20 seconds had elapsed, an experimenter asked the participant three questions, one per SA level. Questions were generated through a process similar to a goal directed task analysis [28]. Example questions include: “Is the servicer satellite currently in the sun? (level 1)”, ‘‘Is the portion of the orbit line in the keep out zone decreasing? ” (level 2), or “If no new action is taken, will the keep out zone be shrinking in 30 minutes?” (level 3). The full list of questions is in the supplementary materials. Participants were instructed to respond as quickly and as accurately as possible and were allowed to use the display to answer. If a trial ended early due to an abort or collision, fewer questions may have been asked.

After each trial, participants assessed their workload through the NASA TLX [35]. To do so, they rated six dimensions of workload on a 21-point scale. This includes mental, physical, temporal, performance, effort, and frustration. At the end of all eight trials, participants then completed comparisons between the different components, allowing for a weighted workload score to be calculated. After completing the TLX survey for a trial, participants were given feedback on their performance. This included information about their burn performance (did they make a good burn selection that could lead to success), end state performance (considering success, abort, or collision), and combined total performance. Each of these was on a scale of ‘poor’, ‘fair’, ‘good’, and ‘excellent’. This performance also corresponded to a monetary bonus participants could earn between $-1.00 and $1.00 per trial. Performance-based earnings were cumulative over all trials.

Finally, after all trials were finished, participants completed the SUS [164], which is a 10-question survey in which participants respond on a 5-point scale. These are combined to give a resulting score from 0 to 100. In addition, participants answered a custom subjective utility survey which had Likert-style questions relating to their perceived understanding of the events, uncertainties, collision likelihood, orbital motion, and operational decision, as well as free-response questions about their experience. The full survey is provided in the supplementary materials, and was customized for each visit. On visit 2, this survey included questions about how their training in the assigned display from visit 1 impacted their performance during visit 2. On visit 1, they also completed a demographics survey including information sex, orbital mechanics familiarity, operational familiarity, and VR familiarity. Their familiarity was coded as a binary 'little to none' or 'moderate to high'. Finally, at the end of visit 2, they completed the balloon risk analog task (BART) [178, 179] to assess their risk-taking behavior.

6.2.4 Statistical Analysis

The primary objective of this aim is to compare the three displays for operational use. Only the data from visit 1 was considered; the data from visit 2 is considered in Aim 4. In both cases, the same statistical pipeline is used for each of the metrics. Analysis was performed to compare the three displays on SA, performance, workload, usability, and subjective utility during the first visit. Each SA level was analyzed independently using the percentage of questions answered correctly across all trials. This resulted in one measure per participant per SA level. This transformation into a percentage was made because the participants answered different numbers of questions due to their individual performance. Performance was calculated for each trial based on the appropriateness of their actions; a full description is in Appendix C.

The study collected 360 trials over 45 participants. One score was collected for each participant for SA level, usability, and utility, for a total of 45 data points per assessment type. For workload and performance, trials were kept separate and so the 360 data points were used. A criterion of $\alpha = 0.05$ was used for significance for all tests, and all assumptions for each statistical test were met.

Prior to statistical analysis, the SA, performance, workload data were inspected for a confounding factor of trial order to capture undesirable learning effects. No effect of trial order was identified for any analysis. However, for performance and workload, there was a dependence on the specific scenario (i.e., regardless of the order presented, some trials had consistently different performance and workload than others). Thus, for these statistical tests, these factors were included in the model. In addition, all data was evaluated for confounds based on participant background including: sex, orbital experience, satellite operational experience, VR familiarity, and BART score. These were included in the models as appropriate.

For each SA level, a linear mixed effects model was used to compare the effect of display modality on SA. The display modality was considered as a fixed effect. The random effects were dependent on SA level. The model was fit using the *lme4* package in R via penalized maximum likelihood estimation [166]. The significance of display modality was assessed using an F test with a type III ANOVA with a Satterthwaite approximation for degrees of freedom and was implemented using *lmerTest* package in R [167]. Any necessary post-hoc tests were done between all pairwise comparisons using estimated marginal means (*emmeans* package in R [168]) with a Tukey p-value correction and Kenward-Roger degrees of freedom correction. Effect sizes were calculated using the *effectsize* package in R [169]

Performance is on a 12-point ordinal scale, and as such, a cumulative linked mixed model approach was taken (using *clmm* in R [180]). This score is a combination of the participant's burn decision and end state (including aspects such as battery level, appropriateness of abort decisions, and use of the light). The full metric calculation is described in the supplementary material. The training display and its interaction with trial difficulty were included as a fixed effect, and the number of balloons collected in the BART task [181], which has been shown to be correlated with risk, was also included as a random effect. A significant interaction was found, indicating that display modality may not provide differences across difficulty levels. Thus, the data was separated by scenario difficulty level and analyzed them separately, keeping the same fixed and random effects. A type III ANOVA was used to compare performance across conditions (*anova.clmm* in the RVAideMemoire package in R [182]). Post-hoc tests were done between all pairwise comparisons using estimated marginal means [168] with a Tukey p-value correction.

As another measure of performance, the overall outcomes were compared across training conditions including the number of scenarios aborted, the number of scenarios resulting in a collision, and the number of successful scenarios. This was done using a Kruskal-Wallis H-test.

For workload, the unique weighted TLX score for each trial was analyzed, resulting in 360 total data points. As with SA, a linear mixed effects model was fit using the training display condition as the fixed effect. The participant and scenario were included as random effects to account for the repeated measures and differences across each scenario. Likewise, to assess usability scores, a linear mixed effects model was fit with the training display as a fixed effect and sex as a random effect. In both cases the same pipeline as SA was followed.

The subjective utility questions were each on a 5-point Likert scale. Thus, ANOVAs could not be used to compare ratings across conditions, and instead, each question was analyzed using a Kruskal-Wallis H-test. Post-hoc tests were done using Dunn's test with a Holm correction [183].

6.3 Results

No significant differences in SA are found among the displays in any of the levels, as seen in Fig. 6.3. The ANOVA comparison between the linear mixed effects models found that all were trending towards, but did not reach, significant differences in the percent of questions answered correctly with level 1 (F(2,23.11) = 2.3, p = 0.12, $\eta^2 = 0.17$), level 2 (F(2,41.83) = 2.03, p = 0.14, $\eta^2 = 0.09$), and level 3 (F(2,41.57) = 2.42, p = 0.10, $\eta^2 = 0.13$).

In the comparison of participant's performance, significant differences are found among the Hard scenarios ($\chi^2(2) = 9.61$, p = 0.008), but no differences among the Easy ($\chi^2(2) = 3.53$, p = 0.17) or Medium ($\chi^2(2) = 0.70$, p = 0.70) difficulty scenarios. Note that for both the Easy and Medium conditions, participant performance was frequently at the maximum of the scale. For the

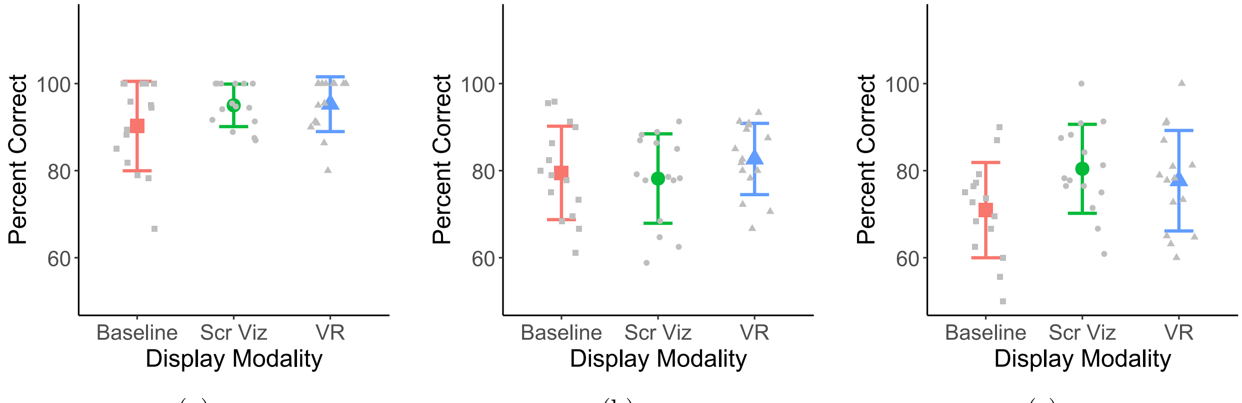


Figure 6.3: The (a) level 1, (b) level 2, and (c) level 3 SA results. All figures show the participant averaged percent of SPAM queries of that level answered correctly. The data mean, standard deviation error bars, and significance is noted between the conditions. Note that the Y axis ranges from 50% to 100%.

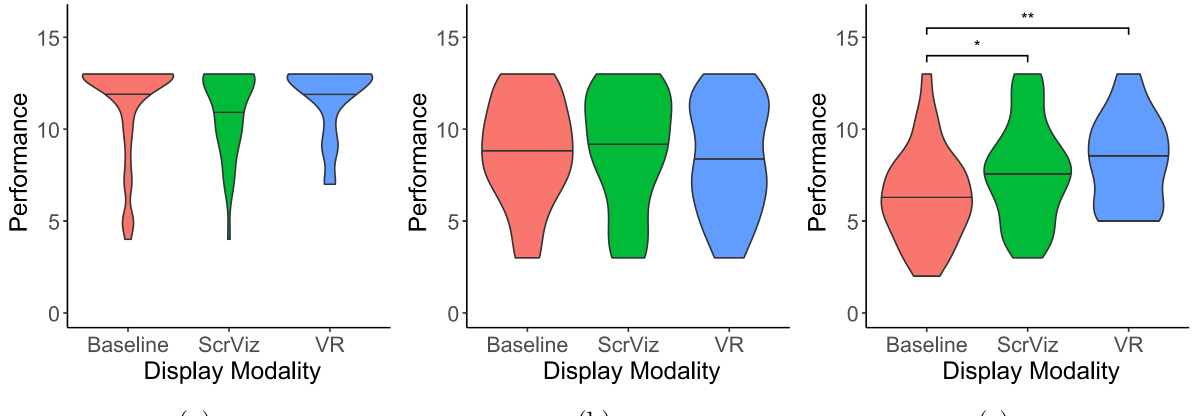
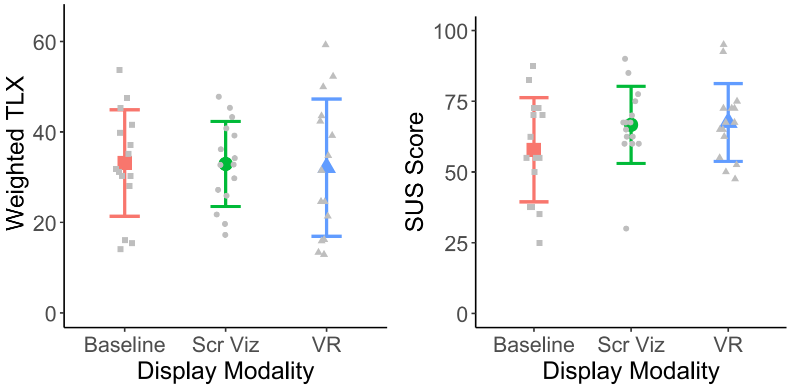


Figure 6.4: Participant's performance on (a) Easy, (b) Medium, (c) Hard trials. The violin plot is overlayed with the median score, and significance is noted between the conditions.

Hard scenarios, differences are seen between Baseline and VR (z = -3.06, p = 0.006) and Baseline and Scr. Viz. (z = -2.36, p = 0.047). For both cases, participants in the Baseline condition performed worse. No differences are found between VR and Scr. Viz. (z = -0.74, p = 0.74). This can be seen in Fig. 6.4. No differences are found in the number of scenarios aborted (H(2) = 0.39, p = 0.82, η^2 = -0.038), the number of scenarios with collisions (H(2) = 4.20, p = 0.13, η^2 = 0.05), or the number of successful scenarios (H(2) = 0.2, p = 0.87, η^2= -0.04).

For workload, no significant differences are found in the ANOVA (F(2,41.7) = 0.03, p = 0.97, η^2 = 0.001) as seen in Fig 6.5a. Additionally, no significant differences are found in usability's



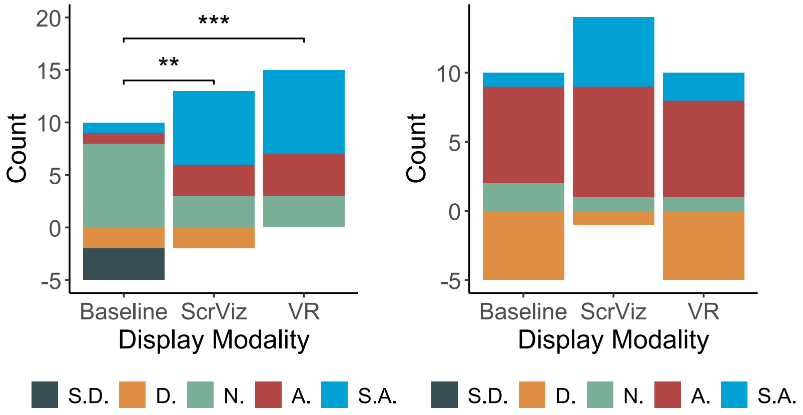
(a) (b)

Figure 6.5: The (a) workload, and (b) usability results. Workload shows the participant-average weighted TLX score, and usability shows the System Usability Scale score. The data mean, standard deviation error bars, and significance arenoted between the conditions.

ANOVA (F(2,42) = 1.82, p= 0.17, η^2 = 0.08), as in Fig. 6.5b. However, the coefficient for the VR term in the linear model approached significance (p = 0.09), which indicates that there are trending differences between the VR and Baseline usability score.

Participants perceived differences in utility between the display conditions in two aspects. Differences are reported in the participants' perceived ability to understand orbital motion (H(2) = 14.6, p = 0.006, η^2= 0.30) and ability to make operational decisions (H(2) = 6.11, p = 0.047, η^2= 0.10). Post-hoc comparisons for orbital motion using the Holm correction found differences in the Baseline and Scr. Viz. (z = -2.86, p = 0.008) and Baseline and VR (z = -3.61, p<0.005) with the Baseline display rated significantly worse in both cases. No differences are found between Scr. Viz. and VR (z = -0.75, p =0.45). Additionally, no significant post-hoc comparisons are found for the ability to make appropriate operational decisions. However, the Baseline to Scr. Viz. comparison was trending significant (z= -2.31, p = 0.061). No other differences in subjectively reported utility are found.

Participants' subjective written comments regarding aspects relevant to the display and/or interface are also analyzed. The VR display was found to be polarizing. Some participants enjoyed it *"It is rather intuitive and easy to understand"*, *"It is a good user interface"*, while others found it hard to use *"I felt like there was no benefit to being in VR and it just makes it clumsier and harder*



(a) (b)

Figure 6.6: The subjective utility questions, showing the results of (a) "I found this system enabled me to understand the relative orbital motion of the satellites" and (b) "I found this system allowed me to make appropriate operation decisions" (S.D. = Strongly Disagree, D. = Disagree, N. = Neither Agree nor Disagree, A. = Agree, S.A. = Strongly Agree)

to use the menu". No participants left insight about the Scr. Viz. display itself. For the Baseline display comments were negative and uniformly indicated a dislike of the 2D nature *"Would be much easier if we could see the orbit paths in 3D vs on a 2D plot", "I had to visualize the 3D space . . . which made the task more demanding".* Similar comments were echoed among the other Baseline participants.

6.4 Discussion

This study compared displays with different degrees of 3D visualizations and immersion in a remote supervision task and is one of the first studies to investigate VR for remote supervision operations. This research helps fill the gap in understanding of how VR may be used in an operational environment that is not full manual control or passive monitoring.

No differences are found in any of the SA levels. This rejects the hypothesis that VR or 3D visualizations will improve SA, as was seen in Aim 2. Previous literature is mixed on the effects of display modality on SA. A comparison of 3D visualizations and traditional displays for an air traffic control supervision task found that 3D visualizations improved SA [174]; likewise, for remote monitoring of autonomous surface vehicles, both VR and 3D visualizations improved SA over a 2D display [19]. Additionally, Aim 2 found that both VR and screen visualization improved SA over the baseline for level 2 and 3 SA, but that VR performed worse for level 1 SA [144]. However, some direct control studies have found no differences in VR with respect to SA [64], while others have found VR improves SA [57]. These differences may be a result of how SA is measured. Previous studies have used subjective and proxy measures for SA, which may not be correlated with those found with SPAM or other objective SA measurements [65, 184, 185]. Likewise, these results may also be attributed to limitations of SPAM, particularly using an accuracy, not response time, based analysis. SPAM may have less sensitivity than other objective measurements [185], influencing the results, but is still an appropriate tool for operational environments. This result may also be a limitation of using a goal-directed task analysis to develop SPAM questions; questions were determined based on what information was needed to perform the task successfully. While in all cases the questions could be answered by any display, some questions did not require the use of any visualization to determine the answer, which may not allow differences in displays to be seen, particularly for level 1 SA. Additionally, this may indicate that not all forms of operations may benefit from the additional immersion or visualization.

Performance was improved by the use of 3D visualizations, both on the screen or in VR, but only in the trials of the hardest difficulty. This result highlights a benefit of including 3D visualizations, either from VR or on a computer screen. This scenario is already a simplified spaceflight operation for the purposes of research. Thus, these promising results may indicate that as more complex real-world scenarios are implemented, the potential benefit of the VR display could be further enhanced. As satellite operations become more complex and challenging for operators, there may be advantages to increasing the fidelity of the visualizations to help operators perform better. The improvements align with previous manual control research that suggests that VR can improve performance over desktop visualizations [53].

No differences were found in workload, which supports the hypothesis and agrees with Aim 2. While processing 3D visualizations on a 2D display may increase mental workload [8–10] and subjective reports appear to indicate an increase in mental demand required to process the 2D data, the overall increase in workload is not reflected in the data. Additionally, while VR may increase physical workload (due to movement or using the controller) [186, 187], this is also not reflected in the overall workload, as participants often placed a low weighting factor on physical workload. Furthermore, there could be a risk that the novelty and physical aspects of the VR environment could have increased workload but this was not found to be the case. Due to the low workload nature of monitoring and supervision [66,175,176], with traditional displays it was unexpected that any one new display could significantly reduce workload overall.

For usability, there were no differences in displays, which agrees with Aim 2. However, this disagrees with other experiments that often subjectively have VR improve usability [19, 53]. Notably, these previous studies are within subjects, meaning subjects were able to use all displays. Using a between-subjects design means that these results are indicative of an objective independent evaluation of usability, rather than a potential comparative assessment of the displays that may be seen in a within-subjects design (i.e., subjects rate their least favorite display lower based on preference). The overall high level of usability that was reported, even for the baseline display, is indicative of the appropriateness of the display design in each condition.

The subjective utility found that 3D visualizations (either on a screen or in VR) can improve the participants' understanding of orbital motion. This is a key aspect for future satellite operations, as complex relative orbits can be difficult to understand intuitively [34]. This may explain some differences in performance. This conclusion is supported by the fact that a majority of the subjective comments about the Baseline display included a dislike of the orbit paths and not being able to visualize them.

This research shares some of the same limitations as Aim 2. While this study included limited operator control over the satellite systems, the scenario was not as operationally complex as real-world scenarios, which may also include aspects such as limited communication, anomalies, more telemetry streams, and slower and prolonged operations. The short time span likely reduced the amount of boredom participants experienced and may have reduced the amount of discomfort the VR headset caused. VR, for prolonged use, may be uncomfortable due to the headset weight and eyestrain [88]. While this study did not elicit these types of responses from participants, it does limit the ability to answer these research questions with increased experimental fidelity.

Future work involves expanding these experiments to more complex operations and using trained operators to understand the effects of VR. It should also include different use cases for VR within operations. While VR may not be promising for use in continuous operations, the subjective responses and utility show that VR may offer benefits as a way to intuitively understand the orbits. This may mean that it could be useful in advanced planning of operations to understand the orbits, visualize the environment during a difficult or sensitive maneuver, or for improving training (Aim 4).

6.5 Summary and Contributions

This study compares the effects of 3D visualization and immersion for remote supervision of a spacecraft operation on SA, performance, workload, usability, and utility. Three displays were designed for a satellite rendezvous task and compared via a human subject experiment. The results of this work indicate that 3D visualizations, whether on a computer screen or in VR, may improve utility and performance in complex scenarios. In future, even more complex operations, these 3D visualizations may be important to include. However, the results show that there is little evidence that immersion through VR provides additional benefits. There are no differences between displays in SA, workload, and usability. While VR has been shown in other research to provide benefits in manual control operations, it does not translate to these findings for remote supervision. These conclusions can also inform other supervisory operations that are emerging, such as transportation, manufacturing, and robotics.

The main contribution to this literature is the study of VR for supervisory control operations, which is understudied. The same experiment and task from this aim are also used in Aim 4.

Chapter 7: Aim 4: Remote Supervision Training

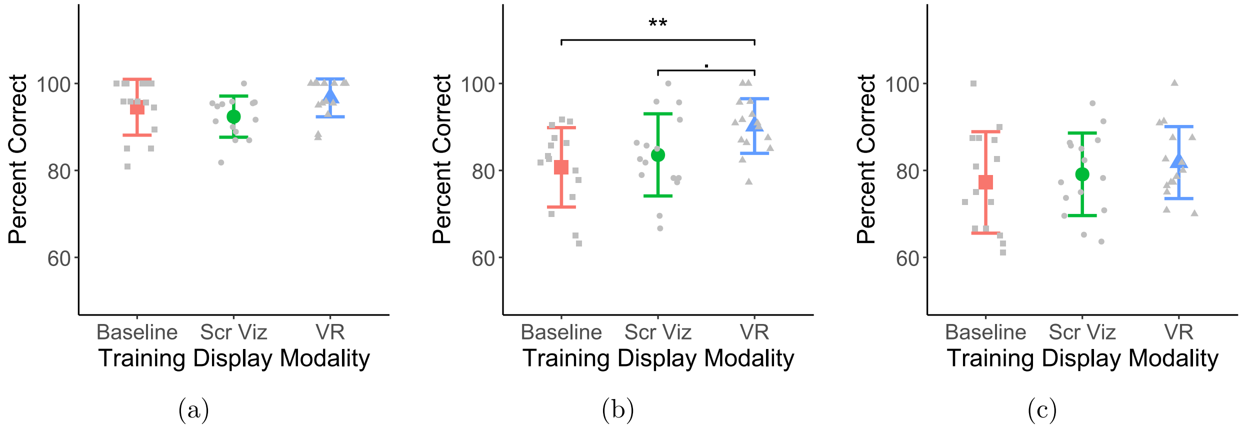
7.1 Introduction

The objective of this aim is to understand how training in alternative display modalities impacts the operations of a satellite supervision task using traditional displays. Many organizations are unlikely to adopt VR displays for operations due to long lead times in technology transitions, increased adoption cost compared to including 3D visualizations on existing computers, and the lack of strong benefits as seen in Aims 2 and 3. However, VR may be promising for training and require less of a barrier to entry, allowing it to be more easily adopted by organizations.

As in Aim 3, the training displays include an immersive VR display, a 3D visualization display, and a baseline display with traditional 2D graphical representations. After training (Aim 3), participants complete the satellite task in the baseline display, which is done on a different day. Their SA, performance, and workload during their second visit are compared to understand the effect of the training modality on operational performance. Additionally, their responses to subjective usability and utility of training are compared. We hypothesize that training with 3D visualizations will improve SA and performance in the second visit, as well as achieve a higher subjective utility. Furthermore, we hypothesize that VR will lead to further improvements in these metrics than 3D visualizations alone. We do not hypothesize there will be a difference in workload scores or usability based on training conditions.

7.2 Methods

This aim uses the same scenario, displays, and experimental design, participants, and statistical pipeline from Aim 3, and their descriptions are included in chapter 6. This Aim is concerned with the second visit, when participants were performing the task in the Baseline display. The first visit is treated as training, which is done through exposure, not through adaptive or modulated difficulty. This allows participants to experience different scenarios and build mental models. During the second visit, all participants are familiarized with the traditional display to reduce the effects



(a) (b) (c)

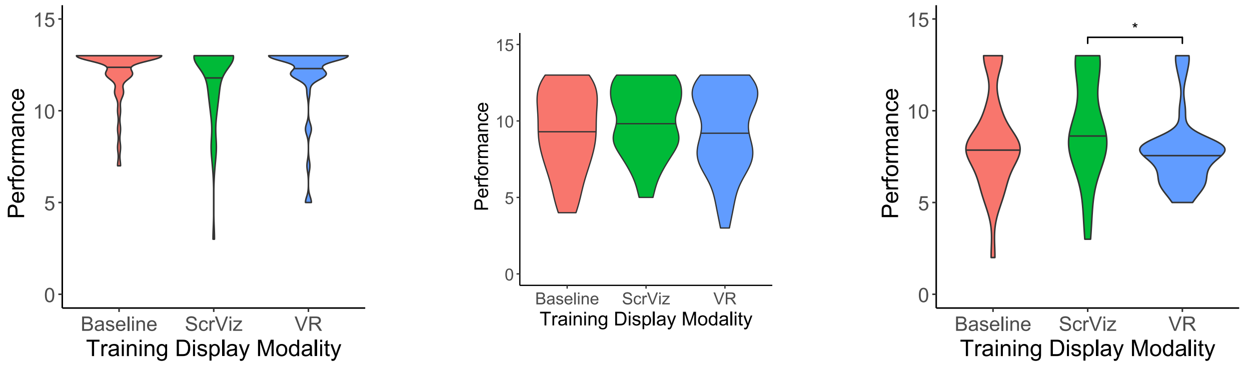
Figure 7.1: The level 1 (a), 2 (b), and 3 (c) SA results. All figures show the participant averaged percent of SPAM queries of that level answered correctly. The data mean, standard deviation error bars, and significance is noted between the conditions.

from learning to use a new display. Comparing participants' visit 2 data, grouping by visit 1 display condition, allows for an understanding of whether training in alternative immersive modalities allows operators to gain context and intuition more easily, facilitating improved operations. The difference between visit 1 and visit 2 data was not studied, as it is more important to understand which training modality leads to the best overall outcomes.

7.3 Results

The results show a significant difference in level 2 SA for the percent of questions answered correctly (F(2, 40.048) = 5.83, p = 0.006, η^2 = 0.23). Post-hoc comparisons found that those who trained in VR improved over those who trained in the Baseline display (t(41.4) = -3.22, p = 0.007, d = -1.20). Those who trained in VR approached, but did not reach, statistically improved performance compared to those who trained in Scr. Viz. (t(36.7) = -2.17, = 0.09, d = -0.98). There was no difference between those who trained in Scr. Viz. and the Baseline screen condition (t(40.9) = -0.52, p = 0.86, d = -0.22). Level 1 SA trended toward, but did not reach, significance (F(2, 42) = 2.45, p = 0.09, η^2 = 0.11). There are no changes in level 3 SA (F(2, 41.61) = 0.73, p = 0.49, η^2 = 0.03). Fig. 7.1 shows the level of SA achieved for all groups across the three levels.

Performance is shown in Fig. 7.2. No difference was found between training modalities for



(a) (b) (c)

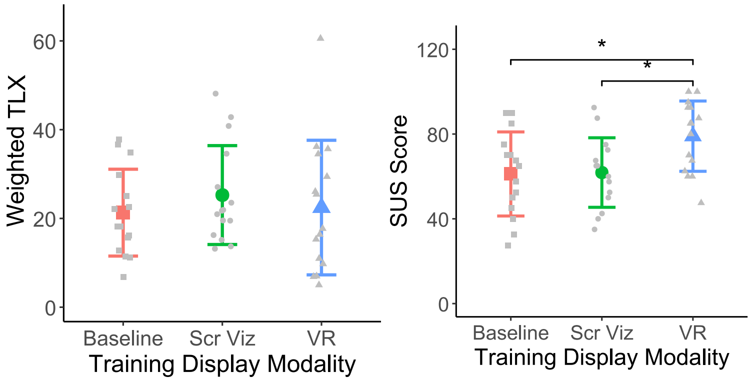
Figure 7.2: Participant's performance on Easy (a), Medium (b), Hard (c) trials. The violin plot is shown with the median score, and significance is noted between the conditions.

the Easy ($\chi^2(2) = 3.84$, p = 0.15) and Medium ($\chi^2(2) = 3.49$, p = 0.17) difficulty trials. However, for the Hard difficulty trials, performance differed depending on the training condition ($\chi^2(2) = 6.78$, p = 0.034). Post-hoc comparisons on the Hard difficulty trials only show differences between those trained in Scr. Viz. and VR (z = 2.55, p = 0.03), with those trained in Scr. Viz. performing better. No differences are found between Baseline and Scr. Viz. (z = -11.66, p = 0.22) and Baseline and VR (z = 0.89, p = 0.65). Additionally, there are no differences in the number of aborts (H(2) = 1.21, p 0.55, η^2= -0.02), number of collisions (H(2) = 2.16, p = 0.33, η^2 = 0.003), or number of successes (H(2) = 0.84), p =0.65, η^2 = -0.03) between participants trained in different conditions

There are no significant difference in workload (F(2, 41.9)= 0.40, p=0.67, η^2 = 0.02). Results are seen in Fig. 7.3a.

Usability was significantly different depending on the screen condition in which participants were trained (F(2, 42.23)= 5.91, p=0.005, η^2 = 0.21), as seen in Fig. 7.3b. Post-hoc tests found participants rated the Baseline screen more usable if they had trained in VR than if they had trained in either the Baseline condition (t(41) = -2.90, p = 0.016, d = -1.06) or the Scr. Viz. condition (t(41) = -2.79, p = 0.021, d = -1.02). No differences are found between how participants rated the usability between those who trained in the Baseline and Scr. Viz. conditions (t(41) = -0.11, p = 0.99, d = -0.04).

Finally, for subjective utility, there are differences between how people perceived the effec-

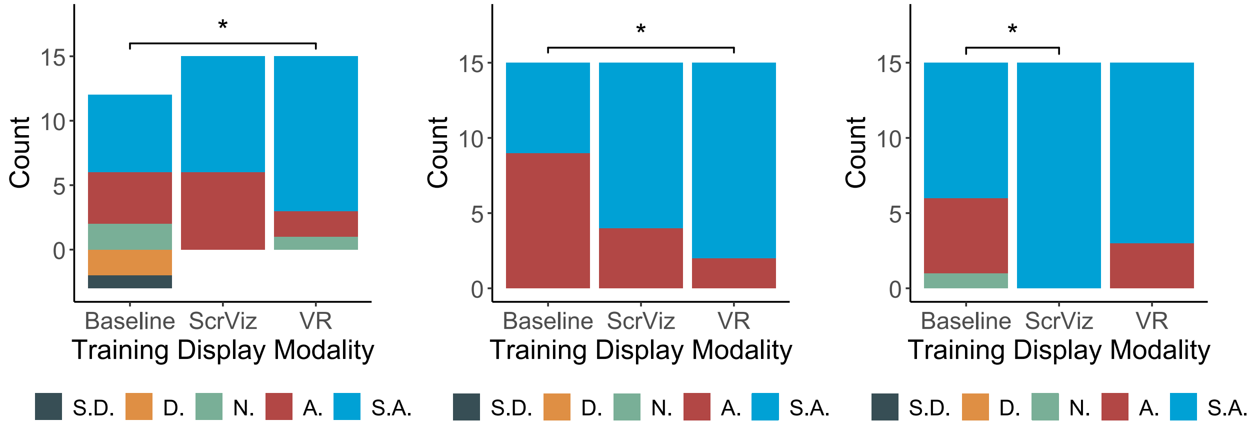


(a) (b)

Figure 7.3: The Workload (a), and usability (b) results. Workload shows the participant-average weighted TLX score, and usability shows the System Usability Scale score. The data mean, standard deviation error bars, and significance is noted between the conditions.

tiveness of their training to help them understand the relative orbital motion (H(2) = 6.39 p = 0.041, $\eta^2 = 0.10$), understand collision likelihood (H(2) = 7.62, p = 0.022, $\eta^2 = 0.13$), and promote event awareness (H(2) = 7.36 p = 0.024, $\eta^2 = 0.13$). In addition, their perception of understanding uncertainties was trending toward significant (H(2)= 5.88, p = 0.053, $\eta^2 = 0.09$). Post-hoc tests found differences in understanding orbital motion between those who trained in the Baseline and VR displays (z = -2.47, p = 0.039), indicating VR was perceived to help participants understand orbital mechanics more, and no differences between Baseline and Scr. Viz. (z = -1.67, p = 0.19) or Scr. Viz. and VR (z = -0.81, p = 0.42). For promoting the understanding of collision likelihood, differences are found between Baseline and VR (z = -2.68, p = 0.022), indicating VR facilitated a perceived improved understanding of how likely a collision is to occur. No differences are found between Baseline and Scr. Viz. (z = -1.91, p = 0.11) or Scr. Viz. and VR (z = -0.76, p = 0.44). For both of these cases, VR had higher ratings than Baseline. Finally, for promoting event awareness, differences are seen between Baseline and Scr. Viz. (z = -2.73, p = 0.019), where visualizations were perceived to improve the participant's understanding of critical events. No differences are found between Baseline and VR (z = -1.41, p = 0.32) or Scr. Viz. and VR (z = 1.32, p = 0.19). Scr. Viz. had higher subjective ratings than Baseline.

Participant subjective comments indicated they found benefits in learning from VR *"The*



(a) (b) (c)

Figure 7.4: The subjective utility questions, showing the results of (a) "I found that the training from the first visit was effective in enabling me to understand the relative orbital motion of the satellites today." (b) "I found that the training from the first visit was effective in enabling my understanding of collision likelihood today.", and (c) "I found that the training from the first visit was effective in enabling me to understand mission critical events today." (S.D. = Strongly Disagree, D. = Disagree, N. = Neither Agree nor Disagree, A. = Agree, S.A. = Strongly Agree)

overall orbit is much easier to see in VR, which is helpful.", *"I felt like I was better able to visualize the trajectories based on the 2D graphs today [on the baseline display] because I had already seen the 3D equivalents"*, and *"My training on the first day helped me greatly in my decision making process. I had developed a process on day one that was still applicable today. I looked for the same flight telemetry metrics to base my decisions on today as I did on day one. The first day also helped me to better visualize the 2D displays in front of me because I had already seen the 3D simulation and knew what to picture the graphs as."*. These were echoed among the other participants. Similar comments were reported among those trained with the 3D visualization in Scr. Viz. *"I think the transition from 3D to 2D made it much easier to understand the 2D display than I feel it would have been if I could only see the 2D perspectives."*, *"The first day of training was more visual and enhanced my ability to understand orbit approaches and the associated nuances."*. Participants did not provide comments about training in the baseline display.

7.4 Discussion

This study compares the effects of training display modality with different degrees of 3D visualizations and immersion for a satellite supervision task, and is one of the first studies to investigate training in VR for supervisory control tasks. This helps fill the gap in understanding how VR may be used for satellite operation training and what benefits it may offer, even when visualizations may not be feasible in operations themselves. The Baseline display training condition is a control and helps to set a reference point for the familiarization that occurs with repeated interactions on the same system by using a traditional satellite operations display set-up, and is also consistent with current satellite operations and training. Further improvements over the Baseline display in any metric indicate additional benefits achieved from training with a system with enhanced displays.

Training in VR improves level 2 SA (comprehension) and approached significance in level 1 SA (perception), indicating a benefit for VR-based training. However, no improvement in level 3 SA (projection) is found. This is in partial support of the hypothesis that training in VR would improve all levels of SA. The lack of significance in level 1 SA is likely due to the fact that all groups had high SA perception scores, making it difficult to discern differences in perception across the groups. This is beneficial and indicates that all groups are able to achieve the high level of perception needed for operations, and is similar to the results seen in Aim 3. The difference in level 2 SA may be attributed to improved mental models of the scenario. Mental models are a contributing factor to SA, especially level 2 and level 3 [45]. While previous work on VR training has not specifically studied SA, biology education research has suggested that learning in VR leads to improved mental models over non-immersive displays [188]; likewise, 3D visualizations have been shown to improve mental models over 2D for electron orbits [189]. This could be one potential mechanism for how VR training is able to improve SA. Additionally, being able to improve level 2 SA through VR is important to achieve safer operations. High SA is critical to maintaining safety and performance [190,191]. Low SA has been attributed to many human-caused accidents in other fields like aviation [192, 193], maritime [194], and nuclear power plant operations [195].

Differences in performance are seen among the Hard trials, where training in the Scr. Viz. display led to better performance than training in VR. In the Easy and Medium trials, participants had a high level of success in all conditions, and no differences were seen. While participants trained in VR performed worse than those trained with Scr. Viz. on the Hard trials, there was no difference in performance between those trained in Baseline or in VR. This is a positive, as it indicates that VR does not harm performance compared to traditional training displays, and so would not be detrimental to use. Some of these differences, or lack of differences, may be attributed to being unfamiliar with the display. The Scr. Viz. display had many elements in common with the Baseline, including the same interface to display alerts, data, and methods to take action; the only difference between these two displays was the visualization of the data. These differences in familiarity may have contributed to some of the differences seen in performance with the VR display. No differences between VR and traditional training on subsequent performance are consistent with findings in literature [196, 197], which have found VR to often be as good as, but not better, for promoting performance. However, previous research has not considered 3D visualizations as an in-between.

Finding no differences in workload is also a positive result and matches the hypothesis. As has been discussed in Aims 2 and 3, any increase in workload would be indicative of a poor training experience that left participants unprepared to understand the Baseline display without increased effort. Supervisory control and monitoring tasks are typically lower workload [66, 175, 176] to begin with, and it was not expected that any one training modality could reduce this further.

The results in support of using VR in training also come from subjective participant assessments. Participants trained in VR gave the Baseline display higher usability scores compared to those trained in the Baseline display, contrary to the hypothesis. This indicates that the prior context developed by people trained in VR enabled them to synthesize and understand the traditional display more easily. How training modality influences the usability of a traditional display has not previously been studied, and represents an important finding of this study. The utility questionnaires and comments further highlight the potential of VR or 3D visualizations for training and help explain some of the differences in other metrics. These results also support the idea that VR and visualizations may be improving mental models of the scenario. As previously mentioned, future satellite operations may be more complex and require quicker decisions to avoid collisions and have harder to understand orbits [198]. Being able to promote understanding of aspects like collision likelihood or orbital motion, which VR did, is critical to improving operators' understanding of the scenario. These are reflected in the comments about the importance of understanding the 3D orbits before transitioning to operating on the Baseline display. In sum, these results indicate that those who trained in VR and Scr. Viz. perceived them as being very useful for understanding the task better in a traditional environment. These perceptions may also be indicative of the inferred improvements in mental models achieved through training in VR. This finding agrees with previous training studies where participants subjectively prefer VR [199, 200].

There are some limitations to the approach taken for this study. The scenario and task used are both sped up and simplified compared to actual operations. This simplification may influence some of the results; for example, in SA level 1 and the easy trials scores are generally high regardless of modality, and no differences are found. Additionally, training took place over one day instead of the typical 3 months to a year that is required for complex operations. Despite this, the results show that VR is able to provide benefits from just one day of training, highlighting its potential. Furthermore, training sessions were not evenly spaced between participants, although they were bounded between 1 and 8 days. No trends were observed between the duration between training sessions and any of the metrics, indicating these differences did not impact the results. Future work should increase the complexity of these trials and include repeated training sessions or more consistent timing between sessions to further understand how VR can provide an impact. While the speed of the trials may influence aspects like boredom, it was consistent between the two visits, allowing us to compare metrics across trials to understand differences based on training modalities.

7.5 Summary and Contributions

This aim compares three training displays, including an immersive VR display, a 3D screen visualization, and a representative traditional display for satellite operations. The results of this work indicate that VR is a promising training mechanism for satellite operations as it improves level 2 SA and usability in traditional displays. No differences were found in the other SA levels and workload. Training in VR also has a higher perceived subjective utility towards understanding critical aspects of satellite operations. VR training may be able to promote improved mental models, safer operations, and improved understanding of traditional display, and future work should continue to assess this and VR's potential for use in training. Additional work is also needed to determine how to incorporate it into training plans effectively. These conclusions can inform training for satellite operations, as well as other supervisory control scenarios like robotics, manufacturing, and transportation.

The main contribution to the literature of this aim is understanding how VR can be useful for supervisory control training, as opposed to the more commonly studied manual control training tasks.

Chapter 8: Aim 5: Supervisory Displays and Trust

8.1 Introduction

The objective of this final aim is to understand how an operator uses the information on the display to make decisions, and how these decisions are related to trust. Unlike the previous aims, this aim focuses on the teaming aspect of operations, where the operator works with an autonomous agent to identify objects in satellite images and does not rely on VR. Instead, 2D visualizations are included as a way to convey information to verify an autonomous system.

This aim seeks to answer three research questions: 1) What information do participants use that makes them more accurate in their decisions? 2) What aspects of gaze do participants exhibit when reviewing information that make their decisions more accurate? 3) What behaviors (i.e., actions or gaze) are related to a participant's trust in the autonomous system?

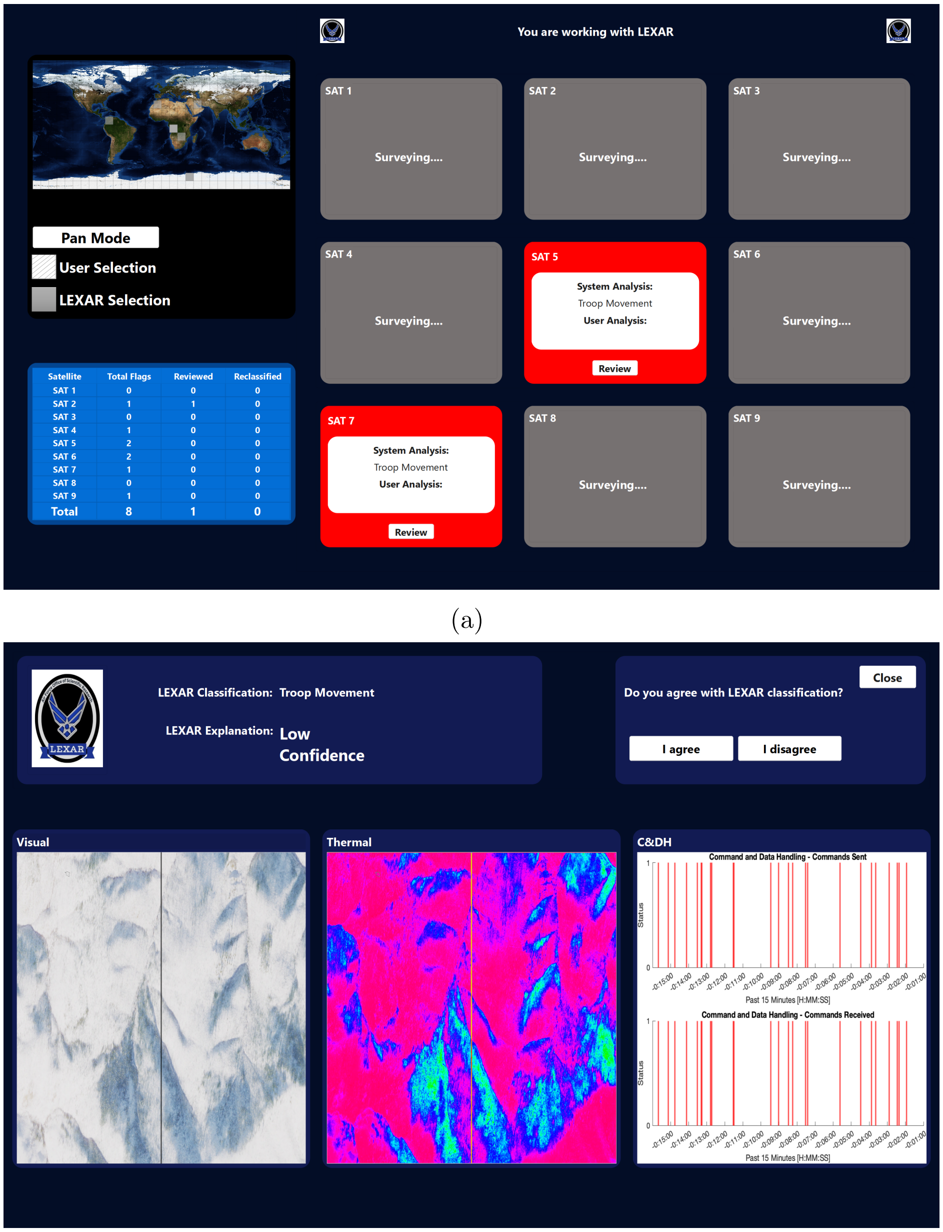
The culmination of these research questions will allow for the understanding of what display components are being used as operators team with an autonomous system to make decisions, and how they view information influences their trust in the autonomous system. This leads to discussions on how to better design displays to promote accuracy and calibrated trust.

8.2 Methods

8.2.1 Task and Display Design

To study decision making and trust in autonomous systems, an operationally relevant human-on-the-loop task was designed, where participants work with an autonomous system to classify satellite data as containing ground troop movement or no ground troop movement. The complete details of the task have been described previously by Sung et al. [201], and the relevant portions are summarized here.

The participants work with nine different ground imaging satellites. Every 30 seconds, two of the satellites flag as having updated information, and the autonomous system classifies these as containing troop movement or no troop movement, as seen in Fig. 8.1a. Participants have the



Satellite	Total Flags	Reviewed	Reclassified
SAT 1	0	0	0
SAT 2	1	1	0
SAT 3	0	0	0
SAT 4	1	0	0
SAT 5	2	0	0
SAT 6	2	0	0
SAT 7	1	0	0
SAT 8	0	0	0
SAT 9	1	0	0
Total	8	1	0

(a)

LEXAR Classification: Troop Movement

LEXAR Explanation: Low Confidence

(b)

Figure 8.1: The user interface the participant used. (a) The home screen, exhibiting 2 satellites which the system has flagged as having been reviewed by the autonomous system. The operator has the option to review or not review the decision as they choose. (b) The review screen, with the visual, thermal, and the command and data handling (C&DH) screens.

option of reviewing the data, but are not required to. Additionally, participants are also tasked with suggesting regions on the globe to image next. This is done to require participants to strategically allocate attentional resources and to encourage less monitoring of the system if they feel it is appropriate.

If the participant opts to review the data, they would see the screen in Fig. 8.1b. This screen contains the autonomous system's recommendation and the system's confidence level in that recommendation. The autonomous system is not always reliable, and the recommended classification is not always correct. On the review screen, the participant can see three pieces of information, or data streams, that can be used to verify the accuracy of the system's recommendation. If the classification is incorrect, a conflict would be present in one, two, or all of the data screens that the participant has access to. There is a visual and a thermal image of the terrain that can be used to assess the autonomous system's accuracy. These images show satellite images of the terrain in either the visual spectrum or with thermal coloring. Troop movements are represented by a singular tank object superimposed into the image. For both these images, a conflict is defined as showing the opposite of the recommendation (e.g., showing a troop in the image when the autonomous system made the determination that there was no troop movement present). Finally, a command and data handling image (C&DH) is included to verify the health of the autonomous system and if it received all the information from the satellite. When the C&DH shows dropped signals, as represented by a mismatch in the vertical lines in the top and bottom sections of this telemetry stream, not all information was transmitted. Participants are instructed that if signals are dropped, the autonomous system is working with incomplete data, and thus, may not be reliable.

There are four different autonomous systems that the participants work with, each having varying reliability and explainability. These are designed to manipulate their trust. Two have low reliability (67.4%), and two had high reliability (83.8%). Additionally, two of these systems have low explainability (where the explanation uses terse, robotic-like language) and two have high explainability (where naturalistic language is used).

8.2.2 Experimental Design

Twelve participants (5 female, 7 males, ages 19-43, median age 23.5 years) completed the study. The study was approved by the University of Colorado Institutional Review Board (protocol # 23-0103). The experiment consisted of 1 training session and 4 testing sessions, where each testing session involved a unique autonomous system.

The training session began with operator background surveys designed to capture individual differences that may influence their trust [202]. These include the “High Expectations” component of the Perfect Automation Schema (PAS) [203], the Automation Induced Complacency Potential (AICP) Scale [204], the Propensity to Trust survey (PT) [205], the “Extraversion” and “Agreeableness” sections of the Big Five Factors of Personality survey [206], the “Masculinity” dimension of the Cultural Values Scale (CVS) [207] and the “Performance Expectancy” as well as “Effort Expectancy” sections of the Unified Theory of Acceptance and Use of Technology (UTAUT) survey [208]. Additionally, a demographics questionnaire was administered that captured age, sex, race, ethnicity, dominant hand, experience with video games, experience with robotic systems, navigational aid use, experience with aerospace-relevant displays, and experience with military monitoring systems.

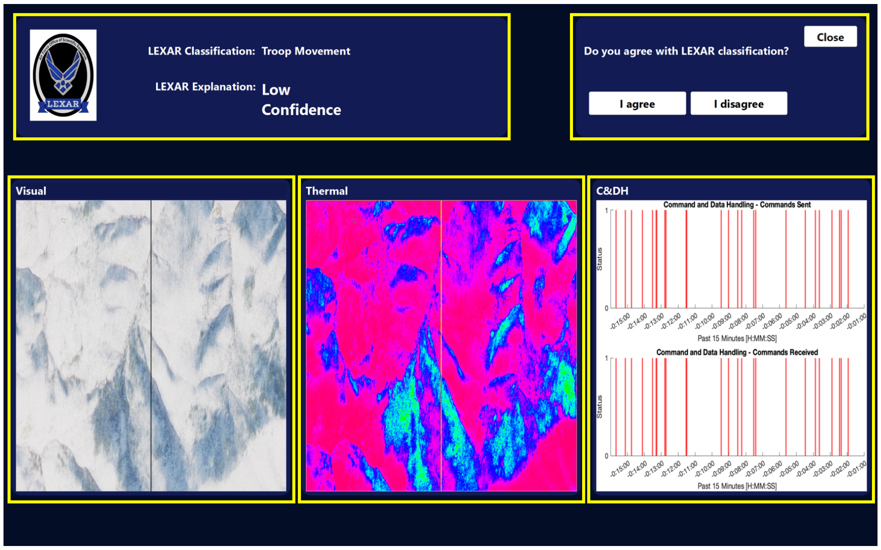
Next, the participants were trained using a PowerPoint slideshow that they could read through at their own pace. This contained relevant information about the task, how to complete it, and how to rate their trust. After training, they completed a quiz to ensure understanding. Finally, participants performed three practice trials with a simulated 100% accurate autonomous system to practice using the display and working with the autonomous system, such as verifying classifications.

During each of the four testing sessions, participants worked with one of the four autonomous systems in a randomized order. Participants were informed that it was a new system each session and that they should not let their feelings and attitude toward a previous system influence their trust or decisions with the new system. When they arrived for the experiment, they completed a brief survey capturing the amount of sleep they had the night before and completed the psychomotor vigilance test (PVT) [179]. They then completed 6 trials. During each trial, 24 satellites were flagged, 2 every 30 seconds, for possible review. Every 45 seconds (8 times per trial), the task paused, and a screen appeared for participants to use a slider to rate their agreement with the statement “I trust this autonomous system” on a continuous scale from “Not at all” to “Completely". After their rating, the trial continued. At the end of a trial, participants filled out the Trust in Autonomous Systems survey [209]. In addition to this, participants wore neurological and physiological sensors, which are not included as part of this data analysis. Participants were paid for their time and could earn bonuses for performance, which included both accuracy in classification and the degree to which they assisted the autonomous agent in identifying areas to take images. Participants received feedback on the team's performance at the end of every trial.

8.2.3 Metrics

In addition to the background surveys described previously, additional metrics were collected that were analyzed in this aim. Observable information about the system and environment were recorded, such as the explainability and reliability of the autonomous system, which satellite was flagged and the time it was flagged, the system's recommendation and whether or not it was correct for that particular satellite, and if inaccurate, what and how many of the 3 data streams were in conflict with the recommendation. Additionally, information about the participant's direct actions was collected. This included whether they reviewed the satellite, the duration and frequency of their reviews, and whether they accepted or rejected the autonomous system's assessment. Gaze metrics were also collected. Areas of interest (AOI) for the gaze metrics are highlighted in Fig. 8.2 and include the visual image, thermal image, C&DH screen, systems analysis (system recommendation and confidence), and buttons (selection of agree or disagree). Gaze metrics for each AOI include how long they were looking at it, the number of times they looked at it, and the number of transitions between each combination of AOIs (e.g., looking first at the visual image and then the thermal image). Beyond this, the coordinates of their gaze can be used for a recurrence quantification analysis (RQA) [210]. RQA is a way to describe dynamical systems, and characterized fixation sequences and discovered repeated scan patterns and fixation locations, which may be related to their decision; the full definition of the RQA terms is in Appendix D.

Additionally, the trust slider values are recorded on a scale from 0 to 1. The differences in subsequent trust sliders is used to capture trust dynamics, or how the trust changes between



LEXAR Classification: Troop Movement

LEXAR Explanation: Low Confidence

Do you agree with LEXAR classification?

Figure 8.2: The review screen, with the visual, thermal, and the command and data handling (C&DH) screens. The yellow boxes indicate the AOIs used for gaze tracking and are not present on the display that the participants reviewed. This includes the system recommendation AOI (top left), the buttons AOI (top right), the visual AOI (bottom left), the thermal AOI (bottom middle), and C&DH AOI(bottom right).

epochs. Including both trust values and trust dynamics is important to capture if calibrated trust is achieved and to understand events that cause trust to change. Within each 45 second epoch, or time between trust slider reports, additional metrics can be computed. This includes the number of satellites that were flagged, the number that were classified during that epoch, the number of satellites reviewed by the participant, how many recommendations they agreed with, how many they rejected, and how many were passively agreed upon (i.e., not reviewed).

8.2.4 Statistical Analysis

In total, 6912 satellites were classified during all testing sessions. Out of these, 6527 satellites were reviewed by the participants, and among them, 1517 had a conflict present (i.e., where the autonomous system was incorrect). In addition, 2304 trust sliders and epochs worth of data were collected.

To understand research question 1, if the participant's accuracy and duration spent reviewing the data streams were influenced by the number of data streams in conflict, several analyses were conducted. For participant accuracy, a generalized linear mixed effects model was created (using *lmer4* in R [167]) to compare the number of conflicts, the autonomous system recommendation, and their interaction effects on accuracy. The participant's score on the PAS was included as a random effect to account for individual differences. To assess significance, an ANOVA was run, followed by simple effect pairwise comparisons using estimated marginal means with a Tukey correction (*emmeans* package [168]). Likewise, for duration, a linear mixed effects model was generated to compare the duration of review based on the number of data streams in conflict, the system recommendation, and the participant's accuracy. In order to meet the assumptions of the model, the duration was log-transformed. To assess significance, an ANOVA was run, followed by estimated marginal means pairwise comparisons.

Additionally, to understand the relative use of a particular data stream and if, when it was in conflict with the recommendation, it affected their accuracy and review duration, a similar analysis as with the number of data streams in conflict was conducted. Instead of using the number of data streams in conflict, the analysis was restricted to times when there was only 1 conflict (i.e., only a single data stream was in conflict) so an analysis of *which* data stream and how participants used it was performed. The analysis of accuracy did not meet the assumptions required for the ANOVA, and instead, a χ^2 test was used. Unlike previously, interactions are not considered due to the limitations of this test. Post-hoc pairwise comparisons were conducted using *prop.test* with a Bonferroni correction. The review duration was analyzed with a linear mixed effects model.

For research question 2, to understand how a participant's behavior differs based on decision accuracy, mutual information was calculated (*sklearn* in Python) between each of the gaze variables and accuracy. This allows for the understanding of what factors differ based on the decision accuracy. Due to the findings from the previous research question, which identified an interaction with the system recommendation, the mutual information was calculated for three separate data sets. This included all reviewed satellites, when the recommendation was troop movements, and when the recommendation was no troop movement. This also allows for an understanding of how people behave differently based on the recommendation.

To identify if gaze is useful for predicting accuracy, a random forest was generated to classify the data as correct or incorrect, using the same 3 data subsets (all, recommended troop, recommended no troop, using *sklearn* in Python with an 80/20 test split). All the variables with non-zero mutual information are included as predictors available to the random forest, as a way of doing feature downselection. From this, the top 10 variables with the highest feature importance were considered. Comparisons were made between the three different models to see what participants are doing differently when they are performing well and when the autonomous system recommendation differs.

For the 3rd research question, an analysis to investigate if there is a relationship between the user's decisions and trust. For each epoch, the total number of times the participant actively reviewed and agreed with the system, disagreed with the system, and never reviewed the system was calculated. For each of these metrics, a repeated measure correlation was run between the trust slider value and the trust dynamics. This was done using *rmcorr* in R [211]. Since each individual may trust each autonomous system differently, groups were defined based on both participants and the autonomous systems with which they worked, resulting in 48 groups in the repeated measures.

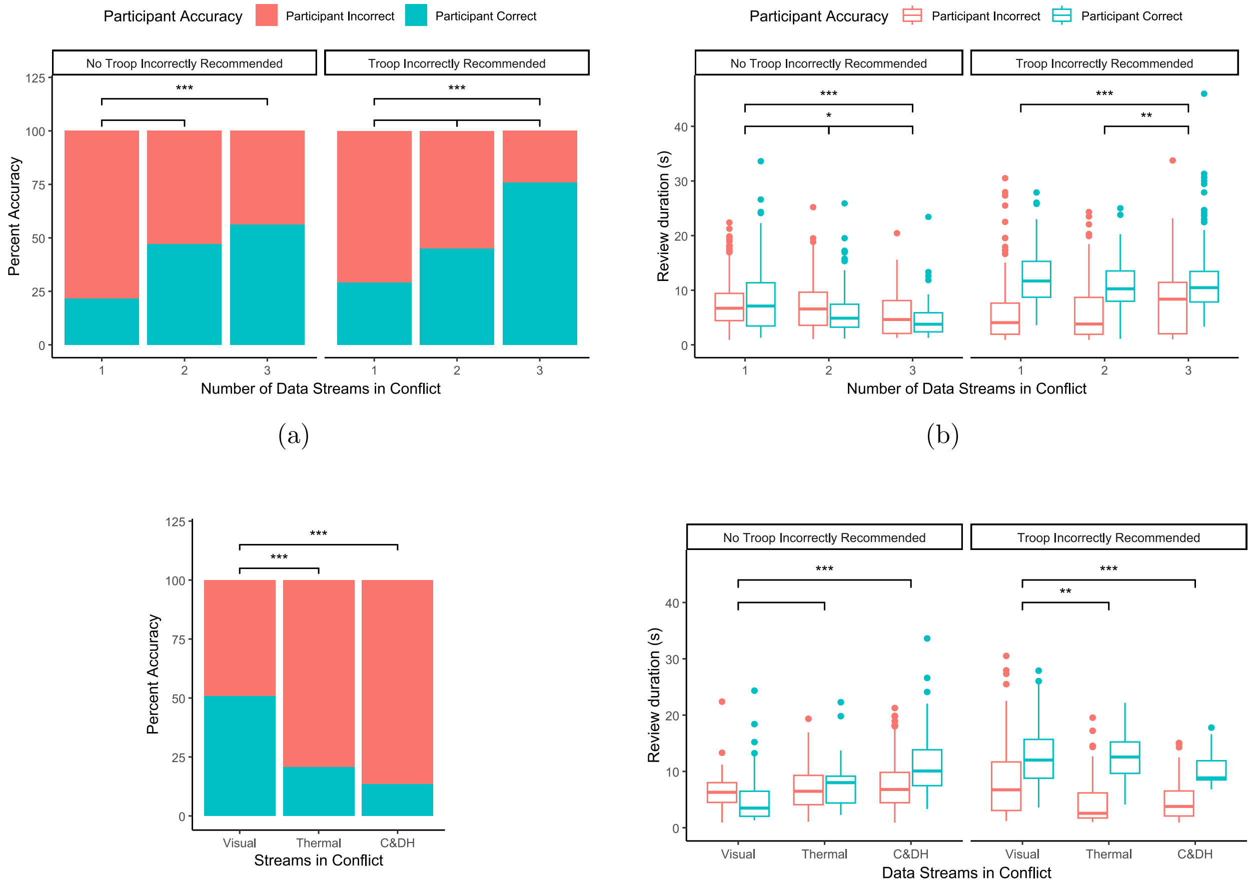
Finally, to understand if gaze and trust are related, a similar process was used to answer this question as was done for understanding how gaze is related to accuracy. A random forest regression was run with the variables containing non-zero mutual information. The top 10 variables by feature importance were kept. This was only run on all the data, as system recommendations should not influence trust. Since trust was only recorded after each epoch, the trust value for each data point was the trust recorded for the previous epoch.

8.3 Results

Overall, the participants were correct 78.8% of the time, which is higher than the 75.6% accuracy of the system alone. For research question 1, to understand the impact of the number and type of conflicts on accuracy and duration, the results comparing the number of conflicts to participant accuracy and review duration can be seen in Figs. 8.3a and 8.3b. A significant interaction between the autonomous system recommendation and number of conflicts is found ($\chi^2(2)$ = 9.38, p =0.009). As the goal is to understand how the number of conflicts influences accuracy, a post-hoc simple effect analysis is conducted with a Tukey correction factor. When the system incorrectly recommends no troop movements, differences in accuracy are found between 1 and 2 conflicts (z = 6.08, $p < 0.005$) and between 1 and 3 conflicts (z = 6.38, $p < 0.005$). No differences are found in accuracy between 2 and 3 conflicts (z = 1.64, p = 0.23). When the system incorrectly recommends troop movements, all three pairwise comparisons are significant; between 1 and 2 conflicts (z = 4.03, $p < 0.005$), 1 and 3 conflicts (z = 10.78, $p < 0.005$) and 2 and 3 conflicts (z = 7.17, $p < 0.005$).

For the duration spent viewing the recommendation screen (Fig. 8.3b), the ANOVA comparing the number of conflicts, participant accuracy, and system recommendation resulted in no three-way interaction ($\chi^2(2)$ = 1.73, p = 0.41), but significant two-way interactions are found. There are interactions between the number of conflicts and the system recommendation ($\chi^2(2)$ = 34.79, $p < 0.005$), the number of conflicts and accuracy ($\chi^2(2)$ = 13.75, $p < 0.005$), and between accuracy and system recommendation ($\chi^2(1)$ = 165.5, $p < 0.005$). Of particular interest is the interaction between accuracy and system recommendation. In both recommendation cases, there is a significant difference in the post-hoc simple effect analysis comparing participant accuracy. However, the directionality is different between when the system incorrectly recommended no troop movement (z = -4.35, $p < 0.005$, when the participant is correct they spend less time reviewing) and when it recommended troop movement (z = 16.08, $p < 0.005$, when the participant is correct they spend more time reviewing). Furthermore, in the post-hoc analysis of the interaction between the number of conflicts and the system recommendation, all pairwise comparisons are significant when there is no troop movement recommended (1 to 2 (z= 2.69, p = 0.02), 1 to 3 (z = 4.63, $p < 0.005$), and 2 to 3 (z = 2.40, p = 0.04). When troop movement is recommended, the pairwise comparison between 1 and 3 conflicts (z = -3.75, $p < 0.005$) and 2 to 3 (z = -4.12, $p < 0.005$) are significant. The comparison between 1 and 2 conflicts is not significant (z = 0.4, p = 0.91).

Looking specifically at the differences between what screen is in conflict when there is only



(a)

(b)

(c)

(d)

Figure 8.3: (a) The percent accuracy compared to the number of data streams in conflicts, split by system recommendation. The accuracy increases with an increasing number of conflicts. (b) The review duration compared to the number of data streams in conflicts, split by system recommendation and by accuracy. (c) The percent accuracy compared to the type of data stream in conflict. Due to the limits of the statistical analysis, this is not split by accuracy (d) The review duration compared to the type of data stream in conflicts, split by system recommendation, and by accuracy.

one conflict, as this informs how participants use the different pieces of data, significant differences between conflict type and accuracy ($\chi^2(2)$=94.4, $p < 0.005$) are found. Pairwise comparisons found differences between visual and thermal conflicts ($p< 0.005$), and visual and C&DH ($p< 0.005$). No differences are found between thermal and C&DH ($p = 0.13$). These results can be seen in Fig. 8.3c.

Like with the number of conflicts, the ANOVA comparing the duration spent viewing the recommendation screen based on conflict type (Fig. 8.3d) has no significant three-way interaction

($\chi^2(2)$ = 1.63, p = 0.45); however, significant two-way interactions existed. Significant interactions are found between conflict type and recommendation ($\chi^2(2)$ = 47.00, $p < 0.005$), conflict type and accuracy ($\chi^2(2)$ = 26.00, $p < 0.005$), and between accuracy and recommendation ($\chi^2(1)$ = 48.85, $p < 0.005$). In contrast with the earlier findings, a difference in the pairwise comparisons between participant accuracy and system recommendations is significant when the system recommends troop movement (z = -9.27, $p < 0.005$). For no troop movement recommendation, the pairwise comparison for accuracy is not significant (z = -0.13, p = 0.89). Additionally, in the post-hoc pairwise comparisons for conflict type and recommendation, when no troop movement is recommended, there are significant differences in duration between visual and thermal (z = -3.98, $p < 0.005$), visual and C&DH (z = -6.5, $p < 0.005$) and no differences in thermal and C&DH (z = -1.662, p = 0.25). A similar trend is observed when the troop is recommended. Significant differences are seen for visual and thermal (z = 3.32, $p < 0.005$) and visual and C&DH (z = 2.81, p = 0.01) comparisons, and no differences in thermal and C&DH (z = -0.45, p = 0.89).

For research question 2, Tab. 8.2 shows the top 10 gaze metrics, by feature importance, that are used in the random forest classification model to distinguish between correct and incorrect decisions. As it is clear from the previous analysis that people behave differently based on the recommendation of the autonomous system, this analysis is over all data to identify gross behavioral differences in accuracy, and then also split by recommendation to determine behavioral differences when the recommendation differs that influence accuracy. Table 8.1 shows the performance metrics for these models. While the accuracy is high, the F1 score, which accounts for false positives and false negatives, is relatively poor. These models are intended to understand what features are most important, not to have high predictive performance. The features selected have high importance out of all the features.

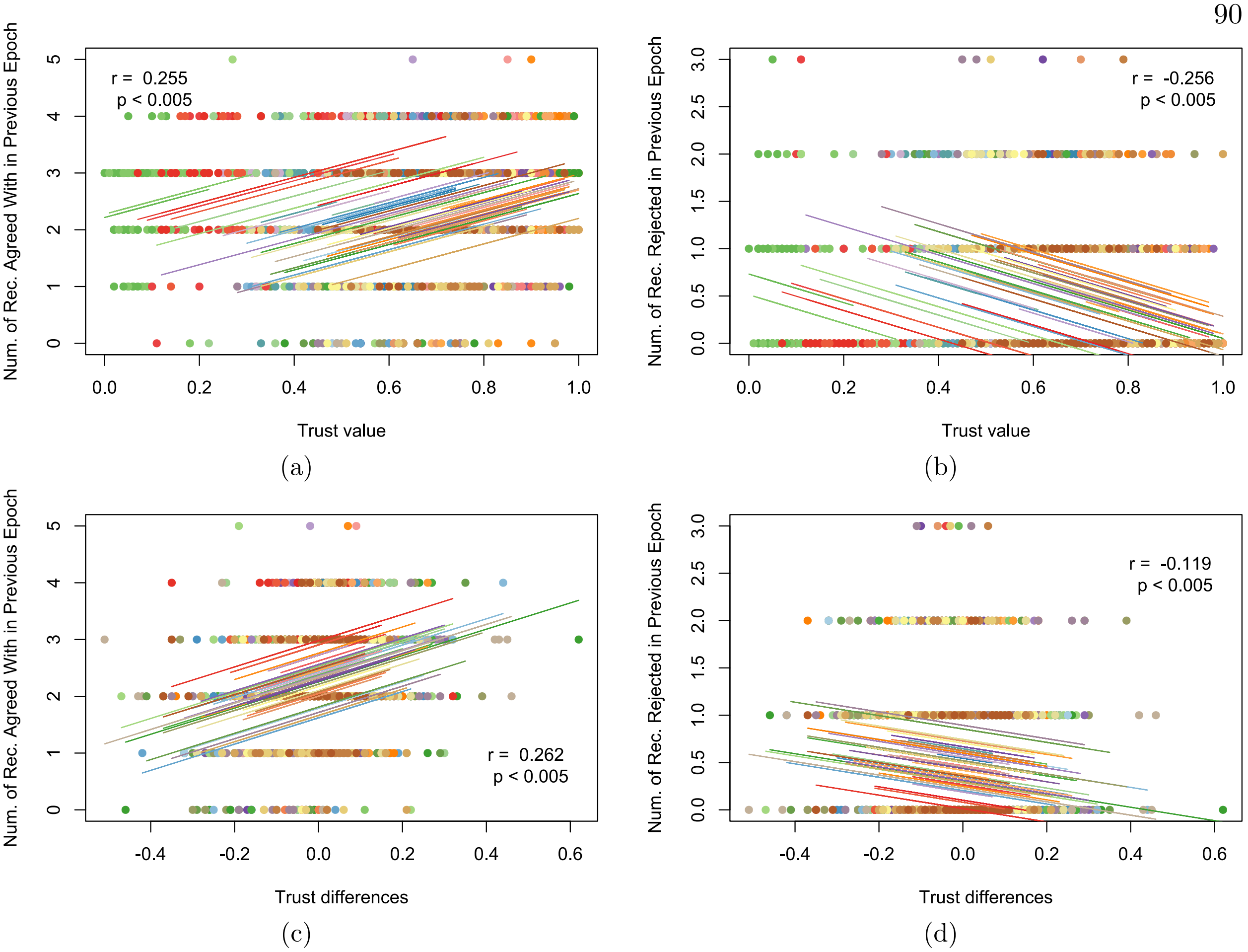
Table 8.1: Random Forest Model Performance

Metric	All data	Recommend Troop	Recommend No Troop
F1-Score	0.45	0.45	0.48
Accuracy	0.80	0.78	0.82

Table 8.2: Top 10 gaze metrics by feature importance for decision accuracy. *Sys.Rec. denotes an interaction with that feature and the system recommendation being important.

	Rank	All data	Recommend Troop	Recommend No Troop
Mutual Information	1	Dur. * Sys.Rec.	Dur.	Total Dur. on Buttons AOI
	2	Entropy	Entropy	Switches from Sys.Rec. to Visual AOI
	3	Dur.	Relative Entropy	Switches from Viz. to Sys.Rec. AOI
	4	Relative Entropy	Total Switches Between AOIs	Percent Recurrence on AOI
	5	Num. Recurrence	Laminarity	Num. Reviews
	6	Det. with AOI	Total Dur. on Viz. AOI	Switches from Sys.Rec. to Th. AOI
	7	Laminarity*Aut.Sys.Rec	Num. recurrence	Relative Entropy
	8	Entropy*Sys.Rec	Percent Recurrence on AOI	Switches from C&DH to Th. AOI
	9	Relative entropy*Sys.Rec	Num. Fixations of Viz. AOI	Switches from Viz. to Th. to C&DH AOI
	10	Det. with AOI*Aut.Sys.Rec	Num. Recurrence on AOI	Switches from Sys.Rec. to Button AOI
Random Forest Classifier	1	Dur.	Dur.	Num. Reviews
	2	Dur. * Sys.Rec.	Num. Recurrence	Dur.
	3	Num. Recurrence	Relative Entropy	Total Dur. on Visual AOI
	4	Relative Entropy	Total Dur. on Visual AOI	Relative Entropy
	5	Total Dur. on Viz. AOI	Entropy	Total Switches Between AOIs
	6	Entropy	Total Dur. on Th. AOI	Num. recurrence
	7	Total Dur. on Viz. AOI*Sys.Rec	Num. Fixations of Viz. AOI	Total Dur. on Th. AOI
	8	Entropy*Sys.Rec	Total Switches Between AOIs	Total Dur. on Buttons AOI
	9	Num. Recurrence*SysRec	Total Dur. on C&DH AOI	Percent of Time on Viz. AOI
	10	Total Switches Between AOIs	Total Dur. on Sys.Rec. AOI	Entropy

Note: AOI = area of interest; Viz = visual AOI; Th = Thermal AOI; C&DH = Command and Data Handling; Sys. Rec. = autonomous system recommendation; Dur = Duration; num = Number of; Det = determinism



(a) (b) (c) (d)

Figure 8.4: Repeated measure correlation plots between (a) trust value and the number of times the participant agreed during the previous epoch (b) trust value and the number of times the participant rejected the recommendation during the previous epoch. (c) trust dynamics and the number of times the participant agreed during the previous epoch (d) trust dynamics and the number of times the participant rejected the recommendation during the previous epoch. For all plots each participant and session has their own intercept, but the slopes are consistent with the correlation found. Correlations and p-values are labeled.

Finally, for research question 3, to understand how behaviors are related to trust, significant, but weak, correlations are found between some of the participants' actions and trust values as seen in Figs. 8.4a and 8.4b and Tab. 8.3. A correlation of r = 0.25 (p <0.005) is found between the number of recommendations the participants agreed with and their trust. For the number of recommendations they rejected, a similar correlation is found of r = -0.26 (p < 0.005). A negligible, but significant, correlation of r = -0.09 (p < 0.005) is found between their trust and the number of recommendations they did not review.

Table 8.3: Correlations between participants' outcomes and their trust.

User Decision	Trust Value (r)	Trust Dynamics (r)
Recommendation Accepted	0.25***	0.26***
Recommendation Rejected	-0.25***	-0.12***

For the trust dynamics, considering the number of recommendations agreed with, the correlation is consistent with previous raw trust values with a correlation of r = 0.26 ($p < 0.005$, Fig. 8.4c). However, for the number of recommendations they rejected, the correlation is weaker at r = -0.12 ($p < 0.005$, Fig. 8.4d). Similar to before, a negligible correlation of r = -0.09 ($p < 0.005$) is found between the trust dynamics and the number of recommendations they did not review.

Finally, Tab. 8.5 shows the top 10 gaze metrics by mutual information and by the random forest regression for trust and trust dynamics. Table 8.4 shows the regression metrics for the prediction of trust and trust dynamics. Regression accuracy is a much harder problem than classification accuracy, and while these models do not demonstrate strong performance, the goal of fitting these models is to elucidate which factors are most predictive of differences in performance. However, given the model's performance, the predictor variables might not be strong predictors in understanding how gaze behavior relates to trust.

8.4 Discussion

The objective of this aim was to analyze how participants' behaviors in viewing information on screens while teaming with an autonomous system lead to differences in accuracy and trust. Analyzing participants' accuracy across different types of conflicts can provide insights into which portions of the screen they are actively checking and which are not used in determining whether a participant agrees with the autonomous system. In general, participant accuracy increases with the number of pieces of conflicting information, confirming the hypothesis. Participants are more likely to overturn the system when more pieces of information contradict the system's recommendation. However, the interaction between the system's recommendation and the number of data streams in conflict, as well as the difference in accuracy with the recommendation, was not hypothesized and

Table 8.4: Random Forest Regression Performance

Metric	Trust Value	Trust Dynamics
MSE	0.032	0.01
R^2	0.28	0.05

may represent differences in behavior based on the recommendation.

The interaction between accuracy and system recommendation is of particular interest when analyzing the review duration. When the system incorrectly recommends troop movement, participants take longer to review in order to make a correct decision. However, when the autonomous system incorrectly recommends no troop movement, participants who are correct take less time to review. It is hypothesized that the differences are primarily due to the nature of the task, where determining the accuracy of the autonomous system is primarily a visual search, particularly for the visual and thermal images. When the system recommends no troop movement but in reality troops are present, people who identify the troop are able to stop as soon as they see the first troop. If they do not spot the error, they may spend more time looking and be ultimately incorrect. On the other hand, when the system recommends a troop and there is none, people who take longer are able to verify that there is no troop present and are more likely to be correct. This difference is noteworthy because it indicates that participant behavior varies depending on the system's recommendation. Previous research suggests that people may respond differently to false alarms and missed detections [46, 212]. In this case, recommending a troop when there is none is analogous to a false alarm, and recommending no troop when there is one is comparable to missed detection. This difference in use based on recommendation has implications for future display designs, as it may mean that display designs should be modified or adjusted based on the recommendation to promote appropriate use and decision making.

The analysis of which data streams participants primarily use further supports these results. Analyzing the data with one conflict provides some insight into the behavior of the participants with regard to each data stream independently. Participants are most accurate when the visual image, which is on the left of the display, is in conflict with the recommendation of the autonomous

Table 8.5: Top 10 gaze metrics by feature importance for trust.

	Rank	Trust value	Trust dynamics
Mutual Information	1	Duration of Time on Buttons AOI	Number of Fixations of Visual AOI
	2	Percent of Time on C&DH AOI	Number of Fixations of C&DH AOI
	3	Total Switches Between AOIs	Percent of Time on Thermal AOI
	4	Duration of Time on Thermal AOI	Percent of Time on Visual AOI
	5	Duration of Time on C&DH AOI	Switches from Buttons to Thermal AOI
	6	Percent of Time on Thermal AOI	Duration of Time on Thermal AOI
	7	Number of Fixations of C&DH AOI	Percent Recurrence on AOI
	8	Number of Fixations of Thermal AOI	Duration of Time on C&DH AOI
	9	Number of Fixations of Buttons AOI	Switches from Visual to C&DH AOI
	10	Duration of Time on Sys.Rec AOI	Number of Fixations of Sys. Rec. AOI
Random Forest	1	Duration of Time on Buttons AOI	Duration
	2	Number of Fixations of Thermal AOI	Percent of Time on Visual AOI
	3	Number of Switches from Buttons to Visual	Center of Recurrence Mass
	4	System Explainability	Number of Fixations of Thermal AOI
	5	Number of Fixations of Buttons AOI	Maximum line length in RQA plot
	6	Total Switches Between AOIs	Percent Recurrence
	7	Percent of Time on Thermal AOI	Number of recurrence
	8	System Confidence	Relative Entropy
	9	Duration	Determinism
	10	Percent of time on C&DH AOI	cluster

system. Inspection of the gaze data revealed that participants commonly spent the most time on this leftmost visual screen, and often started their scan of the display there. Participants typically reviewed all screens to some extent, just with shorter durations. This may be a result of participants' cultural background; in Western English reading cultures, scan patterns are generally from left to right [213]. This may explain the increased accuracy and duration on the visual screen in particular. Participants were not trained to have a particular scan pattern, but were trained to know that all information was important in making their determination. There are a few potential hypothesized explanations for why they did not evenly split their time. Given that the visual and thermal screens present similar types of information, participants may not have felt the need to thoroughly review both sources, instead directing their attention to the information they felt most relevant, and spending less time on the source they felt was less relevant [214, 215]. Additionally, they may have felt some time pressure during the task to make a decision quickly and did not further check the rest of the data [216] if the visual screen matched their expectations, particularly if they had a high level of trust in the autonomous system. Finally, this could indicate that participants felt more comfortable and understood how to process visual information as opposed to the other screens, despite initial training. This is supported by participants spending the least amount of time on the C&DH screen, which is also the most novel. Future work should investigate if these differences in accuracy are due to a natural scan pattern, familiarity, or perceived importance.

The random forest models provide further granularity into scan patterns and how differences in gaze behavior enable differences in participant accuracy. Feature importance was focused on to identify key behaviors as to how participants were using the display differently. Other features that are not listed here may still be required to do well. However, if all participants already do those behaviors it will not show up in the selection. There are some variables that stand out related to their gaze metrics, including duration and entropy metrics. Fixation duration has been shown to be linked with cognitive deliberation, where an increase in the number of fixations and fixation durations is related to either an attempt to gather more detailed information [217], or more difficulty extracting relevant information [218]. Participants who reviewed longer did better.

Likewise, entropy represents the uncertainty, or randomness, in eye movements; higher entropy may indicate broader visual search behavior or be representative of checking over the entire images and reviewing all the data sources thoroughly. Participants who had higher entropy, on average, did better, which is opposite of previous literature. Increased entropy has been shown to be correlated with reduced situation awareness and poor performance in target detection [219]. However, this only considered AOI based entropy and not entropy over the entire screen. It may also be the case that there is a non-linear relationship between entropy and accuracy which is unable to be captured with the modeling techniques used. Additionally, variables such as the review duration, entropy, the number of recurrences, and duration spent on the visual screen all interact with the system's recommendation, reinforcing that there is a difference in how participants are behaving based on the system's recommendation. For all these measures, the difference between correct and incorrect responses is more pronounced when the system recommends troop movement. The only metric related to specific AOIs is the time spent on the visual data stream AOI, which reinforces the fact that participants are primarily relying on it to make their decision. Additionally, larger entropy values may indicate that either the participant is unable to find the troop in the image when one is recommended or is engaging in a more thorough search when no troop is recommended. While these models have high accuracy, they have low F1 scores. This is likely due to the overall imbalance of data points, as most of the time, participants are accurate and there are no conflicts. Thus, it is important to acknowledge that the goal of this analysis was not to build a highly predictive model, but rather use the model building process to identify key behaviors that lead to improved accuracy.

Combined with the above findings, these gaze metrics suggest important implications for display design [220,221]. The visual AOI and the time spent reviewing it seem to play an important role in participants' decisions. In similar displays, it may be important to place the key information where participants initiate their scan pattern; in this situation, it would be on the left side. Under time pressure, this would allow the most critical information to be encountered by the participant first. Similar methodology using gaze tracking may be useful towards understanding scan patterns for a particular display [106]. Alternatively, or in conjunction, training can be implemented to ensure participants have consistent and appropriate scan patterns for the task, or to ensure that they are taking the actions that will allow their accuracy to increase (i.e., longer verification time). This also suggests that for the level of reliability of the system, participants are actively choosing to verify and not rely only on the system recommendation so spending the resources to include verification data may be important.

Considering the relationship between the participants' actions, gaze behavior, and trust dynamics for research question 3, weak correlations are found between participants' actions and both their trust levels and trust dynamics. These results are consistent with previous literature that suggests that behavioral metrics may not be reliable proxies for trust alone [46, 50, 51]. However, these types of behavioral metrics have been used in conjunction with other performance or background metrics to accurately model trust, signaling that these weak correlations can still provide benefits [222, 223]. It was hypothesized that trust should increase with increasing recommendations they agreed with, and decrease with an increasing number of rejections. The directions of the correlations follow this expectations. Some potential reasons for the weak correlations, and the notably weaker correlation with trust dynamics and rejecting the recommendations may be due to bottoming out of the trust scale or due to achieving calibrated trust, where trust changes may be minimal. However, visual inspection of the trust dynamics over time did not show either of these to be the case. Rather, these differences may be due to the reason why the participant rejected the system. Participants were told that the C&DH was a way to see if the autonomous system received all the data. If the participants determined that missing C&DH data was the reason to reject the recommendation, they still might trust the system to make the correct decision when it has complete data. This might strongly affect the correlation between rejecting the recommendation, explaining the weaker correlation here. However, it is unable to be determined the reason for rejection, and future work should further study this to understand if the behavioral metrics of trust may also be influenced by factors outside the autonomous system's control. Additionally, a limitation is that these results may be influenced by using a repeated measures correlation as the data is not continuous.

Notably, there is no correlation between instances where participants did not verify the recommendation and trust or trust dynamics. It was hypothesized that if participants did not review the data, the trust dynamics should not be affected, which is supported. However, it was also hypothesized that a higher trust score should be correlated with less verification; this hypothesis was not supported. This result is likely due to the fact that there were few instances they chose not to review the autonomous systems' recommendation since there was an incentive to verify the system's recommendations to improve team performance, and no penalty for the verification other than time pressure of not completing other tasks, which may have encouraged participants to review all recommendations if possible. Verification typically increases time and cognitive workload [51], but without a penalty [224] or an increased workload, there may not be the trade-off required to see the expected correlation between verification behavior and trust. Additionally, participants may have felt that the cost of verification is less than the cost of reliance [51].

Additionally, low R^2 values indicate that gaze metrics alone may not be useful for understanding trust in the autonomous system. Unlike previous research [113–116], this work did not find gaze alone to be indicative of trust. However, differences in task complexity and types of information processed may explain some of these differences. Additionally, personality traits are intentionally not included as predictors, as the focus is primarily on gaze and display components. Trust has been shown to have inter-individual differences due to prior experiences with trust, personality, or identity, and demographics [202,205,225]. Trust also may be affected by factors such as the explainability and transparency of the system, which do emerge as relatively important features in the analysis. Gaze features may be useful in explaining trust in conjunction with some of these other metrics [222,223,226]. Therefore, these findings do not support using gaze-based data alone to make recommendations about display designs and their impact on trust.

There are some limitations in this research. First, there is an inherent class imbalance since the autonomous systems had to correct the majority of the time (i.e., the autonomous system was never working against their teammate, meaning the threshold for accuracy had to be above at least 50%). The majority (78.8%) of the data is from when the system made a correct recommendation, and thus where participants often correctly accepted the recommendation (90.6% of the time). This impacts the classification analysis methods used to identify what features are important. Additionally, as mentioned previously, there are few instances where participants chose not to review the recommendation due to the lack of a penalty to verify and insufficient time pressure. While this is beneficial for having a larger dataset to understand decision making and what data streams are used, it may have influenced the correlations between trust and participant actions, potentially impacting trust results. While the trust slider scale is not validated on its own, previous analysis showed strong correlations with Jian's Trust in Autonomous System survey, allowing us to achieve a more dynamic measurement without a loss in resolution [227]. Additionally, binning scores, such as creating a proportion for the number of recommendations agreed with per epoch, may reduce time sensitivity, inducing measurement errors and weakening the correlation [117]. However, this is a common metric used in the literature as it allows for parametric analysis and less frequent trust queries, and affects most of the related literature. Finally, the layout of the display used for this study is static and is consistent across all participants. This may influence the results related to information use. Specifically, the participants' reliance on data streams may be attributed to the position of the information on the display within a left-to-right scan pattern rather than the inherent usefulness of the data. Without changing the order of the data streams, it is unclear if the important streams are due to the location or the types of information they contain. Future work should investigate changing the order of the data to understand if participants are relying heavily on the visual source because of its leftmost position, it is easiest to understand, or it is preferred by participants. Future work should also consider more complex gaze metrics, including a quantification and comparison between scan patterns.

Additionally, in this aim visualizations are used as a way to convey information to verify the autonomous system recommendation during the trusting task. Visualizations may be beneficial over text-based displays as an visualizations can integrate multiple pieces of information in an interpretable manner. Compared to the same amount of information in multiple lines of text, visualizations may not be as cognitively demanding, enabling smoother integration of information.

Future work should further study this idea and quantify the benefits of including visualizations for trusting tasks, beyond the benefits of visualization explored in previous aims.

8.5 Summary and Contributions

This aim sought to understand what information operators use when making decisions, using gaze behaviors, and how this relates to their trust. The findings indicate that an operator's decision accuracy is influenced by both the type of conflicting information as well as the system's recommendation. It also found that certain gaze behaviors, such as review duration, the time spent on the visual screen, and the gaze entropy, can be correlated to their accuracy, but that behaviors cannot be used alone to understand trust. Operator's gaze patterns, and the system recommendation, should be taken into account when designing future displays, but additional work should continue to identify how information and its presentation on a display may influence an operator's trust in autonomous decision support systems. The methodology developed for this study reinforces the idea of using gaze and visual metrics to further understand how operators use a display, and these methods could help in future research on display design and trust to understand which display components and transparency features are important, or to improve training.

Chapter 9: Conclusion

9.1 Summary of Findings and Contributions

This dissertation explored aspects of display design to address challenges with future spaceflight supervisory control operations and training. Supervisory control is often understudied, particularly for novel display designs, and is an increasingly important control modality for exploration, manufacturing, and transportation operations.

In Aim 1, a systemized literature review was conducted on menus and text in VR, developing a set of operationally relevant VR display design principles. This helps address the lack of dedicated VR-specific display design principles geared towards operational use. These findings were extended and used in the development of the VR displays throughout the dissertation. The developed design space can be applied towards other operational contexts, and optimizing the displays for VR may make the benefits of VR realized.

In Aim 2, VR and 3D visualizations were compared to a traditional display for the monitoring of spaceflight operations, finding that while 3D visualizations (either on a screen or in VR) were important, VR did not provide additional benefits. Specifically, 3D visualizations were found to improve SA, which is critical for safe operations and may lead to improvements in collision prediction and anomaly detection.

Aim 3 extended Aim 2 into supervisory control operations where the operator had some, but limited, control authority. The findings were similar to those of Aim 2, where 3D visualizations (either screen or VR) offered improvements over traditional displays, but VR alone did not provide additional benefits. 3D visualizations were found to improve performance on the complex Hard trials and resulted in an increased perceived ability to understand the relative orbital trajectories. In future, even more complex, mission paradigms (i.e., refueling) 3D visualizations may be even more critical to improve operators' actions. Combined, these aims advance the literature on VR for monitoring and supervisory control operations.

Aim 4 considered how these three displays could be used as a training tool for spacecraft operations and found that VR is a promising training modality. Training in VR yielded improved SA in subsequent operational tasks. Additionally, it had higher subjective utility ratings, particularly for VR as a way to improve the understanding of orbital motion and collision likelihood during training. Aim 4 contributes towards the understanding of VR for supervisory control training. VR training may lead to improved operator mental models, conceptual understanding of traditional displays, and safer operations, counteracting challenges associated with traditional displays during remote operations.

Finally, Aim 5 considered only screen displays for a spaceflight human-autonomy teaming challenge and looked at what behaviors went into making correct decisions and leading to calibrated trust, addressing the gap in understanding how a screen is used for decision-making and trust. This found that an operator's decision accuracy is influenced by both the type of conflicting information and the recommendation type, indicating that displays may benefit from different designs based on the recommendation. Their gaze behaviors, including review duration, time spent on the visual screen, and gaze entropy, can be correlated to accuracy and weakly correlated to trust. Operator gaze patterns should be taken into account when designing novel displays, especially in situations where accurate decisions are important. The developed framework and methodology may be applicable towards designing novel displays and improving training.

9.2 Limitations and Future Work

In addition to the specific future work and limitations discussed in each of the aims, there are some applicable to all of the experimental aims. The first is the population studied. The participants may not be representative of the trained operators who would use these types of displays and instead have varying backgrounds in orbital mechanics and satellite operations. While subject background knowledge is accounted for in the statistical analysis to reduce these effects, having a more representative subject pool would be ideal. This likely affects Aims 2, 3, and 5, which study aspects of operations. Future work can consider whether these results translate to the highly trained population of operators that typically work with these systems, and commensurately increase the task complexity for those analyses. Previous work has found that expertise may influence subjective ratings, performance, and workload when using novel displays [174], particularly this found that expert users may be less willing to adapt to new displays. Aim 4, which studied training, will likely be less affected by this, as the participant population is representative of the type of people who could become operators. However, the participants had different baseline knowledge prior to the experiment, which may have affected their training experience. Additionally, these results may change through repeated interaction with a system, both for training and operations. Participants only experienced one session in VR and had some, but minimal, familiarization time with the display prior. Using these systems over longer periods of time may lead to more comfort and familiarity, affecting the results and performance. Repeated interactions may also help shape operator perception of the display. This may overcome some of the issues with subjective ratings, particularly those influenced positively by the novelty of experiencing VR or negatively from a change from familiar displays for more expert users.

Beyond this, for all the experimental aims, simplified spaceflight applications were considered. Although complexity was increased throughout the dissertation, and aspects of the operation task were rooted in physics and realism, there are ways to make this research more complex and realistic. This could include increasing task complexity and fidelity for Aims 3 and 4 by including solar charging with respect to sun angles, physics-based uncertainty rates, the introduction of anomalies, or varying the magnitude and directions of the burn. Increasing the complexity may allow for more benefits to be seen for VR and visualizations, especially since performance differences were only seen among the Hard difficulty trials. For Aim 5, complexity can be increased by including more realistic photos and troop movements, more realistic time pressures, and a more realistic sense of risk.

Additionally, VR technology is rapidly improving, particularly in areas like increased resolution and frame rates, and improved comfort that provides a more usable and comfortable experience. While participants did not experience cybersickness and the majority did not have VR-specific issues, some participants noted that VR was uncomfortable to wear, increased physical effort to use, and was harder to read. Furthermore, commercialization and subsequent standardization of VR displays may influence these guidelines and usability. If certain features become ingrained as default, this may affect future operator experiences, and best practices may have to evolve alongside that. Improvements and commercialization of VR display may allow for improvements in the VR experience, particularly with regards to usability. As technology improves, future work can reevaluate these results. Furthermore, as future generations that may grow up using technologies like VR, tablets, or phones, enter the operator workforce, this may influence the results. For example, younger generations that grew up learning to use rapidly improving and changing technologies may be faster to adapt to and accept VR. As such, it would be valuable to investigate how generational shifts in users affect the applicability of these results. Furthermore, generational-based differences in operators may motivate the use of training strategies for different operator backgrounds. If, for instance, younger generations are found to be less prone to information overload in VR or grow up using VR in educational settings, the introduction of new information streams during their training may be able to be accelerated.

Finally, the results found may also be able to translate to other supervisory control paradigms such as air traffic control, maritime management, manufacturing, power generation, or exploration, especially those that present similar challenges to the spaceflight application studied here. These types of paradigms are critical and important for aspects of daily life. Future work can extend this research into these fields to understand if the findings are true across applications or application-specific. Additionally, as VR training shows promise for complex operations and understanding of orbits, future work should study it as an educational tool for other applications that require similar mental models. For example, could VR provide benefits to educate students on complex cislunar geometries, or schoolchildren on complex spatial concepts like 3D geometry or atomic orbits?

Spaceflight has implications for modern society. Prediction and monitoring for natural disasters via weather and Earth science satellites can inform communities of potential threats and save lives. Global Positioning satellites are important for navigation, agriculture, transportation, and military uses. Science satellites have far-reaching implications for understanding Earth, our solar system, and the universe. Being able to effectively operate and team with these systems is critical, and it is in society's best interest to be able to avoid unnecessary collisions and facilitate repairs, and perform maintenance on them. This work on improvements towards display designs, whether it be increased 3D visualizations, VR, or a better understanding of how current displays are used, can facilitate these important future operations and use of spacecraft.

Bibliography

[1] T. B. Sheridan, "Human Supervisory Control of Automation," in *Handbook of Human Factors and Ergonomics.* New York, United States: John Wiley & Sons, Incorporated, 2012.

[2] L. Sim, M. L. Cummings, and C. A. Smith, "Past, present and future implications of human supervisory control in space missions," *Acta Astronautica*, vol. 62, no. 10, pp. 648–655, May 2008.

[3] S. Chien, R. Sherwood, D. Tran, B. Cichy, G. Rabideau, R. Castano, A. Davis, D. Mandl, S. Frye, B. Trout, S. Shulman, and D. Boyer, "Using Autonomy Flight Software to Improve Science Return on Earth Observing One," *Journal of Aerospace Computing, Information, and Communication*, vol. 2, no. 4, pp. 196–216, 2005, publisher: American Institute of Aeronautics and Astronautics _eprint: https://doi.org/10.2514/1.12923.

[4] D. Rijlaarsdam, T. Hendrix, P. T. T. González, A. Velasco-Mata, L. Buckley, J. P. Miquel, O. A. Casaled, and A. Dunne, "The Next Era for Earth Observation Spacecraft: An Overview of CogniSAT-6," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 2450–2463, 2025.

[5] J. Y. C. Chen, E. C. Haas, and M. J. Barnes, "Human Performance Issues and User Interface Design for Teleoperated Robots," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1231–1245, Nov. 2007, conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews).

[6] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, Mar. 2004, publisher: SAGE Publications Inc.

[7] S. Warm, G. Matthews, and V. Finomore, "Vigilance, workload, and stress," *Performance Under Stress*, pp. 115–141, Jan. 2008.

[8] M. Bualat, T. Fong, M. Allan, X. Bouyssounouse, T. Cohen, L. Fluckiger, R. Gogna, L. Kobayashi, G. Lee, S. Lee, C. Provencher, E. Smith, V. To, H. Utz, D. W. Wheeler, E. Pacis, and D. Schreckenghost, "Surface Telerobotics: Development and Testing of a Crew Controlled Planetary Rover System," in *AIAA Space 2013 Conference and Exposition.* San Diego, CA: American Institute of Aeronautics and Astronautics, Sep. 2013.

[9] A. Dan and M. Reiner, "EEG-based cognitive load of processing events in 3D virtual worlds is lower than processing events in 2D displays," *International Journal of Psychophysiology*, vol. 122, pp. 75–84, Dec. 2017.

[10] N. B. S. Woods, David D., "Situation Awareness: A Critical But Ill-Defined Phenomenon," in *Situational Awareness.* Routledge, 2011, num Pages: 14.

[11] J. K. Hawley, A. L. Mares, and C. A. Giammanco, "Training for Effective Human Supervisory Control of Air and Missile Defense Systems," 2006.

[12] J. S. Tittle, A. Roesler, and D. D. Woods, "The Remote Perception Problem," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 46, no. 3, pp. 260–264, Sep. 2002, publisher: SAGE Publications Inc.

[13] M. Slater, V. Linakis, M. Usoh, and R. Kooper, "Immersion, presence and performance in virtual environments: an experiment with tri-dimensional chess," in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology - VRST '96*. Hong Kong: ACM Press, 1996, pp. 163–172.

[14] D. Whitney, E. Rosen, E. Phillips, G. Konidaris, and S. Tellex, "Comparing Robot Grasping Teleoperation Across Desktop and Virtual Reality with ROS Reality," in *Robotics Research*, ser. Springer Proceedings in Advanced Robotics, N. M. Amato, G. Hager, S. Thomas, and M. Torres-Torriti, Eds. Cham: Springer International Publishing, 2020, pp. 335–350.

[15] N. E. Seymour, A. G. Gallagher, S. A. Roman, M. K. O'Brien, V. K. Bansal, D. K. Andersen, and R. M. Satava, "Virtual Reality Training Improves Operating Room Performance," *Annals of Surgery*, vol. 236, no. 4, pp. 458–464, Oct. 2002.

[16] F. Aïm, G. Lonjon, D. Hannouche, and R. Nizard, "Effectiveness of Virtual Reality Training in Orthopaedic Surgery," *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, vol. 32, no. 1, pp. 224–232, Jan. 2016.

[17] S. G. Wheeler, H. Engelbrecht, and S. Hoermann, "Human Factors Research in Immersive Virtual Reality Firefighter Training: A Systematic Review," *Frontiers in Virtual Reality*, vol. 2, 2021.

[18] G. Sun, X. Wanyan, X. Wu, and D. Zhuang, "The Influence of HUD Information Visual Coding on Pilot's Situational Awareness," in *2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 1, Aug. 2017, pp. 139–143.

[19] M. Lager and E. A. Topp, "Remote Supervision of an Autonomous Surface Vehicle using Virtual Reality," *IFAC-PapersOnLine*, vol. 52, no. 8, pp. 387–392, 2019.

[20] T. O'Neill, N. McNeese, A. Barron, and B. Schelble, "Human-Autonomy Teaming: A Review and Analysis of the Empirical Literature," *Human Factors*, vol. 64, no. 5, pp. 904–938, Aug. 2022.

[21] K. T. Wynne and J. B. and Lyons, "An integrative model of autonomous agent teammate-likeness," *Theoretical Issues in Ergonomics Science*, vol. 19, no. 3, pp. 353–374, May 2018.

[22] M. Johnson, J. M. Bradshaw, P. Feltovich, C. Jonker, B. van Riemsdijk, and M. Sierhuis, "Autonomy and interdependence in human-agent-robot teams," *IEEE Intelligent Systems*, vol. 27, no. 2, pp. 43–51, Mar. 2012.

[23] J. K. Hawley, A. L. Mares, and C. A. Giammanco, "The Human Side of Automation: Lessons for Air Defense Command and Control:," Defense Technical Information Center, Fort Belvoir, VA, Tech. Rep., Mar. 2005.

[24] T. B. Sheridan, *Telerobotics, automation, and human supervisory control*, ser. Telerobotics, automation, and human supervisory control. Cambridge, MA, US: The MIT Press, 1992.

[25] J. C. F. de Winter and P. A. Hancock, "Why human factors science is demonstrably necessary: historical and evolutionary foundations," *Ergonomics*, vol. 64, no. 9, pp. 1115–1131, Sep. 2021, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00140139.2021.1905882.

[26] T. B. Sheridan, L. Charny, M. B. Mendel, and J. B. Roseborough, “Supervisory Control, Mental Models And Decision Aids,” in *Analysis, Design and Evaluation of Man–Machine Systems 1988*, ser. IFAC Symposia Series, J. Ranta, Ed. Amsterdam: Pergamon, Jan. 1989, pp. 175–181.

[27] M. Cummings, L. Huang, H. Zhu, D. Finkelstein, and R. Wei, “The Impact of Increasing Autonomy on Training Requirements in a UAV Supervisory Control Task,” *Journal of Cognitive Engineering and Decision Making*, vol. 13, no. 4, pp. 295–309, Dec. 2019.

[28] M. R. Endsley, “Direct measurement of situation awareness: Validity and use of SAGAT,” in *Situation awareness analysis and measurement.* Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2000, pp. 147–173.

[29] ——, “Design and Evaluation for Situation Awareness Enhancement,” *Proceedings of the Human Factors Society Annual Meeting*, vol. 32, no. 2, pp. 97–101, oct 1988, publisher: SAGE Publications.

[30] M. Endsley and W. Jones, “Situation Awareness Information Dominance & Information Warfare.” Wright-Patterson Air Force Base, OH: United States Air Force Armstrong Laboratory, Tech. Rep. AL/CF-TR-19970156, 1997, section: Technical Reports.

[31] R. Opromolla, D. Grishko, J. Auburn, R. Bevilacqua, L. Buinhas, J. Cassady, M. Jäger, M. Jankovic, J. Rodriguez, M. A. Perino, and B. Bastida-Virgili, “Future in-orbit servicing operations in the space traffic management context,” *Acta Astronautica*, vol. 220, pp. 469–477, Jul. 2024.

[32] J. P. Davis, J. P. Mayberry, and J. P. Penn, “On-Orbit Servicing: Inspection, Repair, Refuel, Upgrade, and Assembly of Satellites in Space,” Aerospace Corporation, Tech. Rep. OTR201900236, 2019.

[33] K. L. Hobbs, S. Phillips, M. Simon, J. B. Lyons, J. Culbertson, H. S. Clouse, N. Hamilton, K. Dunlap, Z. S. Lippay, J. Aurand, Z. I. Bell, T. Hammack, D. Ayres, and R. Lim, “The safe trusted autonomy for responsible space program,” 2025.

[34] F. Sellmaier and H. Frei, “Operations of On-Orbit Servicing Missions,” in *Spacecraft Operations*, F. Sellmaier, T. Uhlig, and M. Schmidhuber, Eds. Cham: Springer International Publishing, 2022, pp. 491–529.

[35] S. G. Hart and L. E. Staveland, “Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research,” in *Advances in Psychology*, ser. Human Mental Workload, P. A. Hancock and N. Meshkati, Eds. North-Holland, Jan. 1988, vol. 52, pp. 139–183.

[36] J. M. Reiter, “Multi-Mission Operator Training Practices,” in *SpaceOps 2012 Conference.* Stockholm, Sweden: American Institute of Aeronautics and Astronautics, Jun. 2012.

[37] G. Dittemore and C. Bertels, “The Final Count Down: A Review of Three Decades of Flight controller Training Methods for Space Shuttle Mission Operations,” in *AIAA SPACE 2011 Conference & Exposition.* Long Beach, California: American Institute of Aeronautics and Astronautics, Sep. 2011.

[38] J. Liebowitz and P. Lightfoot, “Training NASA Satellite Operators: An Expert System Consultant Approach,” *Educational Technology*, vol. 27, no. 11, pp. 41–47, 1987, publisher: Educational Technology Publications, Inc.

[39] F. Stathopoulos and K. Oezdemir, “Human Factor and Knowledge Management in 24/7 Multi-Mission Satellite Operations,” in *2018 SpaceOps Conference*. Marseille, France: American Institute of Aeronautics and Astronautics, May 2018.

[40] S. Kolbeck, C. Amodio, and V. Burkhardt, “Job analysis and collaborative training for spacecraft operators in future control rooms.” AIAA, May 2018.

[41] C. Decoust, M. Magalhaes, P. Rolland, K. Pizolato, and M. Farias, “From Submarine to Satellite Operations: A Training Success Story.” AIAA, May 2018.

[42] D. Adams, *The Hitchhiker’s Guide to the Galaxy.* Harmony Books, 1979.

[43] W. B. Rouse and N. M. Morris, “on looking into the black box: prospects and limits in the search for mental models,” Tech. Rep. AD-A 159 080, May 1985.

[44] N. M. Morris and W. B. Rouse, “The effects of type of knowledge upon human problem solving in a process control task,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-15, no. 6, pp. 698–707, Nov. 1985.

[45] M. Endsley, *Situation Awareness Measurement: How to Measure Situation Awareness in Individuals and Teams.* Human Factors and Ergonomics Society, 2021.

[46] J. Meyer, R. Wiczorek, and T. Günzler, “Measures of Reliance and Compliance in Aided Visual Scanning,” *Human Factors*, vol. 56, no. 5, pp. 840–849, Aug. 2014.

[47] N. Moray and T. and Inagaki, “Attention and complacency,” *Theoretical Issues in Ergonomics Science*, vol. 1, no. 4, pp. 354–365, Jan. 2000.

[48] J. E. Bahner, A.-D. Hüper, and D. Manzey, “Misuse of automated decision aids: Complacency, automation bias and the impact of training experience,” *International Journal of Human-Computer Studies*, vol. 66, no. 9, pp. 688–699, Sep. 2008.

[49] J. C. Walliser, E. J. De Visser, and T. H. Shaw, “Application of a System-Wide Trust Strategy when Supervising Multiple Autonomous Agents,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 60, no. 1, pp. 133–137, Sep. 2016.

[50] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, “Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs,” *Journal of Cognitive Engineering and Decision Making*, vol. 2, no. 2, pp. 140–160, Jun. 2008.

[51] N. Ezer, A. D. Fisk, and W. A. Rogers, “Age-Related Differences in Reliance Behavior Attributable to Costs Within a Human-Decision Aid System,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 50, no. 6, pp. 853–863, Dec. 2008.

[52] T. B. Sheridan, “Musings on Telepresence and Virtual Presence,” *Presence: Teleoperators and Virtual Environments*, vol. 1, no. 1, pp. 120–126, Feb. 1992.

[53] A. Elor, T. Thang, B. P. Hughes, A. Crosby, A. Phung, E. Gonzalez, K. Katija, S. H. D. Haddock, E. J. Martin, B. E. Erwin, and L. Takayama, "Catching Jellies in Immersive Virtual Reality: A Comparative Teleoperation Study of ROVs in Underwater Capture Tasks," in *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*. Osaka Japan: ACM, Dec. 2021, pp. 1–10.

[54] A. Naceri, D. Mazzanti, J. Bimbo, Y. T. Tefera, D. Prattichizzo, D. G. Caldwell, L. S. Mattos, and N. Deshpande, "The Vicarios Virtual Reality Interface for Remote Robotic Teleoperation," *Journal of Intelligent & Robotic Systems*, vol. 101, no. 4, p. 80, Apr. 2021.

[55] E. Dima, K. Brunnstrom, and T. Qureshi, "View Position Impact on QoE in an Immersive Telepresence System for Remote Operation," in *International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, p. 3.

[56] M. Wilde, Z. K. Chua, and A. Fleischner, "Effects of Multivantage Point Systems on the Teleoperation of Spacecraft Docking," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 2, pp. 200–210, Apr. 2014, conference Name: IEEE Transactions on Human-Machine Systems.

[57] A. Hosseini and M. Lienkamp, "Enhancing telepresence during the teleoperation of road vehicles using HMD-based mixed reality," in *2016 IEEE Intelligent Vehicles Symposium (IV)*, Jun. 2016, pp. 1366–1373.

[58] M. Kraus, N. Weiler, D. Oelke, J. Kehrer, D. A. Keim, and J. Fuchs, "The Impact of Immersion on Cluster Identification Tasks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 525–535, Jan. 2020, conference Name: IEEE Transactions on Visualization and Computer Graphics.

[59] M. Whitlock, S. Smart, and D. A. Szafir, "Graphical Perception for Immersive Analytics," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, Mar. 2020, pp. 616–625, iSSN: 2642-5254.

[60] M. Kraus, J. Fuchs, B. Sommer, K. Klein, U. Engelke, D. Keim, and F. Schreiber, "Immersive Analytics with Abstract 3D Visualizations: A Survey," *Computer Graphics Forum*, vol. 41, no. 1, pp. 201–229, 2022, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14430.

[61] R. Etemadpour, E. Monson, and L. Linsen, "The Effect of Stereoscopic Immersive Environments on Projection-Based Multi-dimensional Data Visualization," in *2013 17th International Conference on Information Visualisation*, Jul. 2013, pp. 389–397, iSSN: 2375-0138.

[62] C. Hurter, N. H. Riche, S. M. Drucker, M. Cordeil, R. Alligier, and R. Vuillemot, "FiberClay: Sculpting Three Dimensional Trajectories to Reveal Structural Insights," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 704–714, Jan. 2019, conference Name: IEEE Transactions on Visualization and Computer Graphics.

[63] D. Whitney, E. Rosen, E. Phillips, G. Konidaris, and S. Tellex, "Comparing Robot Grasping Teleoperation Across Desktop and Virtual Reality with ROS Reality," in *Robotics Research*, N. M. Amato, G. Hager, S. Thomas, and M. Torres-Torriti, Eds. Cham: Springer International Publishing, 2020, vol. 10, pp. 335–350, series Title: Springer Proceedings in Advanced Robotics.

[64] J. M. Read and J. J. Saleem, "Task Performance and Situation Awareness with a Virtual Reality Head-Mounted Display," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 61, no. 1, pp. 2105–2109, Sep. 2017.

[65] M. R. Endsley, "Measurement of Situation Awareness in Dynamic Systems," *Human Factors*, vol. 37, no. 1, pp. 65–84, Mar. 1995, publisher: SAGE Publications Inc.

[66] B. L. Hooey, D. B. Kaber, J. A. Adams, T. W. Fong, and B. F. Gore, "The Underpinnings of Workload in Unmanned Vehicle Systems," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 452–467, Oct. 2018, conference Name: IEEE Transactions on Human-Machine Systems.

[67] F. Sittner, O. Hartmann, S. Montenegro, J.-P. Friese, L. Brubach, M. E. Latoschik, and C. Wienrich, "An Update on the Virtual Mission Control Room," in *37th Annual Small Satellite Conference*, Logan Utah, Aug. 2023.

[68] R. Gad, S. Martin, M. Olbrich, M. Fischer, S. Baci, F. Ruecker, and R. Sfaxi, "Towards Leveraging Augmented and Virtual Reality for Spacecraft Mission Operations at ESOC," in *17th International Conference on Space Operations*, Dubai, UAE, Mar. 2023.

[69] K. Tsigkounis, A. Komninos, N. Politis, and J. Garofalakis, "Monitoring Maritime Industry 4.0 Systems through VR Environments," in *CHI Greece 2021: 1st International Conference of the ACM Greek SIGCHI Chapter*. Online (Athens, Greece) Greece: ACM, Nov. 2021, pp. 1–8.

[70] M. Cordeil, T. Dwyer, and C. Hurter, "Immersive solutions for future Air Traffic Control and Management," in *Proceedings of the 2016 ACM Companion on Interactive Surfaces and Spaces*. Niagara Falls Ontario Canada: ACM, Nov. 2016, pp. 25–31.

[71] A. L. Gorbunov and E. E. Nechaev, "Augmented Reality Technologies in Air Transport Control Systems," in *2022 Systems of Signals Generating and Processing in the Field of on Board Communications*, Mar. 2022, pp. 1–5, iSSN: 2768-0118.

[72] F. van den Oever, M. Fjeld, and B. Sætrevik, "A Systematic Literature Review of Augmented Reality for Maritime Collaboration," *International Journal of Human–Computer Interaction*, vol. 40, no. 15, pp. 4116–4131, Aug. 2024, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2023.2209838.

[73] Torres, J. P. Molina, A. S. García, and P. González, "Prototyping of Augmented Reality interfaces for air traffic alert and their evaluation using a Virtual Reality aircraft-proximity simulator." IEEE Computer Society, Mar. 2024, pp. 817–826.

[74] J. Rohacs, D. Rohacs, and I. Jankovics, "Conceptual development of an advanced air traffic controller workstation based on objective workload monitoring and augmented reality," *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, vol. 230, no. 9, pp. 1747–1761, Jul. 2016, publisher: IMECHE.

[75] R. Reisman and D. Brown, "Design of Augmented Reality Tools for Air Traffic Control Towers," in *6th AIAA Aviation Technology, Integration and Operations Conference (ATIO)*. Wichita, Kansas: American Institute of Aeronautics and Astronautics, Sep. 2006.

[76] S. Bagassi, M. Corsi, F. De Crescenzio, R. Santarelli, A. Simonetti, L. Moens, and M. Terenzi, "Virtual/augmented reality-based human–machine interface and interaction modes in airport control towers," *Scientific Reports*, vol. 14, no. 1, p. 13579, Jun. 2024, publisher: Nature Publishing Group.

[77] S. Kalamkar, V. Biener, F. Beck, and J. Grubert, "Remote Monitoring and Teleoperation of Autonomous Vehicles—Is Virtual Reality an Option?" in *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2023, pp. 463–472, iSSN: 2473-0726.

[78] A. D. Garcia, J. Schlueter, and E. Paddock, "Training Astronauts using Hardware-in-the-Loop Simulations and Virtual Reality," in *AIAA Scitech 2020 Forum.* Orlando, FL: American Institute of Aeronautics and Astronautics, Jan. 2020.

[79] R. E. Mayer, G. Makransky, and J. Parong, "The Promise and Pitfalls of Learning in Immersive Virtual Reality," *International Journal of Human–Computer Interaction*, vol. 39, no. 11, pp. 2229–2238, Jul. 2023, publisher: Taylor & Francis.

[80] M. Li, Z. Sun, Z. Jiang, Z. Tan, and J. Chen, "A Virtual Reality Platform for Safety Training in Coal Mines with AI and Cloud Computing," *Discrete Dynamics in Nature and Society*, vol. 2020, p. e6243085, Oct. 2020, publisher: Hindawi.

[81] B. Xie, H. Liu, R. Alghofaili, Y. Zhang, Y. Jiang, F. D. Lobo, C. Li, W. Li, H. Huang, M. Akdere, C. Mousas, and L.-F. Yu, "A Review on Virtual Reality Skill Training Applications," *Frontiers in Virtual Reality*, vol. 2, 2021.

[82] M. Thompson, C. Uz-Bilgin, M. S. Tutwiler, M. Anteneh, J. C. Meija, A. Wang, P. Tan, R. Eberhardt, D. Roy, J. Perry, and E. Klopfer, "Immersion positively affects learning in virtual reality games compared to equally interactive 2d games," *Information and Learning Sciences*, vol. 122, no. 7/8, pp. 442–463, Jul. 2021, publisher: Emerald Publishing Limited.

[83] G. Makransky and L. Lilleholt, "A structural equation modeling investigation of the emotional value of immersive virtual reality in education," *Educational Technology Research and Development*, vol. 66, no. 5, pp. 1141–1164, 2018, place: Germany Publisher: Springer.

[84] L. v. Dammen, T. T. Finseth, B. H. McCurdy, N. P. Barnett, R. A. Conrady, A. G. Leach, A. F. Deick, A. L. Van Steenis, R. Gardner, B. L. Smith, A. Kay, and E. A. Shirtcliff, "Evoking stress reactivity in virtual reality: A systematic review and meta-analysis," *Neuroscience and Biobehavioral Reviews*, vol. 138, p. 104709, Jul. 2022.

[85] T. Kojic, M. Vergari, S. Möller, and J.-N. Voigt-Antons, "Assessing User Experience of Text Readability with Eye Tracking in Virtual Reality," in *Virtual, Augmented and Mixed Reality: Design and Development: 14th International Conference, VAMR 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26 – July 1, 2022, Proceedings, Part I.* Berlin, Heidelberg: Springer-Verlag, Jun. 2022, pp. 199–211.

[86] M. L. van Emmerik, S. C. de Vries, and J. E. Bos, "Internal and external fields of view affect cybersickness," *Displays*, vol. 32, no. 4, pp. 169–174, Oct. 2011.

[87] S. Davis, K. Nesbitt, and E. Nalivaiko, "A Systematic Review of Cybersickness," in *Proceedings of the 2014 Conference on Interactive Entertainment.* Newcastle NSW Australia: ACM, Dec. 2014, pp. 1–9.

[88] Y. Chen, X. Wang, and H. Xu, “Human factors/ergonomics evaluation for virtual reality headsets: a review,” *CCF Transactions on Pervasive Computing and Interaction*, vol. 3, no. 2, pp. 99–111, Jun. 2021.

[89] D. B. Van de Merwe, L. Van Maanen, F. B. Ter Haar, R. J. E. Van Dijk, N. Hoeba, and N. Van der Stap, “Human-Robot Interaction During Virtual Reality Mediated Teleoperation: How Environment Information Affects Spatial Task Performance and Operator Situation Awareness,” in *Virtual, Augmented and Mixed Reality. Applications and Case Studies*, ser. Lecture Notes in Computer Science, J. Y. Chen and G. Fragomeni, Eds. Cham: Springer International Publishing, 2019, pp. 163–177.

[90] J. Wentzel, M. Lakier, J. Hartmann, F. Shazib, G. Casiez, and D. Vogel, “A Comparison of Virtual Reality Menu Archetypes: Raycasting, Direct Input, and Marking Menus.” *IEEE transactions on visualization and computer graphics*, vol. PP, Jun. 2024.

[91] T. Luong, Y. F. Cheng, M. Möbus, A. Fender, and C. Holz, “Controllers or Bare Hands? A Controlled Evaluation of Input Techniques on Interaction Performance and Exertion in Virtual Reality,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 11, pp. 4633–4643, Nov. 2023.

[92] J. Andersson and Y. Hu, “Exploring the Impact of Menu Systems, Interaction Methods, and Sitting or Standing Posture on User Experience in Virtual Reality,” C. Gittens, A. Hogue, and A. Cannavo, Eds., 2023.

[93] C. Mckenzie and A. Glazier, “Designing screen interfaces for VR,” Mountain View, May 2017.

[94] R. Yao, T. Heath, A. Davies, T. Forsyth, N. Mitchell, and P. Hoberman, “Oculus VR Best Practices Guide,” Oculus, Tech. Rep., 2014.

[95] J. Y. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, and M. Barnes, “Situation Awareness-Based Agent Transparency:,” Defense Technical Information Center, Fort Belvoir, VA, Tech. Rep., Apr. 2014.

[96] J. Koo, J. Kwac, W. Ju, M. Steinert, L. Leifer, and C. Nass, “Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance,” *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 9, no. 4, pp. 269–275, Nov. 2015.

[97] X. J. Yang, V. V. Unhelkar, K. Li, and J. A. Shah, “Evaluating Effects of User Experience and System Transparency on Trust in Automation,” in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. Vienna Austria: ACM, Mar. 2017, pp. 408–416.

[98] R. Luo, D. Na, and X. J. and Yang, “Evaluating Effects of Enhanced Autonomy Transparency on Trust, Dependence, and Human-Autonomy Team Performance over Time,” *International Journal of Human–Computer Interaction*, vol. 38, no. 18-20, pp. 1962–1971, Dec. 2022.

[99] J. E. Mercado, M. A. Rupp, J. Y. C. Chen, M. J. Barnes, D. Barber, and K. Procci, “Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management,” *Human Factors*, vol. 58, no. 3, pp. 401–415, May 2016.

[100] K. Stowers, N. Kasdaglis, M. A. Rupp, O. B. Newton, J. Y. C. Chen, and M. J. Barnes, "The IMPACT of Agent Transparency on Human Performance," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 3, pp. 245–253, Jun. 2020.

[101] M. E. Gruber, P. A. Hancock, D. J. Barber, R. Wohleber, and J. B. Lyons, "The Impact of Transparency on Human-Autonomy Teaming," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 68, no. 1, pp. 1783–1788, Sep. 2024, publisher: SAGE Publications Inc.

[102] K. Van De Merwe, S. Mallam, and S. Nazir, "Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 66, no. 1, pp. 180–208, Jan. 2024.

[103] K. Stowers, N. Kasdaglis, M. Rupp, J. Chen, D. Barber, and M. Barnes, "Insights into Human-Agent Teaming: Intelligent Agent Transparency and Uncertainty," in *Advances in Human Factors in Robots and Unmanned Systems*, P. Savage-Knepshield and J. Chen, Eds. Cham: Springer International Publishing, 2017, pp. 149–160.

[104] T. Wang and N. Lau, "Level of detail in visualization for human autonomy teaming: Speed, accuracy, and workload effects," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 67, no. 1, pp. 549–555, Sep. 2023.

[105] N. Moacdieh and N. Sarter, "Data density and poor organization: Analyzing the performance and Attentional effects of two aspects of display clutter," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 59, no. 1, pp. 1336–1340, Sep. 2015.

[106] C. Westin, C. Borst, and B. Hilburn, "Strategic Conformance: Overcoming Acceptance Issues of Decision Aiding Automation?" *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 41–52, Feb. 2016.

[107] E. B. Entin and E. E. Entin, "The Effects of Decision Time and Decision-Aid Accuracy on Workload and Performance," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 41, no. 1, pp. 177–181, Oct. 1997.

[108] C.-C. Ting and S. Gluth, "Unraveling information processes of decision-making with eye-tracking data," *Frontiers in Behavioral Economics*, vol. 3, Aug. 2024, publisher: Frontiers.

[109] S. M. Smith and I. Krajbich, "Gaze Amplifies Value in Decision Making," *Psychological Science*, vol. 30, no. 1, pp. 116–128, Jan. 2019, publisher: SAGE Publications Inc.

[110] W.-C. Li, J. Zhang, T. Le Minh, J. Cao, and L. Wang, "Visual scan patterns reflect to human-computer interactions on processing different types of messages in the flight deck," *International Journal of Industrial Ergonomics*, vol. 72, pp. 54–60, Jul. 2019.

[111] U. Ahlstrom and F. J. Friedman-Berg, "Using eye movement activity as a correlate of cognitive workload," *International Journal of Industrial Ergonomics*, vol. 36, no. 7, pp. 623–636, Jul. 2006, num Pages: 14 Place: Amsterdam Publisher: Elsevier Web of Science ID: WOS:000238878900002.

[112] R. Lavine, J. Sibert, M. Gokturk, and B. Dickens, “Eye-tracking measures and human performance in a vigilance task,” *Aviation, Space, and Environmental Medicine*, vol. 73, no. 4, pp. 367–371, Apr. 2002.

[113] Y. Lu and N. Sarter, “Eye Tracking: A Process-Oriented Method for Inferring Trust in Automation as a Function of Priming and System Reliability,” *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 6, pp. 560–568, Dec. 2019.

[114] S. Hergeth, L. Lorenz, R. Vilimek, and J. F. Krems, “Keep Your Scanners Peeled: Gaze Behavior as a Measure of Automation Trust During Highly Automated Driving,” *Human Factors*, vol. 58, no. 3, pp. 509–519, May 2016, publisher: SAGE Publications Inc.

[115] F. Walker, J. Wang, M. H. Martens, and W. B. Verwey, “Gaze behaviour and electrodermal activity: Objective measures of drivers’ trust in automated vehicles,” *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 64, pp. 401–412, Jul. 2019.

[116] Y. W. Kim and Y. G. and Ji, “Designing for Trust: How Human-Machine Interface Can Shape the Future of Urban Air Mobility,” *International Journal of Human–Computer Interaction*, vol. 41, no. 2, pp. 1190–1203, Jan. 2025.

[117] S. C. Kohn, E. J. de Visser, E. Wiese, Y.-C. Lee, and T. H. Shaw, “Measurement of Trust in Automation: A Narrative Review and Reference Guide,” *Frontiers in Psychology*, vol. 12, p. 4138, 2021.

[118] A. Barthou, A. Kemeny, G. Reymond, F. Merienne, and A. Berthoz, “Driver trust and reliance on a navigation system: Effect of graphical display,” in *Les Collections de l’INRETS*. Paris, France: INRETS, Sep. 2010, pp. 199–210.

[119] M. Yeh and C. D. Wickens, “Display Signaling in Augmented Reality: Effects of Cue Reliability and Image Realism on Attention Allocation and Trust Calibration,” *Human Factors*, vol. 43, no. 3, pp. 355–365, Sep. 2001, publisher: SAGE Publications Inc.

[120] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, “The PRISMA 2020 statement: An updated guideline for reporting systematic reviews,” *BMJ*, vol. 372, p. n71, Mar. 2021.

[121] I. Lediaeva and J. J. LaViola, “Evaluation of Body-referenced graphical menus in virtual environments,” in *Graphics Interface 2020, GI 2020, May 28, 2020 - May 29, 2020*, ser. Proceedings - Graphics Interface, vol. 2020-May. Toronto, Virtual, Online, ON, Canada: Canadian Information Processing Society, 2020.

[122] P. Monteiro, H. Coelho, G. Goncalves, M. Melo, and M. Bessa, “Comparison of Radial and Panel Menus in Virtual Reality,” *IEEE Access*, vol. 7, pp. 116 370–116 379, 2019.

[123] M. Mundt and T. Mathew, “An Evaluation of Pie Menus for System Control in Virtual Reality,” 2020, fraunhofer Inst Kommunikat Informat Verarbeitung.

[124] R. Pandey and K. Sorathia, “A Pilot Study on Hand-Referenced Menu User Interfaces for Head-Mounted Display Virtual Reality,” 2024, pp. 260–263.

[125] A. Santos, T. Zarraonandia, P. Díaz, and I. Aedo, "A Comparative Study of Menus in Virtual Reality Environments," in *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces.* Brighton United Kingdom: ACM, Oct. 2017, pp. 294–299.

[126] D. Bowman and C. Wingrave, "Design and evaluation of menu systems for immersive virtual environments," in *Proceedings IEEE Virtual Reality 2001*, Mar. 2001, pp. 149–156.

[127] P. M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," *Journal of Experimental Psychology*, vol. 47, no. 6, pp. 381–391, 1954.

[128] P. Dickinson, A. Cardwell, A. Parke, K. Gerling, J. Murray, and IEEE Computer Society, "Diegetic Tool Management in a Virtual Reality Training Simulation," 2021, pp. 131–139.

[129] K. Koehle, M. Hoppe, A. Schmidt, and V. Maekelae, "Diegetic and Non-diegetic Health Interfaces in VR Shooter Games," C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, and K. Inkpen, Eds., vol. 12934, 2021, pp. 3–11.

[130] Q. Marre, L. Caroux, and J.-C. Sakdavong, "Video Game Interfaces and Diegesis: The Impact on Experts and Novices' Performance and Experience in Virtual Reality," *International Journal of Human–Computer Interaction*, vol. 37, no. 12, pp. 1089–1103, Jul. 2021, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2020.1870819.

[131] E. Nava and A. Jalote-Parmar, "Visualization Techniques in VR for Vocational Education: Comparison of Realism and Diegesis on Performance, Memory, Perception and Perceived Usability," K. Santosh, A. Patel, A. Ghosh, and K. Patel, Eds., vol. 2030, 2024, pp. 104–116.

[132] D. Queck, I. Albert, G. Volkmar, R. Malaka, and M. Herrlich, "Physiological Data Placement Recommendations for VR Sport Applications," J. Chen and G. Fragomeni, Eds., vol. 14027, 2023, pp. 72–85, univ Kaiserslautern Landau RPTU.

[133] P. Salomoni, C. Prandi, M. Roccetti, L. Casanova, L. Marchetti, and G. Marfia, "Diegetic user interfaces for virtual environments with HMDs: a user experience study with oculus rift," *Journal on Multimodal User Interfaces*, vol. 11, no. 2, pp. 173–184, Jun. 2017, studio Evil Srl.

[134] R. Rzayev, P. Ugnivenko, S. Graf, V. Schwind, and N. Henze, "Reading in VR: The Effect of Text Presentation Type and Location," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* Yokohama Japan: ACM, May 2021, pp. 1–10.

[135] P. A. Yamin, J. Park, and H. K. Kim, "In-vehicle human–machine interface guidelines for augmented reality head-up displays: A review, guideline formulation, and future research directions," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 104, pp. 266–285, Jul. 2024.

[136] C. I. Johnson and R. E. Mayer, "An eye movement analysis of the spatial contiguity effect in multimedia learning," *Journal of Experimental Psychology: Applied*, vol. 18, no. 2, pp. 178–191, 2012.

[137] C. D. Wickens and C. M. Carswell, "The proximity compatibility principle: Its psychological foundation and relevance to display design," *Human Factors*, vol. 37, no. 3, pp. 473–494, 1995.

[138] R. L. Newman and K. W. Greeley, “Helmet-Mounted Display Design Guide,” NASA, Tech. Rep. TR-97-11, Nov. 1997.

[139] T. H. Harding and W. McLean, “Head Mounted Display Guidelines for Future Vertical Lift Aircraft,” United States Army Aeromedical Research Laboratory, Tech. Rep. USAARL-TECH-TR–2023-19, Mar. 2023.

[140] L. Bensch, T. Nilsson, J. Wulkop, P. de Medeiros, N. D. Herzberger, M. Preutenborbeck, A. Gerndt, F. Flemisch, F. Dufresne, G. Albuquerque, A. Cowley, and ACM, “Designing for Human Operations on the Moon: Challenges and Opportunities of Navigational HUD Interfaces,” 2024, fraunhofer FKIE Arts & Metiers Inst Technol.

[141] Meta, “Design Immersive Experience,” https://developers.meta.com/horizon/design/accessibility/, 2025.

[142] R. W. Proctor and T. V. Zandt, *Human Factors in Simple and Complex Systems*, 3rd ed. Boca Raton: CRC Press, Oct. 2017.

[143] C. D. Wickens, J. G. Hollands, S. Banbury, and R. Parasuraman, *Engineering psychology and human performance*, fourth edition ed. Boston: Pearson, 2013.

[144] S. L. Buchner, A. Rindfuss, J. Wood, H. Schaub, and A. P. Hayman, “Impacts of 3D visualizations and virtual reality in display designs for remote monitoring of satellite operations,” *Frontiers in Virtual Reality*, vol. 6, Feb. 2025, publisher: Frontiers.

[145] P. W. Kenneally, S. Piggott, and H. Schaub, “Basilisk: A Flexible, Scalable and Modular Astrodynamics Simulation Framework,” *Journal of Aerospace Information Systems*, vol. 17, no. 9, pp. 496–507, Sep. 2020, publisher: American Institute of Aeronautics and Astronautics.

[146] I. Iacovides, A. Cox, R. Kennedy, P. Cairns, and C. Jennett, “Removing the HUD: The Impact of Non-Diegetic Game Elements and Expertise on Player Involvement,” in *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, ser. CHI PLAY ’15. New York, NY, USA: Association for Computing Machinery, Oct. 2015, pp. 13–22.

[147] J. O’Hara and S. Fleger, “Human-System Interface Design Review Guidelines,” Tech. Rep. BNL–216211-2020-FORE, 1644018, Jul. 2020.

[148] “ARAC WG Report FAR/JAR 25.1322 & AC/ACJ 25.1322,” Federal Aviation Administration, Tech. Rep., Sep. 2004.

[149] C. D. Wickens, L. C. Thomas, and R. Young, “Frames of Reference for the Display of Battlefield Information: Judgment-Display Dependencies,” *Human Factors*, vol. 42, no. 4, pp. 660–675, Dec. 2000, publisher: SAGE Publications Inc.

[150] J. Wood, M. C. Margenet, P. Kenneally, H. Schaub, and S. Piggott, “Flexible Basilisk Astrodynamics Visualziation Software Using the Unity Rendering Engine,” in *AAS Guidance and Control Conference*, Breckenridge, CO, Feb. 2018.

[151] C. D. Wickens, “Cognitive Factors in Display Design,” *Journal of the Washington Academy of Sciences*, vol. 83, no. 4, pp. 179–201, 1993, publisher: Washington Academy of Sciences.

[152] H. A. Rosyid, A. Y. Pangestu, and M. I. Akbar, "Can Diegetic User Interface Improves Immersion in Role-Playing Games?" in *2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, Oct. 2021, pp. 200–204.

[153] L. Caroux and K. Isbister, "Influence of head-up displays' characteristics on user experience in video games," *International Journal of Human-Computer Studies*, vol. 87, pp. 65–79, Mar. 2016.

[154] J. F. Golding, "Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness," *Brain Research Bulletin*, vol. 47, no. 5, pp. 507–516, Nov. 1998.

[155] C. J. Hainley, K. R. Duda, C. M. Oman, and A. Natapoff, "Pilot Performance, Workload, and Situation Awareness During Lunar Landing Mode Transitions," *Journal of Spacecraft and Rockets*, vol. 50, no. 4, pp. 793–801, Jul. 2013, publisher: American Institute of Aeronautics and Astronautics.

[156] J. A. Karasinski, S. K. Robinson, K. R. Duda, and Z. Prasov, "Development of real-time performance metrics for manually-guided spacecraft operations," in *2016 IEEE Aerospace Conference*. Big Sky, MT, USA: IEEE, Mar. 2016, pp. 1–9.

[157] F. T. Durso, C. A. Hackworth, T. R. Truitt, J. Crutchfield, D. Nikolic, and C. A. Manning, "Situation Awareness as a Predictor of Performance for En Route Air Traffic Controllers," *Air Traffic Control Quarterly*, vol. 6, no. 1, pp. 1–20, Jan. 1998, publisher: American Institute of Aeronautics and Astronautics.

[158] F. T. Durso, C. A. Hackworth, and T. R. Truitt, "Situation awareness as a predictor of performance in en route air traffic controllers," 1999.

[159] S. Loft, D. Morrell, and S. Huf, "Using the situation present assessment method to measure situation awareness in simulated submarine track management," *International Journal of Human Factors and Ergonomics*, vol. 2, pp. 33–48, Dec. 2013.

[160] S. Loft, V. Bowden, J. Braithwaite, D. B. Morrell, S. Huf, and F. T. Durso, "Situation Awareness Measures for Simulated Submarine Track Management," *Human Factors*, vol. 57, no. 2, pp. 298–310, Mar. 2015, publisher: SAGE Publications Inc.

[161] J. C. Cunningham, H. Battiste, S. Curtis, E. C. Hallett, M. Koltz, S. L. Brandt, J. Lachter, V. Battiste, and W. W. Johnson, "Measuring Situation Awareness with Probe Questions: Reasons for not Answering the Probes," *Procedia Manufacturing*, vol. 3, pp. 2982–2989, Jan. 2015.

[162] T. Mirchi, K.-P. Vu, J. Miles, L. Sturre, S. Curtis, and T. Z. Strybel, "Air Traffic Controller Trust in Automation in NextGen," *Procedia Manufacturing*, vol. 3, pp. 2482–2488, 2015.

[163] M. Fujino, J. Lee, T. Hirano, Y. Saito, and M. Itoh, "Comparison of SAGAT and SPAM for Seeking Effective Way to Evaluate Situation Awareness and Workload During Air Traffic Control Task," in *Proceedings of the 2020 HFES 64th International Annual Meeting*, Virtual, 2020.

[164] J. Brooke, "SUS: A 'Quick and Dirty' Usability Scale," in *Usability Evaluation In Industry*. CRC Press, 1996, num Pages: 6.

[165] F. Durso, T. Truitt, C. Hackworth, J. Crutchfield, D. Ohrt, D. Nikolić, P. Moertl, and C. Manning, “Expertise and chess: A pilot study comparing situation awareness methodologies,” Jan. 1995.

[166] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting Linear Mixed-Effects Models Using lme4,” *Journal of Statistical Software*, vol. 67, pp. 1–48, Oct. 2015.

[167] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, “lmerTest Package: Tests in Linear Mixed Effects Models,” *Journal of Statistical Software*, vol. 82, pp. 1–26, Dec. 2017.

[168] R. V. Lenth, B. Bolker, P. Buerkner, I. Giné-Vázquez, M. Herve, M. Jung, J. Love, F. Miguez, H. Riebl, and H. Singmann, “emmeans: Estimated Marginal Means, aka Least-Squares Means,” Jun. 2023.

[169] M. S. Ben-Shachar, D. Lüdecke, and D. Makowski, “effectsize: Estimation of Effect Size Indices and Standardized Parameters,” *Journal of Open Source Software*, vol. 5, no. 56, p. 2815, Dec. 2020.

[170] D. L. Oltrogge and D. A. Vallado, “Debris Risk Evolution And Dispersal (DREAD) for post-fragmentation modeling,” in *2019 15th Hypervelocity Impact Symposium.* Destin, FL, USA: American Society of Mechanical Engineers, Apr. 2019, p. V001T10A009.

[171] C. D. Wickens, J. S. McCarley, A. Alexander, L. C. Thomas, M. Ambinder, and S. Zheng, “Attention-Situation Awareness (A-SA) Model of Pilot Error,” NASA Ames Research Center, Moffett Field, CA, Tech. Rep. AHFD-04-15/NASA-04-5, Jan. 2005.

[172] F. Lu, S. Davari, L. Lisle, Y. Li, and D. A. Bowman, “Glanceable AR: Evaluating Information Access Methods for Head-Worn Augmented Reality,” in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, Mar. 2020, pp. 930–939, iSSN: 2642-5254.

[173] L. Caroux, M. Delmas, M. Cahuzac, M. Ader, B. Gazagne, and A. Ravassa, “Head-up displays in action video games: the effects of physical and semantic characteristics on player performance and experience,” *Behaviour & Information Technology*, vol. 42, no. 10, pp. 1466–1486, Jul. 2023, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0144929X.2022.2081609.

[174] G. Rottermanner, V. A. de Jesus Oliveira, P. Lechner, P. Graf, M. Kreiger, M. Wagner, M. Iber, C.-H. Rokitansky, K. Eschbacher, V. Grantz, V. Settgast, and P. Judmaier, “Design and Evaluation of a Tool to Support Air Traffic Control with 2D and 3D Visualizations,” in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, Mar. 2020, pp. 885–892, iSSN: 2642-5254.

[175] B. M. Huey and C. D. Wickens, “Workload Transition: Implications for Individual and Team Performance,” National Research Council, Tech. Rep., 1993.

[176] R. Grier, “How high is high? A metanalysis of NASA TLX global workload scores,” vol. 59, Oct. 2015.

[177] H. Weiss, A. Liu, A. Byon, J. Blossom, and L. Stirling, “Comparison of Display Modality and Human-in-the-Loop Presence for On-Orbit Inspection of Spacecraft,” *Human Factors*, p. 00187208211042782, Sep. 2021, publisher: SAGE Publications Inc.

[178] C. W. Lejuez, J. P. Read, C. W. Kahler, J. B. Richards, S. E. Ramsey, G. L. Stuart, D. R. Strong, and R. A. Brown, "Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART)," *Journal of Experimental Psychology. Applied*, vol. 8, no. 2, pp. 75–84, Jun. 2002.

[179] M. Basner, A. Savitt, T. M. Moore, A. M. Port, S. McGuire, A. J. Ecker, J. Nasrini, D. J. Mollicone, C. M. Mott, T. McCann, D. F. Dinges, and R. C. Gur, "Development and Validation of the Cognition Test Battery for Spaceflight," *Aerospace Medicine and Human Performance*, vol. 86, no. 11, pp. 942–952, Nov. 2015.

[180] R. H. B. Christense, "ordinal—Regression Models for Ordinal Data," 2023.

[181] S. C. Reed, F. R. Levin, and S. M. Evans, "Alcohol increases impulsivity and abuse liability in heavy drinking women," *Experimental and Clinical Psychopharmacology*, vol. 20, no. 6, pp. 454–465, 2012, place: US Publisher: American Psychological Association.

[182] Maxime HERVE, "RVAideMemoire: Testing and Plotting Procedures for Biostatistics," Mar. 2011, institution: Comprehensive R Archive Network Pages: 0.9-83-7.

[183] Alexis Dinno, "dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums," Aug. 2014, institution: Comprehensive R Archive Network Pages: 1.3.6.

[184] M. R. Endsley, "The Divergence of Objective and Subjective Situation Awareness: A Meta-Analysis," *Journal of Cognitive Engineering and Decision Making*, vol. 14, no. 1, pp. 34–53, Mar. 2020, publisher: SAGE Publications.

[185] ——, "A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM," *Human Factors*, vol. 63, no. 1, pp. 124–150, Feb. 2021, publisher: SAGE Publications Inc.

[186] C.-J. Chao, S.-Y. Wu, Y.-J. Yau, W.-Y. Feng, and F.-Y. Tseng, "Effects of three-dimensional virtual reality and traditional training methods on mental workload and training performance," *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol. 27, no. 4, pp. 187–196, 2017, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hfm.20702.

[187] P. Millais, S. L. Jones, and R. Kelly, "Exploring Data in Virtual Reality: Comparisons with 2D Data Visualizations," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems.* Montreal QC Canada: ACM, Apr. 2018, pp. 1–6.

[188] L. C. Thomas and C. D. Wickens, "Visual Displays and Cognitive Tunneling: Frames of Reference Effects on Spatial Judgments and Change Detection," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 45, no. 4, pp. 336–340, Oct. 2001, publisher: SAGE Publications Inc.

[189] S.-C. Chen, M.-S. Hsiao, and H.-C. She, "The effects of static versus dynamic 3D representations on 10th grade students' atomic orbital mental model construction: Evidence from eye movement behaviors," *Computers in Human Behavior*, vol. 53, pp. 169–180, Dec. 2015.

[190] M. R. Endsley, "Predictive Utility of an Objective Measure of Situation Awareness," *Proceedings of the Human Factors Society Annual Meeting*, vol. 34, no. 1, pp. 41–45, Oct. 1990, publisher: SAGE Publications.

[191] ——, “Toward a Theory of Situation Awareness in Dynamic Systems,” *Human Factors*, vol. 37, no. 1, p. 33, 1996.

[192] D. G. Jones and M. R. Endsley, “Sources of situation awareness errors in aviation,” *Aviation, Space, and Environmental Medicine*, vol. 67, no. 6, pp. 507–512, Jun. 1996.

[193] F. T. Durso, T. R. Truitt, C. A. Hackworth, J. M. Crutchfield, and C. A. Manning, “En Route Operational Errors and Situational Awareness,” *The International Journal of Aviation Psychology*, vol. 8, no. 2, pp. 177–194, Apr. 1998.

[194] A. L. Baker, S. M. Fitzhugh, D. E. Forster, and K. E. Schaefer, “Communication Metrics for Human-Autonomy Teaming: Lessons Learned from us Army Gunnery Field Experiments,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, pp. 1157–1161, Sep. 2021, publisher: SAGE Publications Inc.

[195] E. Solberg, E. Nystad, and R. McDonald, “Situation awareness in outage work – A study of events occurring in U.S. nuclear power plants between 2016 and 2020,” *Safety Science*, vol. 158, p. 105965, Feb. 2023.

[196] L. M. Daling and S. J. Schlittmeier, “Effects of Augmented Reality-, Virtual Reality-, and Mixed Reality–Based Training on Objective Performance Measures and Subjective Evaluations in Manual Assembly Tasks: A Scoping Review,” *Human Factors*, vol. 66, no. 2, pp. 589–626, Feb. 2024, publisher: SAGE Publications Inc.

[197] A. D. Kaplan, J. Cruit, M. Endsley, S. M. Beers, B. D. Sawyer, and P. A. Hancock, “The Effects of Virtual Reality, Augmented Reality, and Mixed Reality as Training Enhancement Methods: A Meta-Analysis,” *Human Factors*, vol. 63, no. 4, pp. 706–726, Jun. 2021, publisher: SAGE Publications Inc.

[198] A. R. Selkowitz, S. G. Lakhmani, and J. Y. C. Chen, “Using agent transparency to support situation awareness of the Autonomous Squad Member,” *Cognitive Systems Research*, vol. 46, pp. 13–25, Dec. 2017.

[199] M. Al-Hor, H. Almahdi, M. Al-Theyab, A. G. Mustafa, M. Seed Ahmed, and S. Zaqout, “Exploring student perceptions on virtual reality in anatomy education: insights on enjoyment, effectiveness, and preferences,” *BMC Medical Education*, vol. 24, no. 1, p. 1405, Dec. 2024.

[200] Y. Chen, Y. He, X. Zou, H. Cai, H. H. E. Yiu, and W.-k. Ming, “Undergraduate nursing students’ preferences for virtual reality simulations in nursing skills training: A discrete choice experiment,” *Digital Health*, vol. 11, p. 20552076251339009, May 2025.

[201] J. Sung, S. Leary, V. S. Hurd, C. Lee, Y. Qin, Z. Kong, T. K. Clark, and A. Anderson, “Operationally Realistic Human-Autonomy Teaming Task Simulation to Study Multi-Dimensional Trust,” in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. Boulder CO USA: ACM, Mar. 2024, pp. 1028–1032.

[202] H. Chung and X. J. Yang, “Predicting Trust Dynamics With Personal Characteristics,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 68, no. 1, pp. 310–316, Sep. 2024.

[203] S. M. Merritt, J. L. Unnerstall, D. Lee, and K. Huber, “Measuring Individual Differences in the Perfect Automation Schema,” *Human Factors*, vol. 57, no. 5, pp. 740–753, Aug. 2015.

[204] S. M. Merritt, A. Ako-Brew, W. J. Bryant, A. Staley, M. McKenna, A. Leone, and L. Shirase, "Automation-Induced Complacency Potential: Development and Validation of a New Scale," *Frontiers in Psychology*, vol. 10, p. 225, 2019.

[205] S. M. Merritt, H. Heimbaugh, J. LaChapell, and D. Lee, "I Trust It, but I Don't Know Why: Effects of Implicit Attitudes Toward Automation on Trust in an Automated System," *Human Factors*, vol. 55, no. 3, pp. 520–534, Jun. 2013, publisher: SAGE Publications Inc.

[206] M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas, "The Mini-IPIP Scales: Tiny-yet-effective measures of the Big Five Factors of Personality." *Psychological Assessment*, vol. 18, no. 2, pp. 192–203, 2006.

[207] B. Yoo, D. Naveen, and T. and Lenartowicz, "Measuring Hofstede's Five Dimensions of Cultural Values at the Individual Level: Development and Validation of CVSCALE," *Journal of International Consumer Marketing*, vol. 23, no. 3-4, pp. 193–210, May 2011.

[208] Venkatesh, Morris, Davis, and Davis, "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly*, vol. 27, no. 3, p. 425, 2003.

[209] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an Empirically Determined Scale of Trust in Automated Systems," *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53–71, Mar. 2000, publisher: Routledge _eprint: https://doi.org/10.1207/S15327566IJCE0401_04.

[210] N. C. Anderson, W. F. Bischof, K. E. W. Laidlaw, E. F. Risko, and A. Kingstone, "Recurrence quantification analysis of eye movements," *Behavior Research Methods*, vol. 45, no. 3, pp. 842–856, Sep. 2013.

[211] J. Z. Bakdash and L. R. Marusich, "Repeated Measures Correlation," *Frontiers in Psychology*, vol. 8, Apr. 2017, publisher: Frontiers.

[212] K. Geels-Blair, R. Stephen, and J. Schwark, "Using System-Wide Trust Theory to Reveal the Contagion Effects of Automation False Alarms and Misses on Compliance and Reliance in a Simulated Aviation Task," *The International Journal of Aviation Psychology*, vol. 23, no. 3, pp. 245–266, Jul. 2013, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10508414.2013.799355.

[213] F. Abed, "Cultural Influences on Visual Scanning Patterns," *Journal of Cross-Cultural Psychology*, vol. 22, no. 4, pp. 525–534, Dec. 1991.

[214] J. M. Henderson, "Human gaze control during real-world scene perception," *Trends in Cognitive Sciences*, vol. 7, no. 11, pp. 498–504, Nov. 2003.

[215] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends in Cognitive Sciences*, vol. 9, no. 4, pp. 188–194, Apr. 2005.

[216] T. Rieger and D. Manzey, "Understanding the Impact of Time Pressure and Automation Support in a Visual Search Task," *Human Factors*, vol. 66, no. 3, pp. 770–786, Mar. 2024, publisher: SAGE Publications Inc.

[217] M. Schulte-Mecklenbeck, A. Kühberger, and R. Ranyard, “The role of process data in the development and testing of process models of judgment and decision making,” *Judgment and Decision Making*, vol. 6, no. 8, pp. 733–739, Dec. 2011.

[218] X. P. Kotval and J. H. Goldberg, “Eye Movements and Interface Component Grouping: An Evaluation Method,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 42, no. 5, pp. 486–490, Oct. 1998.

[219] Y. Lee, K.-T. Jung, and H.-C. Lee, “Use of gaze entropy to evaluate situation awareness in emergency accident situations of nuclear power plant,” *Nuclear Engineering and Technology*, vol. 54, no. 4, pp. 1261–1270, Apr. 2022.

[220] J. J. Hendrickson, “Performance, preference, and visual scan patterns on a menu-based system: implications for interface design,” in *Proceedings of the SIGCHI conference on Human factors in computing systems Wings for the mind - CHI ’89.* ACM Press, 1989, pp. 217–222.

[221] P. Kearney, L. Wen-Chin, Y. Chung-San, and G. Braithwaite, “The impact of alerting designs on air traffic controller’s eye movement patterns and situation awareness,” *Ergonomics*, vol. 62, no. 2, pp. 305–318, Feb. 2019, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00140139.2018.1493151.

[222] S. L. Buchner, J. R. Kintz, J. Y. Zhang, N. T. Banerjee, T. K. Clark, and A. P. Hayman, “Assessing Physiological Signal Utility and Sensor Burden in Estimating Trust, Situation Awareness, and Mental Workload,” *Journal of Cognitive Engineering and Decision Making*, vol. 19, no. 2, pp. 154–173, Jun. 2025, publisher: SAGE Publications.

[223] J. R. Kintz, N. T. Banerjee, J. Y. Zhang, A. P. Anderson, and T. K. Clark, “Estimation of Subjectively Reported Trust, Mental Workload, and Situation Awareness Using Unobtrusive Measures,” *Human Factors*, Nov. 2022, publisher: SAGE Publications Inc.

[224] R. Pak, F. Nicole, P. Margaux, B. Brock, and L. and Sturre, “Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults,” *Ergonomics*, vol. 55, no. 9, pp. 1059–1072, Sep. 2012.

[225] J. Ayoub, L. Avetisyan, M. Makki, and F. Zhou, “An Investigation of Drivers’ Dynamic Situational Trust in Conditionally Automated Driving,” *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 3, pp. 501–511, Jun. 2022.

[226] A. O. Rindfuss, “Modeling trust and trust dynamics from physiological signals for operational supervisory control during human-autonomy teaming,” Master’s thesis, University of Colorado at Boulder, 2025.

[227] S. Leary, Y. Qin, Z. Kong, T. Clark, and A. Anderson, “Validating Rapid Trust Measurements in Spaceflight-Relevant Human-Autonomy Teaming Applications,” in *IAF Human Spaceflight Symposium.* Milan, Italy: International Astronautical Federation (IAF), 2024, pp. 583–587.

Appendix A: Detailed Results

A.1 Aim 2

Table A.1: Summary of Metrics including mean and standard deviation for each condition.

	Baseline		Scr. Viz.		VR	
	M	SD	M	SD	M	SD
Level 1 SA [%]	83.39	5.40	78.81	6.50	71.82	5.31
Level 1 SA - Satellite States [%]	78.24	11.93	71.37	5.55	5992	9.02
Level 1 SA - Time to Event [%]	86.59	5.39	83.70	9.99	79.68	5.43
Level 2 SA [%]	78.35	12.84	89.61	8.99	87.46	7.15
Level 3 SA [%]	74.37	6.74	82.19	4.25	81.38	7.81
Workload	40.25	12.87	34.43	12.67	37.12	15.04
Usability	59.09	26.68	69.54	13.77	61.81	12.10

Table A.2: Summary of the subjective utility questionnaires results, including the median for each condition.

	Baseline	Scr. Viz.	VR	$\chi^2(2)$	p-value
Servicer Uncertainty	Neutral	Agree	Agree	5.34	0.069
Debris Uncertainty	Agree	Agree	Agree	2.43	0.30
Collision Likelihood	Agree	Agree	Agree	2.29	0.32
Easy to Find information	Agree	Agree	Agree	0.85	0.65
Event Awareness	Agree	Strongly Agree	Agree	0.41	0.82

A.2 Aim 3

Table A.3: Summary of Metrics including mean and standard deviation for each condition. The performance data is on a Likert scale so the median is reported.

	Baseline		Scr. Viz.		VR	
	M	SD	M	SD	M	SD
Level 1 SA [%]	90.24	10.27	95.00	4.89	95.27	6.29
Level 2 SA [%]	79.47	10.73	78.18	10.27	90.22	6.28
Level 3 SA [%]	70.93	10.95	80.42	10.22	77.69	11.55
Performance Easy	13	-	12	-	12	-
Performance Medium	9	-	9	-	8	-
Performance Hard	6	-	8	-	8	-
Workload	33.13	19.09	32.92	17.05	32.12	21.28
Usability	57.83	18.41	66.67	13.62	67.50	13.73

Table A.4: Summary of the subjective utility questionnaires results, including the median for each condition.

	Baseline	Scr. Viz.	VR	$\chi^2(2)$	p-value
Easy to Find Info	Agree	Agree	Strongly Agree	3.37	0.18
Event Awareness	Agree	Strongly Agree	Strongly Agree	2.98	0.23
Uncertainties	Agree	Agree	Agree	3.30	0.19
Collision Likelihood	Agree	Strongly Agree	Agree	3.48	0.18
Orbital Motion	Neutral	Agree	Strongly Agree	14.58	< 0.005
Operational Decisions	Agree	Agree	Agree	6.11	0.047

A.3 Aim 4

Table A.5: Summary of Metrics including mean and standard deviation for each condition. The performance data is on a Likert scale so the median is reported.

	Baseline		Scr. Viz.		VR	
	M	SD	M	SD	M	SD
Level 1 SA [%]	94.53	6.42	92.37	4.74	96.69	4.35
Level 2 SA [%]	80.71	9.13	83.58	9.45	90.22	6.27
Level 3 SA [%]	77.24	11.68	79.11	9.49	81.80	8.28
Performance Easy	13	-	13	-	13	-
Performance Medium	9	-	11	-	9	-
Performance Hard	8	-	9	-	8	-
Workload	21.31	16.66	23.26	17.43	22.44	19.00
Usability	61.17	19.84	61.83	16.41	79.00	16.58

Table A.6: Summary of the subjective utility questionnaires results, including the median for each training condition.

	Baseline	Scr. Viz.	VR	$\chi^2(2)$	p-value
Event Awareness	Strongly Agree	Strongly Agree	Strongly Agree	7.46	0.024
Uncertainties	Agree	Strongly Agree	Strongly Agree	5.88	0.052
Collision Likelihood	Agree	Strongly Agree	Strongly Agree	7.62	0.022
Orbital Motion	Neutral	Strongly Agree	Strongly Agree	6.40	0.041
Operational Decisions	Strongly Agree	Strongly Agree	Strongly Agree	0.71	0.70

A.4 Aim 5

Table A.7: Top 10 gaze metrics by mutual information. *Sys.Rec. denotes an interaction with that feature and the system recommendation being important.

All data		Recommend Troop		Recommend No Troop	
Dur * Sys.Rec.	0.027	Dur	0.049	Total Dur on Buttons AOI	0.037
Entropy	0.025	Entropy	0.033	Switches from Sys.Rec. to Viz. AOI	0.030
Dur	0.0023	Relative Entropy	0.032	Switches from Viz. to Sys.Rec. AOI	0.027
Relative Entropy	0.023	Total Switches Between AOIs	0.030	Percent Recurrence on AOI	0.028
Number of recurrence	0.020	Laminarity	0.029	Num. Reviews	0.027
Det. with AOI	0.019	Total Dur on Viz. AOI	0.028	Switches from Sys.Rec. to Th. AOI	0.027
Laminarity*Aut.Sys.Rec	0.019	Num. recurrence	0.028	Relative Entropy	0.027
Entropy*Sys.Rec	0.017	Percent Recurrence on AOI	0.023	Switches from C&DH to Th. AOI	0.023
Relative entropy*Sys.Rec	0.017	Num. Fixations of Viz. AOI	0.024	Switches from Viz. to Th. to C&DH AOI	0.019
Det. with AOI*Aut.Sys.Rec	0.016	Num. recurrence on AOI	0.021	Switches from Sys.Rec. to Button AOI	0.017

Note: AOI = area of interest; Viz = visual AOI; Th = Thermal AOI; C&DH = Command and Data Handling; Sys. Rec. = autonomous system recommendation; Dur = Duration; num = Number of; Det = determinism

Table A.8: Top 10 gaze metrics by feature importance for decision accuracy and their feature importance score. *Sys.Rec. denotes an interaction with that feature and the system recommendation being important. Feature importance is out of a percent of all possible features available for the model

All data		Recommend Troop		Recommend No Troop	
Duration	0.120	Duration	0.152	Number of Reviews	0.11
Duration * Sys.Rec.	0.055	Number of recurrence	0.091	Duration	0.084
Number of recurrence	0.054	Relative Entropy	0.084	Total Duration on Visual AOI	0.052
Relative Entropy	0.037	Total Duration on Visual AOI	0.066	Relative Entropy	0.046
Total Duration on Visual AOI	0.033	Entropy	0.061	Total Switches Between AOIs	0.042
Entropy	0.029	Total Duration on Thermal AOI	0.048	Number of recurrence	0.035
Total Duration on Visual AOI*Sys.Rec	0.028	Number of Fixations of Visual AOI	0.033	Total Duration on Thermal AOI	0.032
Entropy*Sys.Rec	0.028	Total Switches Between AOIs	0.030	Total Duration on Buttons AOI	0.032
Number of recurrence*SysRec	0.027	Total Duration on C&DH AOI	0.028	Percent of Time of Visual Screen	0.032
Total Switches Between AOIs	0.022	Total Duration on Sys.Rec. AOI	0.026	Entropy	0.03

Table A.9: Top 10 gaze metrics by feature importance for trust.

Trust value		Trust Difference	
Duration of Time on Buttons AOI	0.089	Number of Fixations of Visual AOI	0.052
Percent of Time on C&DH AOI	0.083	Number of Fixations of C&DH AOI	0.0456
Total Switches Between AOIs	0.079	Percent of Time on Thermal AOI	0.0419
Duration of Time on Thermal AOI	0.079	Percent of Time on Visual AOI	0.037
Duration of Time on C&DH AOI	0.072	Switches from Buttons to Thermal AOI	0.036
Percent of Time on Thermal AOI	0.072	Duration of Time on Thermal AOI	0.036
Number of Fixations of C&DH AOI	0.068	Percent Recurrence on AOI	0.035
Number of Fixations of Thermal AOI	0.060	Duration of Time on C&DH AOI	0.034
Number of Fixations of Buttons AOI	0.060	Switches from Visual to C&DH AOI	0.033
Duration of Time on Sys.Rec AOI	0.060	Number of Fixations of Sys. Rec. AOI	0.032

Table A.10: Top 10 gaze metrics by feature importance for trust.

Trust value		Trust Difference	
Duration of Time on Buttons AOI	0.265	Duration	0.249
Number of Fixations of Thermal AOI	0.137	Percent of Time on Visual AOI	0.068
Number of Switches from Buttons to Visual	0.088	Center of Recurrence Mass	0.066
System Explainability	0.088	Number of Fixations of Thermal AOI	0.046
Number of Fixations of Buttons AOI	0.077	Maximum line length in RQA plot	0.037
Total Switches Between AOIs	0.074	Percent Recurrence	0.035
Percent of Time on Thermal AOI	0.046	Number of recurrenc	0.033
System Confidence	0.045	Relative Entropy	0.029
Duration	0.037	Determinism	0.027
Percent of time on C&DH AOI	0.024	cluster	0.026

Appendix B: Aim 2: Custom Surveys and Questionnaires

The following surveys are questionnaires were created for the Aim 2 experiment. In addition, the System Usability Scale [164] and NASA TLX [35] were used.

B.1 Demographics

Please fill out this questionnaire so that we can obtain some information about your level of experience and demographics.

What is your sex?
○Male
○Female
○Non-binary/third gender
○Prefer not to say

What is your age? (years)

How much sleep did you get last night? (hours)

Have you had alcohol in the past 6 hours?
○No
○Yes

What is the highest level of education you have completed?
○High school
○Some undergraduate education (not complete)
○Undergraduate degree
○Some graduate education (not complete)
○Graduate degree

What kind of content have you experienced in virtual reality?
□Traditional Media (2D movies, TV, Video)
□3D/360 Media Gaming in VR
□Simulation (flight/driving, Google Earth)
□Product or industrial visualization (ex. CAD)
□Previous human subject experiments
□None
□Other

How would you describe your familiarity with virtual reality environments?
○Little to no experience
○Moderate to high experience

Do you have experience with orbital mechanics?
□Currently in a class
□Have completed a class
□Have experience through work/design team.
□Have experience through video games (e.g., KSP)
□None

□Other

How would you describe your familiarity with orbital mechanics?
○Little to no experience
○Moderate to high experience

Do you have any experience with spacecraft operations?
○Yes
○No

How would you describe your familiarity with spacecraft operations?
○Little to no experience
○Moderate to high experience

B.2 SPAM queries

The following lists the possible SPAM queries asked to operators. Queries were randomly selected from this list, however, queries were only allowed to be selected in times were they were appropriate to the situation (e.g. "Will you receive new information before the next collision?" would only be asked if there was a collision potential.) Queries with a '/' indicate that either option may be selected.

Level 2:

- Is the next sensor update scheduled to occur before/after the next potential collision?
- Is the servicer satellite approaching/going away from the debris in all 3 axis?
- Is the servicer satellite going faster/slower than the debris?
- Is the servicer battery level currently increasing/decreasing?
- Is the debris/servicer satellite tumbling?
- Is there currently enough fuel to complete a burn?
- Is the servicer currently in the sun/Earth's shadow?
- Is the uncertainty associated with the servicer's orbit currently increasing/constant?
- Is the servicer satellite approaching/going away from the debris in the out of plane direction?

Level 3:

- Is there a risk of the servicer running out of fuel/battery before the next action?
- Will you meet the criteria to perform an mission abort at the time of the next action?
- Will you have enough fuel at the time of the next burn to complete the burn?
- If there is not a burn, will the servicer cross orbits with the debris satellite?
- Will the servicer's next crossing in the along track direction be in front/behind of the debris?
- Will the servicer's next crossing in the radial direction be closer to/further from Earth than debris?
- Will there be sufficient battery to complete the next hour of the mission?
- Is there a high, medium, or low likelihood of a future collision?
- Will you receive new information before the next potential collision?

B.3 Utility

Please read EACH of the following statements, and indicate the extent to which you agree with each:

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I found this system enabled me to understand the uncertainty associated with the servicer.	○	○	○	○	○
I found this system enabled me to understand the uncertainty associated with the debris.	○	○	○	○	○
I found this system promoted my understanding of collision likelihood.	○	○	○	○	○
I found the necessary information easy to find.	○	○	○	○	○
I felt that I was aware of mission critical events.	○	○	○	○	○

Appendix C: Aims 3 and 4: Custom Surveys and Questionnaires

The following surveys are questionnaires were created for the Aims 3 and 4 experiment. In addition, the System Usability Scale [164] and NASA TLX [35] were used.

C.1 Demographics

Please fill out this questionnaire so that we can obtain some information about your level of experience and demographics.

How much sleep did you get last night? (hours)

Have you had alcohol in the past 6 hours?
○No
○Yes

What sex were you assigned at birth, on your original birth certificate?
○Male
○Female
○Don't know
○Prefer not to say

What is your current gender?
○Male
○Female
○Non-binary/third gender
○I identify as:
○Prefer not to say

What is your age? (years)

What is the highest level of education you have completed?
○High school
○Some undergraduate education (not complete)
○Undergraduate degree
○Some graduate education (not complete)
○Graduate degree

What kind of content have you experienced in virtual reality?
□Traditional Media (2D movies, TV, Video)
□3D/360 Media Gaming in VR
□Simulation (flight/driving, Google Earth)
□Product or industrial visualization (ex. CAD)
□Previous human subject experiments
□None
□Other

How would you describe your familiarity with virtual reality environments?
○Little to no experience

○Moderate to high experience

Do you have experience with orbital mechanics?
□Currently in a class
□Have completed a class
□Have experience through work/design team.
□Have experience through video games (e.g., KSP)
□None
□Other

How would you describe your familiarity with orbital mechanics?
○Little to no experience
○Moderate to high experience

Do you have any experience with spacecraft operations?
□Military Work □Commercial Space Work □Research □None
□Other

How would you describe your familiarity with spacecraft operations?
○Little to no experience
○Moderate to high experience

C.2 SPAM queries

The following lists the possible SPAM queries asked to operators. Queries were randomly selected from this list, however, queries were only allowed to be selected in times where they were appropriate to the situation (e.g., "Is the next planned burn more/less than 30 minutes away?" would only be asked if there was a planned burn.) Queries with a '/' indicate that either option may be selected.

Level 1:

- Is the servicer's battery greater/less than 50%?
- Is the servicer's fuel greater/less than 50%?
- Is the servicer currently in the sunlight/shadow?
- Is the servicer's flashlight on/off?
- Are there currently any cautions/warnings?
- Is the next planned burn more/less than 30 minutes away?
- Is there more/less than 30 minutes until the lighting conditions change?
- Is the servicer above/below the client satellite?
- Is the servicer ahead/behind of the client satellite?
- Is there a potential collision in the next 30 minutes?
- Does the servicer's current trajectory enter the keep out zone?

Level 2:

- Is the servicer's battery level currently increasing/decreasing?
- Is the keep out zone currently increasing/decreasing in size?

- Is the servicer's flashlight currently aimed towards the client?
- Is there currently enough battery/fuel/time to complete an abort burn?
- Does the orbit enter the keep out zone from ahead/behind of the client satellite?
- Does the orbit enter the keep out zone from above/below the client satellite?
- Is the time to collision increasing/decreasing?
- Is the portion of the orbit line in the keep out zone increasing/decreasing?

Level 3:

- If no new unplanned actions are taken will there be a collision alert in 30 minutes?
- If no new unplanned actions are taken, will there be enough battery/fuel to complete an abort burn in 30 minutes?
- If no burns occur, will the servicer enter the keep out zone in the next 30 minutes?
- If there is not a burn, will the servicer cross orbits with the debris satellite?
- If no new unplanned actions are taken, will the servicer's battery be greater/less than 50% in 30 minutes?
- If no action is taken, will you receive a new low battery/fuel caution/warning within the next 30 minutes?
- If no new action is taken, will the keep out zone be growing/shrinking in 30 minutes?
- If no new unplanned action is taken, will the servicer's battery be increasing/decreasing in 30 minutes?
- Will the servicer be in the sun/shadow at the early/default/late burn location?

C.3 Utility

Please read EACH of the following statements, and indicate the extent to which you agree with each. The system refers to the display that you worked with today.

	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
I found the necessary information easy to find in this system.	○	○	○	○	○
I found this system allowed me to be aware of mission critical events.	○	○	○	○	○
I found this system enabled me to understand the uncertainties associated with the mission.	○	○	○	○	○
I found this system promoted my understanding of collision likelihood.	○	○	○	○	○
I found this system enabled me to understand the relative orbital motion of the satellites.	○	○	○	○	○
I found this system allowed me to make appropriate operation decisions.	○	○	○	○	○

Please provide any additional comments on this display and your experience performing the task.

Please read EACH of the following statements, and indicate the extent to which you agree with each. The training from the first visit refers to your previous visit.

	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
I found that the training from the first visit was effective in enabling me to understand mission critical events today.	○	○	○	○	○
I found that the training from the first visit was effective in enabling me to understand the uncertainties today.	○	○	○	○	○
I found that the training from the first visit was effective in enabling my understanding of collision likelihood today.	○	○	○	○	○
I found that the training from the first visit was effective in enabling me to understand the relative orbital motion of the satellites today.	○	○	○	○	○
I found that the training from the first visit was effective in enabling me to make appropriate operation decisions today.	○	○	○	○	○

Please provide any additional comments about how your training the first day impacted your performance today.

Did you prefer the display from visit 1 or 2? Why?

C.4 Performance

Participants' performance score on each trial was composed of two parts: burn decision and satellite end state. Burn decision relates to how effective the participants' burn selection enables overall mission success. During training, participants were instructed to always select a burn location, so no burn selection is an indication of either poor SA or poor understanding of appropriate actions. A 'Poor' burn decision consists of a burn location where it would be impossible to service the satellite. 'Fair' and 'Good' burn decisions are locations where servicing a satellite was possible, but different burn decisions would be more effective. 'Excellent' burn decisions were selections that put the participant in the best position to service the client. Not every scenario has all possible outcomes and for each trial, the location of a 'Poor', 'Fair', 'Good', or 'Excellent' may differ. The satellite end state refers to the final serviced state (collided, aborted, or serviced). The granularity of the serviced state was added by evaluating participants' abort attempts, time to collision when aborted, use of flashlight, and end battery level. These different actions and metrics provide indirect information about the participants' understanding of the task, awareness of the environment, and optimization of the outcome. Like the burn decision, not every trial had the full spectrum of possible scores for the satellite end state. The granularity of the scores is intended to objectively evaluate subject actions. These performance subdivisions allow for the delineation of performance but can be generalizable across scenarios. It should be noted that all scenarios cannot achieve all of the performance values.

Table C.3: Burn Decision

Outcome	Score
No Selection	1
Poor	2
Fair or Good	3
Excellent	4

Table C.4: End State

Outcome	Score
Collision - No Abort Attempt	1
Collision - Abort Attempt	2
Abort with $>$ 15 minutes to collision	3
Serviced with $<$ 5 minutes to collision or $<$ 25 % battery	4
Abort - light use $<$ 75%	5
Abort - light use $>$ 75%	6
Serviced - Battery $<$ 30 %	7
Serviced - Battery 30-50 %	8
Serviced - Battery $>$ 50 %	9

$$\text{Total Performance} = \text{End State(battery, outcome)} + \text{ Burn Decision}$$

Appendix D: Aim 5: Gaze Metric Definitions

The following gaze metrics were found in the models:

- Duration: Duration spent reviewing the recommendation
- Total Duration on AOI: Sum of all the fixation durations on that AOI
- Switches from AOI_1 to AOI_2: Number of times participant's fixation went from AOI_1 to AOI_2. In a unidirectional manner
- Total Switches: Total times the participant switched fixations between AOIs
- Number of fixations on AOI: Number of times the participant fixated on an AOI
- Number of Reviews: Number of times the participant opened the review screen for a particular recommendation
- Number of Recurrence: $R = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} r_{ij}$ The sum of recurrences
- Percent Recurrence: $REC = 100\frac{2R}{N(N-1)}$, represents for a sequence of N fixations the percentage of recurrent fixation
- Entropy: $ENT = -\sum(plog_2(p))$, where p is the probability of an event. Shannon entropy of the system. This is maximized when the distribution of a set of observations is uniformly distributed and is 0 when all observations share the same value.
- Relative Entropy: $RelENT = \frac{ENT}{log_2(MaxLine - minLine + 1)}$, where MaxLine is the maximum length of a diagonal, and minLine is the minimum length of a diagonal.
- Determinism: $DET = 100\frac{|D_L|}{R}$, where $|D_L|$ is the set of diagonal lines. The proportion of recurrent points forming diagonal lines and represents repeating gaze patterns
- Laminarity: $LAM = 100\frac{|H_L| + |V_L|}{2R}$, where H_L is the set of horizontal lines in the RQA plot and represents areas first scanned in detail and refixated briefly on later in time, and V_L is the set of verticle lines that area fixated first in a single fixation and rescanned over consecutive fixations at a later time. In general, laminarity indicates that specific areas of a scene are repeatedly fixed.
- Cluster: Number of recurrence clusters normalized by size of recurrence triangle