“It's One Thing After Another”: Supporting the Management of Frequent and Nested Interruptions in High-Tempo Multi-Agent Operations

by

Karanvir Singh Panesar

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2023

Doctoral Committee:

Professor Nadine Sarter, Chair
Professor Ella Atkins, Virginia Tech
Associate Professor Leia Stirling
Associate Professor Xi Jessie Yang

“Chance favors only the prepared mind.”

Pasteur, Louis (1854)

Karanvir Singh Panesar

kpanesar@umich.edu

ORCID iD: 0000-0002-9707-756X

Dedication

To mamma and papa.

Acknowledgements

I would like to thank the many people without whom this dissertation would not have been realized. First, I would like to thank my advisor Dr. Nadine Sarter for the unwavering support and endless patience these past five years. She has taught me how to think critically, explore alternate perspectives, and be a better researcher. I have learned from her a sense of uncompromised academic honesty, drive, and excellence. Not least of all, her love and passion for aviation is one that is contagious—I have, now, an appreciation and interest in aviation and aviation safety that is far stronger than I did five years ago.

I would also like to thank my committee members Dr. Ella Atkins, Dr. Leia Stirling, and Dr. Xi Jessie Yang. Both inside and outside the classroom, they have been instrumental in motivating and guiding this research in the right direction. I am grateful for their constructive feedback, challenging questions, and valuable support throughout the course of this dissertation.

I would also like to acknowledge the Rackham Graduate School, the Industrial and Operations Engineering (IOE) department, the Federal Aviation Administration, and the Air Force Office of Scientific Research. I am grateful to each of them for providing the means and funds for me to conduct research during my time at UM.

I would like to extend my thanks to the wonderful staff in the IOE department. Thank you to Olof Minto (Mint), Christopher Konrad, and Edwin Tang for the countless times I needed help with setting up computer workstations and figuring out solutions to obscure software problems. Thank you to Paul Pridham for help with building new components for the flight simulator and being a patient listener. I'm also very thankful to Chuck Woolley, Elizabeth Michalski, Matt Irelan, Rodd Capps, Tina Picano Sroka, Teresa Maldonado, and Sherri Cranson for their assistance on many occasions. Their willingness to help and their speedy responses were invaluable in keeping everything running smoothly and made me feel at home in the IOE department. Special thanks to Sheryl Ulin for her generous support and advice with planning events, and for the memorable conversations in hallways and words of encouragement. I would also like to thank all the College of Engineering undergraduate and graduate students for their willingness and patience to participate in my experiments, despite the 3.5-hour time commitment.

I am appreciative of all the THInC lab members for making my PhD journey easier. Thank you to Yuzhi Wan, Yidu Lu, Kevin Lieberman, Robert Thomas, Emma Faulhaber, Akshay Bhardwaj, and Hannah Baez for creating a welcoming lab environment and being a constant source of inspiration, support, and encouragement. I am grateful to have had the chance to learn from and work alongside such an amazing and talented group of friends. I am also extremely thankful for the undergraduate and graduate student programmers and researchers—Vikram Mathew, Christopher Lee, Rohith Kesamneni, Kevin Wang, Hanna Dong, and Madison Heyer—who helped with simulation development and data analysis, often under tight deadlines.

I would like to thank the amazing group of friends and fellow students—Kamolnat Tabattanon, Yifan Li, Teerachart Soratana, Patrik Schuler, and Daniel Felipe Otero-Leon —who were always there when needed and made my experience at UM both memorable and enjoyable. A special thanks to Kati Moug and Ypsilanti Rugby for introducing me to the wonderful sport of Rugby and giving me some of my favorite memories in Michigan.

Last but not least, I would like to thank my mom, dad, and sister for their love and encouragement. This dissertation would not have been possible without them by my side.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

The growing number of semi-autonomous machine agents in many safety-critical application domains brings with it an increase in the frequency with which human supervisory controllers need to interrupt ongoing tasks and lines of reasoning to handle unexpected and/or time-critical problems and requests. These disruptions may occur at inopportune times, such as being interrupted again immediately after completing a previous interrupting task (serial interruption) or while still handling a previous interruption (nested interruption). Frequent interruptions can lead to errors and delays which threaten safety in time-sensitive event-driven domains such as aviation and medicine. Successful teaming of operators with multiple machine agents therefore requires a better understanding of, and support for attention allocation and interruption management (IM). The goals of this dissertation were to 1) identify and analyze the challenges that operators encounter at various stages of interruption management (IM) when handling frequent and nested interruptions, and 2) develop and evaluate a set of candidate displays to address the observed challenges.

To this end, three human-subject experiments were conducted. The first two experiments focused on identifying the difficulties faced by operators when detecting, interpreting, and switching between frequent and nested interruptions in a supervisory command and control task. Frequent and nested interruption notifications were less likely to be acknowledged, compared to less frequent and non-nested ones. Participants also struggled with the appropriate scheduling of incoming tasks and took longer to switch to nested interrupting tasks of higher urgency, compared to both single and serial interruptions. The longer switch time resulted from delays at

the earlier stages of detection and interpretation of notifications as well as a resistance to switch away from the ongoing task, even for highly urgent interrupting tasks. In the third and final study, two candidate displays were developed and tested to address issues with poor scheduling of pending tasks. The first display involved automatic sorting of incoming task notifications by level of urgency; the second candidate consisted of a color- and location-based visualization of the relative urgency levels for the ongoing and interrupting task to support task prioritization and switching. The visualization of relative task urgency improved overall performance, led to decision-making accuracy, and resulted in more efficient prioritization of ongoing and interrupting tasks. At the same time, it involved a greater risk of failing to notice misclassifications made by imperfectly reliable automation.

The theoretical contribution of this research is a better understanding of the process of interruption management. More specifically, challenges and performance breakdowns experienced in the detection, interpretation, and integration of frequent and nested interruptions were identified. In contrast to what is suggested by current IM models, our findings show that interruption management is not a linear process, but one where behavior and performance at one stage depends on anticipated and experienced difficulties at both earlier and subsequent stages. In addition to identifying and analyzing challenges with handling frequent interruptions, this work also addresses said challenges and provides empirically based guidance on the design of interruption-resilient interfaces. From an applied perspective, findings from this line of work will help reduce the attentional demands and improve the safety and performance of human-machine teams, and the well-being of human operators in a variety of complex event-driven application domains like aviation and healthcare.

Chapter 1 – Introduction

The growing number of semi-autonomous machine agents in many application domains brings with it an increase in the frequency with which human supervisory controllers need to interrupt ongoing tasks and lines of reasoning to handle unexpected and/or time-critical problems and requests. For example, emerging forms of air-based transportation such as Urban Air Mobility (UAM; see **Figure 1.1**) are expected to accommodate up to 1200 pilotless aircraft operating simultaneously (FAA, 2020; Mueller et al., 2017; for more information, see **Table 1.1**). Managing such large-scale operations will require a few ground-based operators to maintain awareness of the status and behavior of large numbers of heterogeneous vehicles. This concept of operations is referred to as Single-Operator Multi-Agent (SOMA).

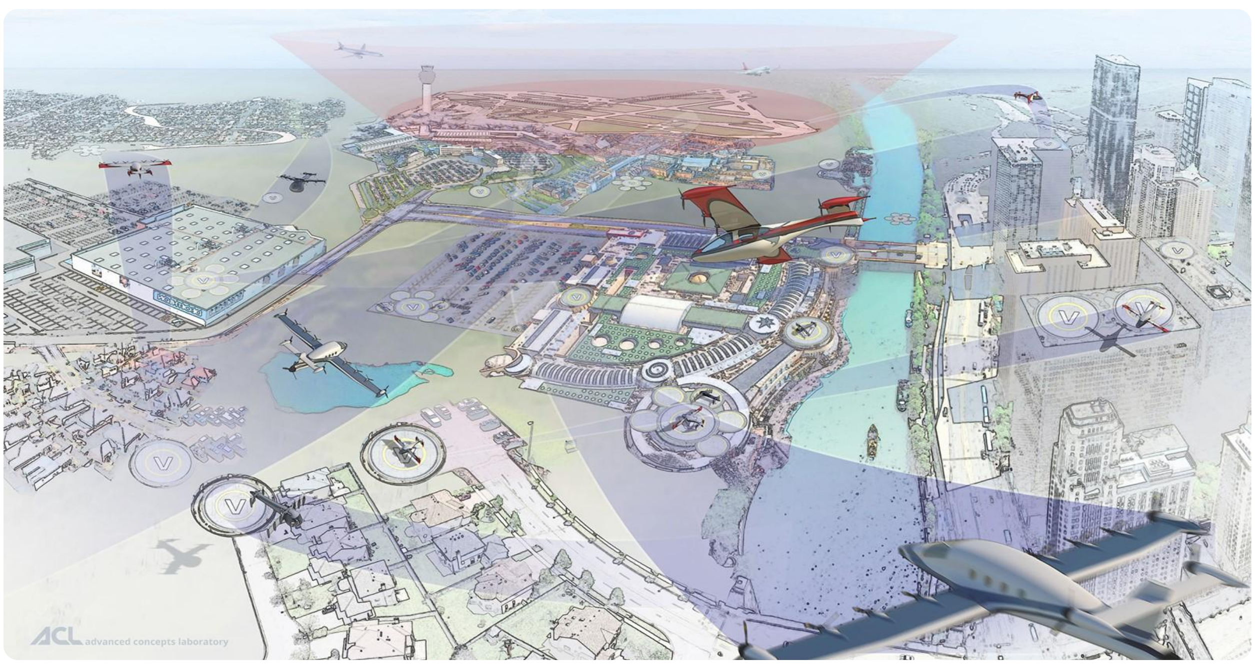

Figure 1.1 Envisioned concept of operations in Urban Air Mobility. Photo Credit: NASA Advanced Concepts Laboratory

Table 1.1 Overview of the Urban Air Mobility concept.

Urban Air Mobility
With the rapid development of electric Vertical Takeoff and Landing (eVTOL) aircraft that are designed to be small, quiet, and easy-to-fly, the aviation industry is nearing a paradigm shift in urban travel and transportation. The concept of Advanced Aerial Mobility (AAM) constitutes a push towards the use of small- and medium-sized pilotless eVTOL aircraft that can travel within and between urban, exurban, and rural areas (National Academies of Sciences, Engineering, and Medicine, 2020). Envisioned concepts of operations in AAM include on-demand and scheduled flight operations like swarm-based search and rescue, rapid aerial disaster response, package delivery, and transportation of people within and between regions. Urban Air Mobility (UAM) is a subset of the AAM concept (Hill et al., 2020; Holden & Goel, 2016; Lascara et al., 2019). UAM aims to make short-distance travel *within* cities and metropolitan areas easier, faster, and more affordable. Elevating city traffic from primarily road-based two-dimensional movement into a three dimensional space with aerial vehicles is expected to save billions of hours that are lost as a result of idling in traffic around busy downtown areas (Schrank et al., 2019). The target of UAM is to allow for the transportation of hundreds of thousands of people across a metropolitan area in pilotless aircraft with minimal to no ground-based supervision within the next two decades (~2040). Introducing fully autonomous air-based operations into the National Airspace System over congested cities like Los Angeles and New York while still achieving an equivalent or greater level of safety and public acceptance as current operations in commercial aviation will require fundamental changes to the design of avionics systems and unprecedented collaboration between the government, regulatory agencies, industry, and academia (Mathur et al., 2019; Panesar et al., 2021). Expected pilot shortages, cost challenges, and regulatory and other barriers must be systematically identified and mitigated to guide the design of future technology, methods, concept of operations, design guidelines, standards, and regulations. In this dissertation, we focus on the attentional challenges that are expected to result from the need for few ground-based operators to oversee a large number of pilotless eVTOL aircraft, particularly in the context of interruptions.

The SOMA concept is in stark contrast to today's operations in which managing even one unmanned aircraft like the RQ-21, a small tactical unmanned aerial vehicle (UAV) designed to support reconnaissance missions, requires a team of multiple trained personnel. In the context of SOMA, the operator will frequently need to interrupt their regular duties, such as health monitoring and flight configuration, to handle emergencies and requests generated by the aircraft. Since events such as loss of communication and bad weather are seldom predictable, they may interfere with the operator's tasks at inopportune times like during periods of high workload. This can be unacceptable in time- and safety-critical domains such as aviation where delays in the range of seconds can result in accidents. Successful teaming of operators with multiple autonomous agents therefore requires a better understanding of, and support for attention allocation and interruption management strategies.

Effective interruption management has been a long-standing challenge and an important research topic in safety-critical domains like aviation and healthcare. Interruptions can affect performance and safety, as illustrated by accidents such as Delta 1141 in 1988 (Wickens et al., 2013) where the First Officer inadvertently skipped a checklist item—lowering the flaps for takeoff—after responding to an Air Traffic Control (ATC) call to change the takeoff runway. As a result, the aircraft was unable to generate enough lift and crashed shortly after takeoff. In healthcare, interruptions are also a common occurrence and a major source of clinical errors (Trbovich et al., 2010; Westbrook et al., 2010). Factors such as interruption frequency, and the duration and complexity of interrupting tasks increase the time needed to return to the interrupted task and reduce the accuracy of recalling task-relevant information (Adamczyk & Bailey, 2004; Bailey & Iqbal, 2008; Borst et al., 2010; Cades et al., 2007; Monk et al., 2004, 2008; for a review of factors involved, see Puranik et al., 2020). It is therefore critical to avoid unnecessary interruptions, and to support operators in handling inevitable ones—the main focus of this dissertation research.

Researchers differ in their definitions of interruptions but most agree that interruptions present a bigger challenge than related concepts like distractions and task-switching (Adler & Benbunan-Fich, 2013; Baethge et al., 2015; Jett & George, 2003). Whereas distractions cause "a momentary deflection of attention from ongoing activities" (Latorella, 1997, p. 36), interruptions require both the deflection of attention *and* a suspension/resumption of behavior from ongoing task activity. Similarly, while frequent task switching is damaging to operator performance (Miyata & Norman, 1986; Monk, 2004), it does not necessarily require the suspension and resumption of the ongoing task (OT) —i.e., the next task is switched to only after completing, rather than momentarily suspending, the current task. Puranik et al. (2020) note five key aspects of past interruption definitions: (1) the suspension of the execution of an ongoing task (behavioral, attentional, or both), (2) the unexpectedness of the interruption, (3) the presence of an interrupting task (IT), (4) the intention to resume the interrupted task, and (5) the source of the interruption (external or internal). Puranik and colleagues consider only the first two aspects necessary to define an interruption. Moving forward, we adopt a definition which recognizes that interruptions a) require a shift of attention and suspension of behavior from the ongoing activity, b) involve the completion of an interrupting task, and c) involve (the intention to) return to the interrupted task. This definition aligns with Van den Berg et al. (1996, p. 236), Brixey et al. (2007, p. E38), and Hirsch et al. (2022, p. 147), and can be summarized as follows:

> *An interruption is a temporary suspension of the behavioral performance of, and attentional focus on, an ongoing task to execute activities that belong to a secondary task. The interrupted task is resumed after a certain lapse of time.*

In the next three sections, we review 1) the steps involved, and challenges associated with switching between ongoing and interrupting task activities, 2) possible ways of mitigating and recovering from the negative effects of task interruptions, and 3) the goals of this dissertation based on gaps in the current literature on interruption management.

1.1 The Interruption Management Process

As described in the Interruption Management Stage Model (IMSM) developed by Latorella (1997), the process of interruption management comprises three main stages. The first stage, detection, consists of the noticing and acknowledgement of an incoming interruption cue; the second stage, interpretation, corresponds to the identification of the cue to determine the priority of the interruption relative to the ongoing task; and the third stage, integration, involves making a decision on whether and when to accept, postpone, or reject the incoming interruption. Early work on interruption management has tended to focus on how operators handle only a few interruptions at each of these stages. However, as workplaces continue to become more complex and involve larger numbers of highly autonomous agents, interruptions have become more frequent and sometimes nested (e.g., Andreasson et al., 2017). Nested interruptions are cases where an ongoing task is interrupted and the interrupting task itself is then interrupted by yet another, potentially more urgent task. Frequent and nested interruptions present new and growing challenges that can lead to performance breakdowns at each of the three stages of interruption management (IM). These potential breakdowns—depicted in red in **Figure 1.2** at points A, B, and C—include a) higher likelihood of missing interruption signals, b) more effortful and potentially inaccurate assessments of the relative levels of priority of ongoing, interrupting, and pending tasks, and c) poor or failed integration of interrupting and interrupted tasks.

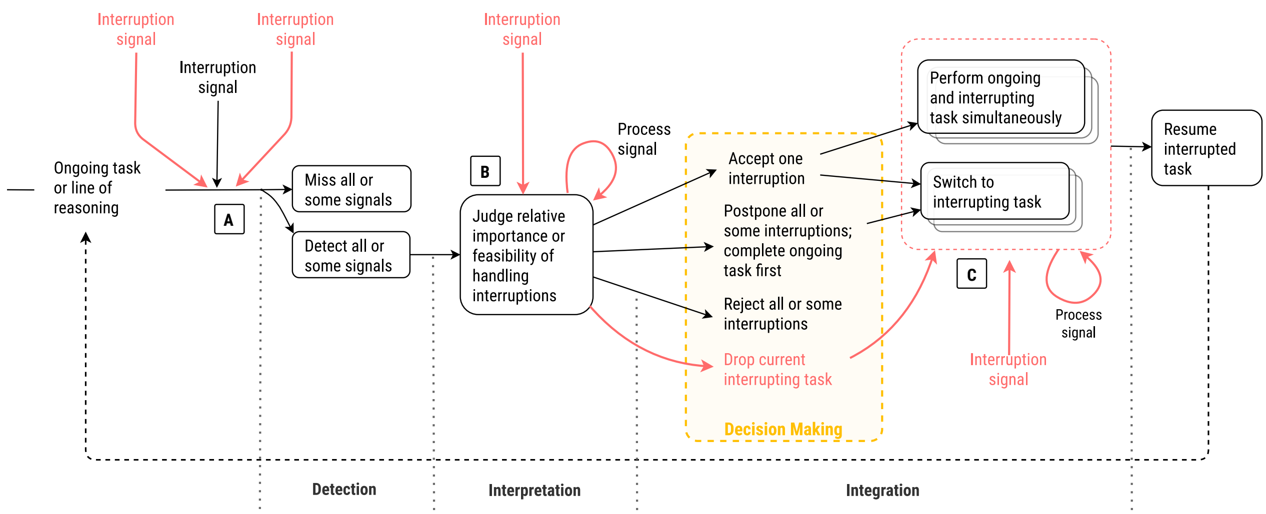

Figure 1.2 Process of interruption management (modified from Sarter, 2013). Items shown in red represent attention demands imposed by frequent and potentially nested interruptions. The yellow dashed rectangle indicates that decision making also is made more challenging when dealing with interruptions that are nested.

Two more recent well-established and widely-used models of working memory and multitasking provide insights into the cognitive processes involved in interruption management: the Memory for Goals (MFG) model pioneered by Altmann and Trafton (2002), and the unified theory of threaded cognition (TC) developed by Salvucci and Taatgen (2011b). According to the MFG model, behavior is driven by the most recently activated goal (i.e., the intention to accomplish a task or take an action) in working memory. These goals decline in activation over time and must be strengthened above an interference threshold in order to activate reliably and direct a person's behavior. If goal activation does not exceed this interference threshold, a person may fail to resume an interrupted task. The theory of threaded cognition re-conceptualizes the MFG concept of goals as task threads and integrates another cognitive resource called the "problem state," which Salvucci and Taatgen (2011a) define as representations in working memory of the context or information needed to complete a task. Threaded cognition explicitly recognizes the potentially concurrent nature of multitasking by stating that each goal or intention (e.g., an item on a to-do checklist) is represented by a separate task thread. Multiple task threads can be active at the same time, although with the constraint that only one task thread can use a cognitive resource (visual, aural, vocal, manual, problem state, declarative, and procedural) at any given time. Consequently, if both interrupting and yet to be completed interrupted tasks require the problem state resource, the two tasks will interfere with each other and likely prevent the rehearsal of information associated with the interrupted task. In other words, when returning to the interrupted task, recalling the problem state will take considerably longer or fail altogether. In the next three sections we discuss the specific challenges associated with the detection, interpretation, and integration stages of IM.

1.1.1 Detection

At the detection stage, an interruption signal may be missed a) due to low salience of the signal (e.g., Boot et al., 2006; Gibson & Peterson, 2001; Jones, 2001), or due to masking effects such as attentional blink where cues presented within a period of half a second or more after another cue are often missed (e.g., Broadbent & Broadbent, 1987; Di Lollo et al., 2005; Egeth et al., 2001; Raymond et al., 1992). Cue salience—"a signal-to-noise measure of the feature contrast (e.g., color, motion, luminance contrast) between the target and the surrounding stimuli relative to the feature variability among the surrounding" (Wickens & McCarley, 2019, p. 28)—has been incorporated into several models of noticing, particularly in the visual modality (e.g., Itti et al., 1998; Le Meur et al., 2006; Posner et al., 1980). Of these models, the most widely used is the **s**alience, **e**ffort, **e**xpectancy, and **v**alue (SEEV) model (Steelman et al., 2011; Wickens et al., 2009; Wickens & McCarley, 2019). The SEEV model posits that the reallocation of attentional resources, and in turn the noticeability of visual cues, depends on a combination of bottom-up, environment-driven factors of salience and effort, and top-down, knowledge-driven factors of expectancy and value. In other words, SEEV predicts that cues are more likely to be noticed if they are more salient, require less effort to be perceived, have a high expectancy/probability of occurring, and are regarded as important to the task at hand.

While the SEEV model has been validated primarily in the visual modality, there is evidence that similar relationships exist in the auditory and tactile modalities. To ensure appropriate relative levels of cue salience, research on multimodal interfaces (i.e., interfaces that distribute information across various sensory channels) suggests the presentation of alerts and alarms (such as interruption notifications) through lesser-used channels like hearing and touch (e.g., Chih-Yuan Ho et al., 2001; Ferris et al., 2006; Riggs et al., 2017; Sarter, 2006). Signals in those modalities are generally more salient than visual cues and do not require an effortful reorientation of the eyes, head, or body (Lu et al., 2011; Proctor & Proctor, 2021; Sarter, 2007; Sklar & Sarter, 1999; van Erp, 2007, Chapter 10; Wickens & Colcombe, 2007).

Salient cues do not guarantee detection. Masking phenomena like attentional blink may result in the signal being missed if it is presented simultaneously with or within a short interval of another cue (e.g., 200-500ms; Dell'Acqua et al., 2006; Martens & Wyble, 2010). This becomes more likely with an increase in the frequency of interruptions. Since attentional blink is known to occur in the visual, auditory, and tactile modalities (for an overview, see Wan, 2019), mitigation may require solutions that go beyond distribution of cues across sensory channels. For example, the artificial postponement of cues may be employed until there is sufficient temporal separation between the signals to ensure reliable detection. Assuming that the incoming interruption cue is sufficiently salient to surpass sensory thresholds, and is not masked by surrounding cues, the perceived stimulus is passed to short-term sensory storage for subsequent processing (Latorella, 1997).

In collaborative work environments, such as aviation, it is often insufficient for operators to only notice an incoming signal. Rather, they need to acknowledge receipt to the sender to maintain a shared awareness of system state (Klimoski & Mohammed, 1994; McComb et al., 2010; Nikolic et al., 2004). Flight crews of commercial airliners, for example, must listen for and immediately acknowledge clearances given by air traffic control (ATC) personnel. In case of emergencies, ATC may request pilots to "report the number of souls on board, when able." In those cases, ATC does not expect an immediate answer to the query because they understand that pilots are likely busy getting the aircraft under control and running appropriate emergency checklists. However, a verbal acknowledgement of the request is still expected.

The need for acknowledgements, by itself, represents a form of interruption as it requires an operator to briefly suspend their ongoing activities. However, models of interruption management, such as IMSM (see **Figure 1.2**), do not include this step as they focus primarily on individual operators. According to the MFG model, the successful activation of a goal—in this case, ‘acknowledge interruption request’—is not guaranteed and depends on contextual factors such as the number of recently completed and pending goals. With more frequent and nested interruptions, an operator likely experiences a higher level of interference in working memory due to the presence of residual goals for recently completed (sub)tasks, and from the need to interpret notifications for pending interrupting tasks. As a result, the operator may fail to reliably activate the goal to acknowledge the interruption signals. Additionally, the increase in procedural (scheduling of task steps) and declarative memory (recall of facts) interference due to the need to interpret a larger number of notifications and from the periodic need to rehearse information stored in the problem state resource can be expected to increase the time needed to acknowledge interruption signals.

1.1.2 Interpretation

Following the detection, and possibly acknowledgement, of an interruption signal, its accurate interpretation requires the operator to gather or recall information related to the interruption, retrieve from declarative memory the rules for mapping interruption characteristics to level of priority (e.g., based on importance, urgency, etc.), compare the level of priority to that of the ongoing task to classify the incoming task’s *relative* priority (i.e., higher, same, or lower), and retrieve from declarative memory the rules/actions associated with the determined level of relative priority. To continue with the example from the previous section, after acknowledging a controller’s request to report the number of passengers on board during an emergency, the interpretation stage requires that the flight crew determine whether the priority of the request is equal, higher, or lower than that of checklists and other emergency-related tasks they must perform and therefore whether it should be handled immediately or postponed until a safe state has been reached and workload has decreased.

When there are multiple pending interruption notifications, as is likely the case with frequent and/or nested interruptions, the interpretation process must be repeated for multiple task pairs, and, at the same time, the task with the current highest level of priority must be maintained in working memory for making subsequent comparisons. Recall that according to MFG, goals and representations stored in working memory, once activated, must be actively rehearsed to prevent them from gradually decaying and being forgotten, which may happen within a few seconds. Because newer goals are thought to interfere with old goals due to an increased interference threshold, both the reliable activation of goals for new tasks, and the maintenance of old goals for resuming old or interrupted tasks, becomes less reliable and more time consuming. During the interpretation stage, operators may therefore forget to interpret the interruption cue, map task characteristics to an incorrect priority level, forget the level of priority of the current task, or recall an incorrect rule based on the identified level of relative priority.

1.1.3 Integration

The third stage—integration—involves three steps: 1) based on the outcome of the interpretation stage, decide whether to accept immediately, postpone, or reject the incoming interruption, 2) switch to and complete the interrupting task(s), and 3) return to and complete the interrupted task(s). In the context of SOMA, operators who are responsible for multiple tasks and/or agents may find it difficult to choose a correct course of action, especially when facing resource tradeoffs or conflicts (e.g., decide to stay on a nearly complete task by delaying a higher priority pending task, or switch to the higher priority task immediately and forgo progress on the current task), and they may be prone to 'thematic vagabonding'—a sort of escape behavior where a person jumps from one topic to another without sufficiently dealing with the first (Dörner, 1980). In other words, frequent interruptions may trigger (unintentional) goal-switching behavior that prevents the operator from efficiently integrating incoming and pending tasks, remembering to complete deferred tasks, and completing the ongoing and interrupting tasks accurately.

Decision making

The first step after classifying the priority of an incoming interruption notification is to decide when and whether to accept the interrupting task. If the incoming task (IT) is of higher priority than the current, the operator may decide to switch to and perform the IT immediately. If instead the interrupting task is of equal priority, it may be better to postpone it and complete the ongoing task first. If the interrupting task has a level of priority below that of the ongoing task, it may be rejected altogether. Notably, the choice of delaying an interrupting task until a later time (e.g., until after reaching a breakpoint or next subtask) may be less appealing as it additionally requires that the operator remember to return to it. Accurate and timely recall of the need to execute a deferred task is a function associated with prospective memory, in which the "execution of retrieved intentions must be briefly delayed until an ongoing activity is completed" (McDaniel et al., 2004, p. 533). Prospective memory has been shown to be a problematic aspect of the interruption management process (Dismukes, 2012; Dodhia & Dismukes, 2009). For example, in a simulator study evaluating the effects of interruptions on deferred tasks, Wilson et al. (2018) report that participants failed to return to a deferred task more often when interrupted by a complex air traffic control task, compared to a blank interruption. This is in line with the

MFG model, which predicts that an increased level of interference and load on working memory from sufficiently demanding interrupting tasks reduces the chance of activating the correct goal (in this case, remembering to return to a deferred and interrupted task).

Switching to and performing the interrupting task.

To coordinate the switch between interrupting and interrupted tasks, McFarlane (1997) proposed a taxonomy of four ways: immediate, negotiated, mediated, and scheduled. An *immediate* interruption is presented without prior notice and must be attended to right away. *Negotiated* interruptions are presented with an advance warning that notifies the operator of a pending interruption and gives the operator some flexibility over when/whether to attend to the interrupting task. In *mediated* interruptions, the timing and/or occurrence of the interruption is controlled by a mediator, such as a software agent, whose goal is to infer different ways to present an interruption based on its properties such as urgency, importance, and associated risk (Adamczyk et al., 2005; e.g., Horvitz et al., 2005; McFarlane & Latorella, 2002). This could be done, for example, by delaying the interrupting task until a breakpoint is reached in the task or until the operator's workload drops below an acceptable threshold. Lastly, *scheduled* interruptions are presented at fixed, pre-determined points in time. Scheduled interruptions allow the operator to anticipate (and thus prepare for) the switch to the interrupting task.

A comparison of the four coordination methods by McFarlane (2002) showed that the negotiated strategy resulted in best overall performance on an emergency response task. However, the immediate strategy elicited faster handling of the interrupting task, which would make it more suitable for time-critical interruptions (such as a request to reroute an aircraft with a passenger experiencing cardiac arrest). In the case of negotiated and mediated interruptions, therefore, a major challenge is to support the operator in noticing and processing incoming interruptions to ensure that interruption lag (the period between the onset of an alert and the initiation of an interrupting task) is minimized while maintaining speed and accuracy on both the ongoing and interrupting tasks. The interruption lag period is critical in supporting this process as it gives operators a chance to deliberately store relevant task information in order to prepare to switch to the interrupting task and facilitate the return to the interrupted task after completing the interrupting task.

Following a switch away from the ongoing task, performance on the interrupting task is thought to be affected primarily by the need to rehearse the information and goals related to the interrupted task, which may interfere with the declarative, problem state, and procedural resources needed by the interrupting task. Recall that, according to threaded cognition, the use of these cognitive resources is sequential—i.e., access is blocked until the production-rule for the current procedural step is fired (~50ms; Salvucci et al., 2009), or until the cued item is retrieved from declarative memory (up to ~200-500ms). Therefore, the periodic rehearsal of the problem state can be expected to interfere with an interrupting task that also relies heavily on procedural and declarative memory resources. For example, Salvucci and Beltowska (2008) found that when asked to perform a memory recall task, participants in a driving study experienced larger deviations from lane center and had longer response times to braking. This indicates that the need to periodically remember to rehearse the problem state information, which relies on the procedural resource, can interfere with simultaneous performance on other procedural tasks and access to declarative memory. Notably, the effects of interference from problem state rehearsal last only a few seconds at the beginning of the interrupting task until repeated memory recall becomes difficult or impossible (Salvucci & Taatgen, 2011a, Chapter 4). Performance on interruptions longer than a few seconds may therefore not be affected at all, given that working memory representations associated with the interrupting task being performed are more recent and have a higher level of activation, especially if performance measures are aggregated over the entire task (Monk et al., 2008).

Returning to and performing the interrupted task

After completing the interrupting task, resumption of, and accuracy on the interrupted ongoing task (OT) is affected by a number of factors which can be grouped into three categories: task-related factors, situational factors, and personal factors (Hirsch et al., 2022). Task-related factors relate to the attributes of task(s) being performed. These include OT/IT task complexity, duration of the interrupting task, and similarity between OT and IT. Situational factors refer to the environment and context in which interruptions are presented and managed. These factors include interruption timing and position, interruption frequency, and nesting level. Lastly, personal factors relate to a person's capacity and efficiency in handling interruptions. These include attributes like working memory capacity, experience, and level of motivation.

Task complexity is an umbrella term that is composed of multiple dimensions (Liu & Li, 2012) and is operationalized differently across experiments (Couffe & Michael, 2017; Williams et al., 2020). In some cases, it is defined by the number of components to be processed, and the degree of uncertainty associated with the task (e.g., Brazzolotto & Michael, 2020; Cades et al., 2007). In other cases, complexity is associated with the severity of time constraints (e.g., Braarud & Kirwan, 2011; Greitzer, 2005). Typically, prior research suggests that more complex primary tasks take longer to resume when returning from an interruption (Magrabi et al., 2010; Monk et al., 2008; Szumowska & Kossowska, 2017). This has been explained by the increased difficulty in encoding and remembering the relevant components needed for resuming the interrupted task. Similarly, when returning from interrupting tasks that are more complex, performance has been found to suffer in terms of resumption lag (the period between the end of the interrupting task and the first action after the return to the interrupted task), total time on task, and accuracy on the interrupted task (Basoglu et al., 2009; Cades et al., 2007; Eyrolle & Cellier, 2000; Gillie & Broadbent, 1989; Hodgetts & Jones, 2006b; Mansi & Levy, 2013; Radović & Manzey, 2022; for a contradictory example, see Zijlstra et al., 1999). Like complexity, duration of the interrupting task is, in general, correlated positively with resumption lag on the interrupted task (Altmann et al., 2017; Borowsky et al., 2016; Fong et al., 2017; Foroughi, Werner, et al., 2016; Hodgetts & Jones, 2006b; Monk & Kidd, 2008). That is, longer interruptions tend to prolong the decay of interrupted-task-related elements stored in working memory and, in turn, lead to longer resumption lag on the interrupted task. This effect is particularly strong when memory rehearsal is inhibited as a result of, for example, interruptions that are both long and highly complex (e.g., Borst et al., 2015). Personal factors may interact with task-related factors to alleviate or exaggerate their effects on resumption lag and accuracy. Increases in working memory capacity, for example, have been shown to reduce resumption lag on the interrupted task, and attenuate the effects of interruption duration (e.g., Foroughi, Malihi, et al., 2016; Foroughi, Werner, et al., 2016).

Similarity (e.g., processing modality) between the interrupting and interrupted tasks has also been shown to increase resumption time and reduce task accuracy (Czerwinski et al., 1991; Lee & Duffy, 2015; Lu et al., 2013). Lee and Duffy (2015), for example, evaluated the effect of cognitive (solving seventh-grade level math word problems) and skill-based (typing a sentence) task pairs on interrupted task performance. They found that participants took longer to complete the interrupted task when both the interrupting and interrupted tasks were cognitive or skill-based. Factors that affect concurrent task performance and time-sharing of attentional resources have been documented extensively under the widely established Multiple Resources Theory (MRT) developed by Wickens (1980; for an overview, see Wickens et al., 2022). MRT describes how multiple tasks or pieces of information can be performed/processed in parallel, without creating interference or decrements in performance, if they draw from different pools of attentional resources (similar to the cognitive resources described in the threaded cognition model in the previous section). These resources are shown in **Figure 1.3** in a four-dimensional structure, with each dimension—modality, processing stage, processing code and response type—being composed of multiple discrete levels. The MRT model predicts that any two tasks will compete for attentional resources to the extent that they occupy the same level on one or more dimensions (e.g., two visual perceptual tasks or two auditory spatial tasks). Note that the modalities dimension could be extended to include three levels, instead of two, with the addition of the tactile modality.

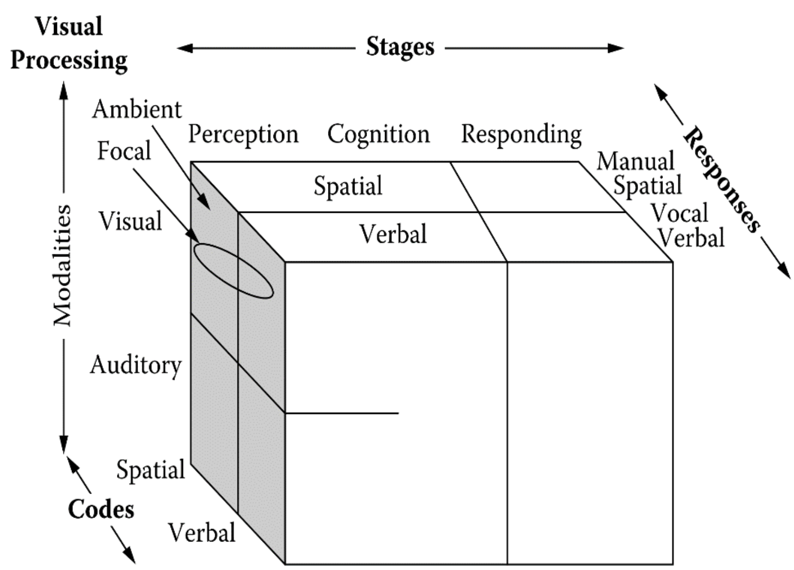

Figure 1.3 A representation of the Multiple Resource Theory (Wickens & McCarley, 2019, p. 132).

Even though MRT was developed specifically to predict task performance while time sharing, there is some evidence that it may be extended to sequential multitasking paradigms like interruption management, where attention is switched between tasks in an 'all-or-none' manner (for a detailed overview, see Wickens et al., 2013, Chapter 10). For example, interrupting and interrupted tasks that share the same processing code lead to slower and less accurate resumption on the interrupted tasks (Brudzinski et al., 2007; Edwards & Gronlund, 1998; Ratwani et al., 2007; Ratwani & Trafton, 2008). In a sequential cell-based number entry task that relied on spatial memory resources to remember which cell to return to, Ratwani and Trafton (2008) report that spatial interruptions (mental rotation) resulted in participants taking longer and being less accurate in returning to the pre-interruption location, compared to a verbal interruption (an arithmetic task). These results suggest that task status and progress (i.e., the problem state) are stored in working memory using, at least partially, distinct attentional resources for spatial and verbal representations. Therefore, interrupting and interrupted tasks that share the same processing code (e.g., verbal-verbal) can be expected to increase resumption lag and reduce task accuracy, compared to task pairs with different processing codes (e.g., verbal-spatial).

Beyond attributes that are intrinsic to the ongoing and interrupting tasks, manipulating situational factors like increasing the number of times an ongoing task is interrupted by a secondary interrupting task, i.e., increased interruption frequency, has been linked to a decline in performance on the interrupted task in terms of both resumption lag and accuracy (Basoglu et al., 2009; Lee & Duffy, 2015; Westbrook et al., 2010; Zijlstra et al., 1999). Santomauro et al. (2021), for instance, found that nurses were twice as likely to administer incorrect dosage when they were interrupted 12 times per hour while performing a medication entry task, compared to a baseline of three. However, other research in healthcare settings has found no effects of interruption frequency on resumption lag or accuracy on the interrupted task (e.g., Drews et al., 2019; Thomas et al., 2017), and researchers in other application domains even demonstrated a *positive* effect of interruption frequency on primary task performance (e.g., Monk, 2004). In a VCR programming task, for example, Monk (2004) found that more frequent interruptions resulted in shorter resumption lags, with no observed costs in resumption errors or overall time on task. Part of the reason for these contradictory findings is likely the varying degrees of task complexity employed across the studies. Interruptions tend to benefit performance (i.e., reduced time on task) on simple primary tasks (e.g., Mark et al., 2008; Monk et al., 2004; Speier et al., 2003) and degrade performance (i.e., higher resumption lag and lower accuracy) on more complex primary tasks (Bailey et al., 2000; e.g., Lee & Duffy, 2015; Magrabi et al., 2010; Westbrook et al., 2010). Baethge et al. (2015) note that the relationship between interruption frequency and task performance follows an inverted U curve (Yerkes & Dodson, 1908), where the cumulative increase in workload, while initially beneficial due to increased motivation and exerted effort, eventually leads to performance breakdowns. Similarly, the timing and position of interruptions during the primary task can also moderate the effects of interruptions on resumption performance and accuracy (e.g., Adamczyk & Bailey, 2004; Bailey & Iqbal, 2008; Botvinick & Bylsma, 2005; Brazzolotto et al., 2022; Cutrell et al., 2001). In other words, even for tasks that are complex, interruptions presented during periods of low workload, such as in between subtasks, are less disruptive to task performance.

Findings are sparse on the effects of nested interruptions on task resumption. In fact, only a few studies have even acknowledged the potentially nested nature of task interruptions (Andreasson et al., 2017; Baethge et al., 2015; Sasangohar, 2015). Andreasson and colleagues (2017), for instance, conducted ethnographic observations and semi-structured interviews of workers at a diesel manufacturing plant, focusing particularly on how workplace activities were performed in the presence of interruptions. The authors observed that the collaborative nature of manufacturing sometimes led to the interruption of the primary task by several interrupting tasks, some of which required the workers' attention before the previous interrupting task could be completed. These nested interruptions often came in the form of scheduled maintenance work being interrupted by more 'acute' requests for repairs which were, at times, interrupted by an even more urgent alert. Similarly, Baethge et al. (2015), Sasangohar (2015), and, more recently, Laarni (2021) have also highlighted the existence of nested interruptions in the office and healthcare settings.

To our knowledge, only two articles have empirically evaluated the effects of nested interruptions on task resumption and accuracy (Saleh, 2019; Sasangohar et al., 2017). In a set of studies conducted by Saleh (2019), participants performed three rule-based tasks that simulated control room operations at varying levels of interruption depth (i.e., nesting level). Participants either completed all tasks sequentially, one of three tasks as a secondary interrupting task, or two of the three tasks as nested interrupting tasks. While the overall completion time for the set of three tasks was higher for nested interruptions, as would be expected due to the higher number of switches between tasks, nested interruptions were not found to result in longer resumption lag. Accuracy on the interrupted task as a function of nesting level was not reported. On the other hand, Sasangohar et al. (2017) conducted a study where intensive care unit nurses performed a medication entry task during which they experienced nested interruptions. Findings from this research show that nested interruptions resulted in longer resumption lags and lower accuracy on the interrupted tasks. The discrepancy in findings between the two lines of research is most likely a result of differences in the nature of the ongoing and interrupting tasks. The higher load on working memory in the medical task in the second study may explain the larger degradation in performance when returning to the interrupted task, compared to the procedural control room task in the first study.

The longer resumption times observed in the case of nested interruptions can be attributed, in part, to a slow or failed recall of problem state and goals from working memory when returning to the interrupted task. Compared to single and serial interruptions, where only the goal and information regarding the ongoing task needs to be maintained, nested interruptions impose a higher load on working memory due to the need to remember information related to both the ongoing and the secondary interrupting tasks. As such, MFG predicts that an increased level of interference in working memory will result in the delayed and potentially failed activation of all or some goals related to the interrupted tasks.

1.2 Supporting Interruption Management and Recovery

In a review of possible strategies for mitigating the negative effects of interruptions, Sasangohar et al. (2013) note two possible approaches: 1) more effective management of interruptions, and 2) support for recovery from interruptions. Interruption management solutions focus on reducing the negative performance effects of interruptions by changing how the interruptions are presented and coordinated with the operator. Interruption recovery solutions, on the other hand, help operators return to the interrupted task in a timely and efficient manner by providing, for example, information about past events.

1.2.1 Interruption Management

MRT suggests that reliable detection can be achieved by employing underutilized sensory channels for presenting the notification. In addition, information presented in non-traditional and less taxed channels can support *preattentive reference*, i.e., the ability for people to process, in parallel, changes in the environment and reorient perceptual systems to potentially interesting conditions (Folk & Remington, 2006; Treisman & Souther, 1985; Woods, 1995). In the context of interruption management, this means that operators are enabled to process partial information about the nature of the interrupting task in one sensory channel without having to interrupt the performance of an ongoing task in a different channel.

Limited guidance is available for the most effective selection and combination of modalities for supporting parallel interpretation of interruptions. Hameed et al. (2006) employed multimodal information presentation to inform participants about the type, importance, and expected length of an interruption of a simulated visual feedwater control task. They found that, compared to the baseline condition of a simple focal visual binary indicator, presenting information about interruptions in the form of peripheral visual and tactile cues led to (a) higher detection rates for interruptions and (b) more appropriate task switching and management of interruptions due to parallel preattentive processing of the interruption information while performing the interrupted task. Furthermore, a review of empirical studies on multimodal interfaces by Sarter (2013) showed that response times to interrupting tasks with high urgency are lower if they are presented in the auditory modality, but low urgency tasks illicit a faster response if presented through touch. Ho et al. (2004) ran a simulated Air Traffic Control Task and tested whether presenting information about the modality, urgency, and source of the interrupting task had any effect on operator behavior or performance. They found that interrupting tasks presented in the auditory modality allowed the operator to continue doing the ATC task in parallel without significant performance costs but performing the interrupting task in the tactile and visual modalities resulted in degraded performance of the interrupted task.

At the integration stage, studies have shown that indicating the imminent onset of an interruption can reduce the resumption lag by allowing for a more graceful and deliberate disengagement from the interrupted task as compared to an immediate/unexpected interruption (Altmann & Trafton, 2004; Hodgetts & Jones, 2006a; Trafton et al., 2003). While the desirable interval of time between the interruption warning and interruption onset is heavily dependent on the complexity of the ongoing and interrupting tasks, several studies have found lengths of at least 5-6 seconds to be beneficial in reducing the resumption lag (e.g., Gold et al., 2013; Mok et al., 2015). More recently, in a dynamic command and control task, Labonté et al. (2019) successfully employed pre-interruption warnings eight seconds before the onset of the interrupting task to reduce resumption lag compared to interruptions presented without a warning. They show, using both eye-tracking and task performance metrics, that pre-interruption warnings result in better encoding of the task state before the onset of the interruption, which in turn reduces the time needed to resume the interrupted task. They note that this technique results in an increase in mental workload before switching to the interrupting task, but the mental workload is reduced after returning to the interrupted task. However, their study only presents one interruption during a 3-minute scenario and does not require the operator to judge the nature of the interrupting task. It is not clear whether the benefits of the pre-interruption warning will extend to time-constrained environments with more numerous interruptions because pre-interruption warnings are only useful if they are reliably detected to begin with.

1.2.2 Interruption Recovery

In many cases, it is not possible to warn the user in advance or provide control over the timing of the interruption due to their unpredictability. Instead of announcing the interruption in advance, interruption recovery methods thus help operators quickly return to the interrupted task by viewing and/or replaying information about recently occurred and critical events (Sasangohar et al., 2014; Scott et al., 2008, 2006) John et al. (2005), for example, tested an interruption recovery aid called CHEX (Change History Explicit) that helps operators recover from interruptions in a dynamic simulated air traffic control task by tracking, logging, and allowing replay of important changes as they occur in a “Change History” table. Their results indicate that easy access to critical past events reduced the time needed to return to the interrupted task, but the ability to replay past events also resulted in resumption lag times *longer* than no support at all. For a time-critical UAV monitoring task, Scott and others (2006) tested event-based static and animated timeline displays to make it easier to retrieve the working context when switching back to the interrupted task. The static timeline displayed bookmarks depicting the occurrence of important events such as UAV status change and appearance/disappearance of threats. The animated display additionally allowed participants to playback the events for a specified time window at an accelerated speed. The authors discovered that both the static and animated timeline displays helped improve interrupted task resumption times, compared to the no assistance condition, for complex tasks only. They noted *increased* interrupted task resumption times for tasks that were less demanding, which indicates that the type of assistance provided needs to factor in the (perceived) complexity of the interrupted task. The authors note that one way to mitigate the observed negative effects of timeline displays might be to reduce visual clutter by limiting the timeline to only contain mission-related items such as the onset of threats, rather than all changes in UAV state.

Because the complexity of encoding system state and the frequency of interruptions can be expected to increase with respect to the number of vehicles in multi-agent supervision, there is an opportunity to extend this work to a) evaluate the effects of managing a non-homogeneous set of vehicles with varying mission requirements, and b) increasing the number of vehicles tested.

1.3 Research Gaps

To date, most work on interruption management has focused on paradigms involving one human operator interfacing with one or few agents at rather low levels of automation and experiencing only a small number of (single) interruptions (e.g., Bogunovich & Salvucci, 2011; Monk, 2004; Monk et al., 2008). Only recently, the increasing problem of frequent and nested interruptions has been highlighted (Andreasson et al., 2017; Baethge et al., 2015; Laarni, 2021). To date, very few researchers (Saleh, 2019; Sasangohar et al., 2017) have empirically examined the concept of nested interruptions and only at the integration stage—none have examined the effects of nested interruptions on performance at the detection and interpretation stages.

Another limitation of much of the existing literature in interruption management is that it has employed a) immediate and forced interruptions that must be completed at the time of onset or b) deferable/negotiable interruptions in simple, non-safety- or time-critical contexts (e.g., Iqbal & Horvitz, 2007; Salvucci & Bogunovich, 2010; Wiberg & Whittaker, 2005). Immediate and forced interruptions reduce or eliminate the need for timely and accurate interpretation of associated cues and proper integration of the interruption into the workflow (e.g., Iqbal & Bailey, 2005; Monk, 2004; Monk et al., 2004; Trafton et al., 2003). Even though design guidelines (e.g., Salvucci & Taatgen, 2011a) on interruption management suggest that forced interruptions should be avoided in favor of user self-interruptions, most empirical work to date still either employs forced interruptions, or negotiated interruptions with simple tasks, and without the need for the operator to consider the relative priority of incoming tasks.

The overall goal of the present research is to complement and expand on earlier work by assessing the effects of frequent and nested interruptions on operator performance and the interruption management process. In the chapters that follow, we report and discuss results from three empirical studies. The specific goal of the first experiment, detailed in **Chapter 2**, is to identify the challenges that operators encounter, and the strategies they employ in complex SOMA operations when managing a large number of negotiated interruptions, some of which are presented as interruptions that are nested within other interrupting tasks. Building on findings from the first study, the second experiment (**Chapter 3**) evaluates more closely the breakdowns that occur in the interpretation and scheduling of interruptions at the interpretation and integration stages. In **Chapter 4**, we summarize key findings and challenges identified in the first two studies and provide a brief discussion of mitigation methods proposed for addressing the identified challenges. In **Chapter 5**, we describe in more detail the selected mitigation methods and empirically evaluate them in a third and final experiment. Finally, in **Chapter 6**, we conclude with remarks on the practical and broader implications of the present research and provide a brief outlook on the future of this work.

Chapter 2 – Identifying Challenges with Interruption Management

2.1 Introduction

As detailed in **Chapter 1**, interruption management can break down at each of the three stages of detection, interpretation, and integration. At the detection stage, the onset of multiple interruptions in close temporal proximity reduces the likelihood of noticing and acknowledging interruption signals. Accurate judgement of task priority at the interpretation stage is made more difficult due to the need for repeated comparisons between pending and incoming tasks. And at the integration stage, deciding when and which task to switch to, and returning to the interrupted task in a timely fashion, is more likely to fail when there are multiple pending tasks. To date, most work on interruption management has focused on paradigms involving only a small number of single interruptions in the context of forced interruptions which reduce the need for operators to prioritize and schedule pending tasks.

The goal of this first study was to complement and expand on earlier work by identifying the difficulties faced by operators when coping with frequent and nested interruptions in a supervisory command and control task. To this end, interruption frequency was operationalized as the number of interruptions presented during an ongoing task. It varied between low and high. Interruptions were presented as single, serial, or nested interruptions. Single interruptions were presented individually, at separate points over the course of an ongoing task. In contrast, serial and nested interruptions were presented in pairs. In the case of serial interruptions, the second interruption was introduced immediately after completion of the first interrupting task. In the case of nested interruptions, the second interruption occurred a few seconds after the *start* of the first interruption, thus ‘interrupting the interruption’. Lastly, task priority was manipulated by changing the level of urgency and was defined as the time available to accept and switch to a pending task.

In line with the theories and literature discussed in the previous chapter, we expected the following outcomes:

H1: *Participants will be less likely to detect and slower to acknowledge interruption notifications during high frequency periods and for nested interruptions compared to low frequency periods, and serial/single interruptions, respectively.*

This expectation is based on the assumption that high frequency and nested interruptions will result in a higher chance of failing to perceive and/or acknowledge interruption signals due to increased interference from active and residual goals. Acknowledgement time is expected to be longer during high frequency and nested conditions as a result of higher procedural and declarative interference from the need to interpret more notifications, and the periodic rehearsal of the goal and problem state needed to return to the interrupted task, respectively.

H2: *Participants will be less accurate at interpreting interruption notifications during high frequency periods and when interruptions are nested, compared to low frequency and single/serial interruptions, respectively.*

MFG predicts that the activation and retention of more goals in working memory increases the effort needed to activate new goals and remember current ones. As a result, we expect that operators will forget to interpret interruption cues, will map task characteristics to incorrect priority levels, and forget/misclassify the level of urgency of the current task. Time needed to process and interpret notifications is expected to be longer because of the need to potentially make comparisons between multiple notification/task pairs.

H3: *Participants will perform worse on the interrupted task during high frequency interruption periods compared to low interruption frequency. This effect will be more pronounced when interruptions are nested. Performance on the interrupting task will be lower during high frequency interruption periods as compared to low frequency periods.*

In the case of frequent and nested interruptions, it is assumed that there will be higher working memory interference due to the need to maintain goals for resuming both the first and the second interrupted tasks. The higher level of interference is expected to increase the chance of failed or delayed activation of goals and subtasks on the interrupted task. Performance on the interrupting task is expected to degrade with more frequent interruptions as a result of the induced urgency to return to the interrupted task.

H4: *Participants will take longer to resume working on the flight request task when returning from nested interruptions and from interruptions with the same processing code as the interrupted task.*

We expect that both nested interruptions and interruptions with the same processing code will limit the rehearsal and retention of task goals and problem state due to the need to encode a larger number of resumption goals, and the use of overlapping processing code resources.

H5: *Participants will cope with more frequent and nested interruptions by simplifying their strategy to switch to interrupting tasks.*

We expect that participants will engage in focused attention to minimize disruptions to performance on the new flight task. This may be done by simplifying the decision-making criteria to always ignore or delay as long as possible interruptions of low and medium urgency, but switch immediately to high urgency interruptions.

2.2 Method

2.2.1 Participants

40 University of Michigan students (22 male, 17 female, one non-binary) between the ages of 18 and 28 years (*M* = 21, *SD* = 2.5) participated in the study. Participant eligibility was limited to ages 18-30 years, which is comparable to the age range of current (and anticipated) UAV operators. Participants were paid $16/hour for completing one three-and-a-half-hour experiment session. A $20 performance bonus was awarded to participants with performance scores in the top quartile. This research was conducted in compliance with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at the University of Michigan (IRB #HUM00195849). Informed consent was obtained from each participant before the experiment.

2.2.2 Experiment Apparatus and Tasks

Participants were tasked with supervising a set of autonomous drones delivering cargo to commercial locations in the Houston, Texas metropolitan area. Participants performed five different tasks: approving new flights, selecting alternate landing sites, diverting to alternate vertiports (i.e., designated landing ports), detecting unauthorized aircraft, and responding to requests for vehicle status information. The experiment was conducted using a multi-UAV simulator developed by the THInC lab based on the Air Force Vigilant Spirit ground control station (Feitshans et al., 2008). It consists of two 24” monitors running at a resolution of 1920 x 1200 each (see **Figure 2.1**). The left-hand monitor displayed the ongoing task of approving new flight requests and a notification area with pending interrupting tasks. The right-hand monitor displayed one of the four interrupting tasks on the left-hand side and UAV health and mission progress information for each of the 20 simulated aircraft on the right side. Lastly, a foot pedal was used for acknowledgement of interruption notifications. The button-activated pedal was located on the floor underneath the experiment workstation. Participants were allowed to reposition the pedal to ensure reliable activation.

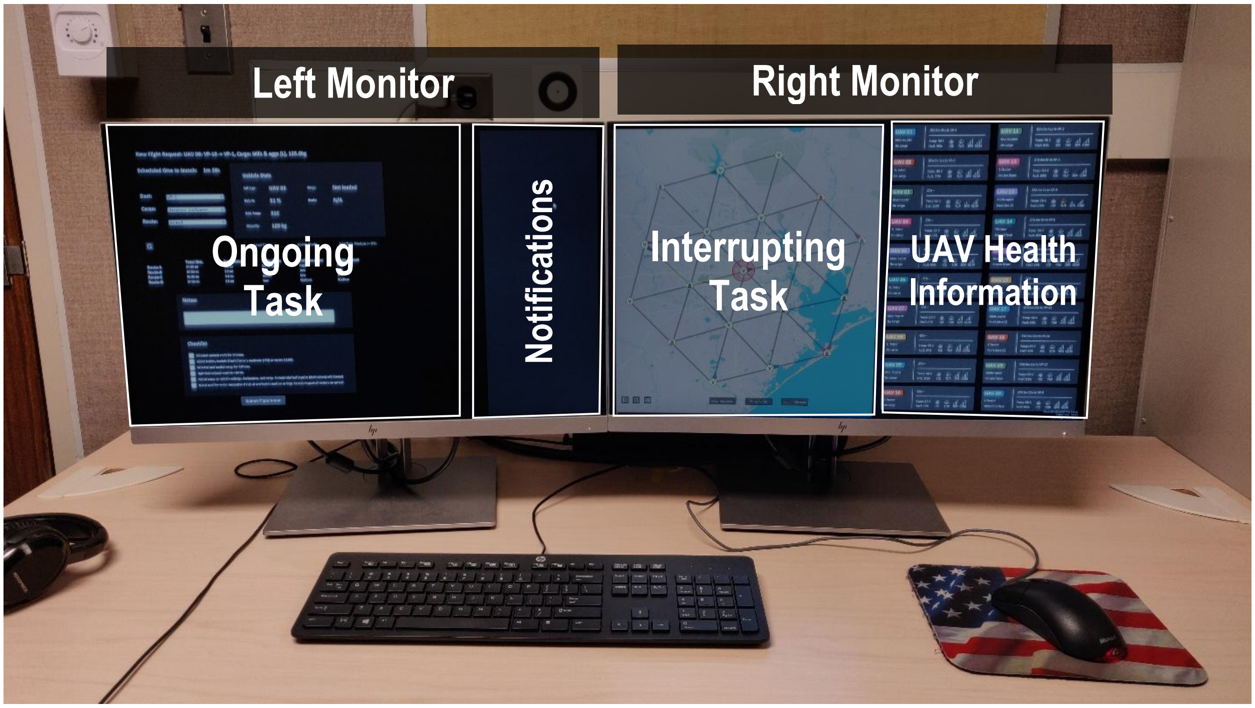

Figure 2.1 Experiment setup showing the task interface on two monitors. The left-hand monitor displays the ongoing task and the notifications area. The right-hand monitor displays one of four interrupting tasks and UAV health information.

Ongoing Task

Throughout each scenario, participants were responsible for one ongoing task (OT) that involved configuring and approving new flight requests (shown in **Figure 2.2**). The flights simulated the transportation of packages, food, and medical cargo between various landing and takeoff sites called vertiports. This task required participants to select the destination vertiport, the cargo to be transported, and an optimal route. The optimal route was selected out of four possible options based on five criteria of minimum travel distance, maximum separation distance, minimum turbulence level, minimum risk of overheat, and minimum noise impact. For all flight request tasks, only one route option was considered optimal and better than the other three route options. The order of importance of the criteria depended on the type of UAV (multicopter or tiltrotor) and the type of cargo (liquid or fragile). Shorter travel distance took higher precedence than separation distance for multicopters while higher separation was more important for tiltrotor UAVs. In cases where the cargo to be transported was liquid or fragile, lower turbulence took precedence over both travel distance and separation distance. Risk of overheat and noise impact were always the last two criteria to be optimized.

Table 2.1 Prioritization order of criteria for selecting the optimal route in the flight request task.

Cargo Type	**UAV Type**	**Prioritization Order**
Not Liquid/Fragile	Multicopter	Total distance → Separation → Turbulence → Risk of overheat → Noise impact
	Tiltrotor	Separation → Total distance → Turbulence → Risk of overheat → Noise impact
Liquid/Fragile	Multicopter	Turbulence → Total distance → Separation → Risk of overheat → Noise impact
	Tiltrotor	Turbulence → Separation → Total distance → Risk of overheat → Noise impact

Following selection of the optimal route, participants were asked to load the appropriate cargo, uplink the selected route, and add extra battery modules if needed (none for low turbulence along route, one for medium turbulence, and two for high turbulence). Those three tasks each took approximately five seconds each to complete before the vehicle status was updated. Participants then typed the UAV callsign, destination vertiport, and name of cargo into a textbox to relay this information to the destination vertiport. In cases where the risk of overheat along the selected route was medium or high, participants were asked to additionally relay a warning message to “inspect all motors on arrival.” Lastly, participants were asked to verify the completion of all subtasks and complete a checklist before submitting the flight request by clicking the “Submit flight intent” button (see **Figure 2.2**). Once submitted, the message “UAV ## flight request submitted!” was displayed for five seconds before it was replaced by the next flight request task.

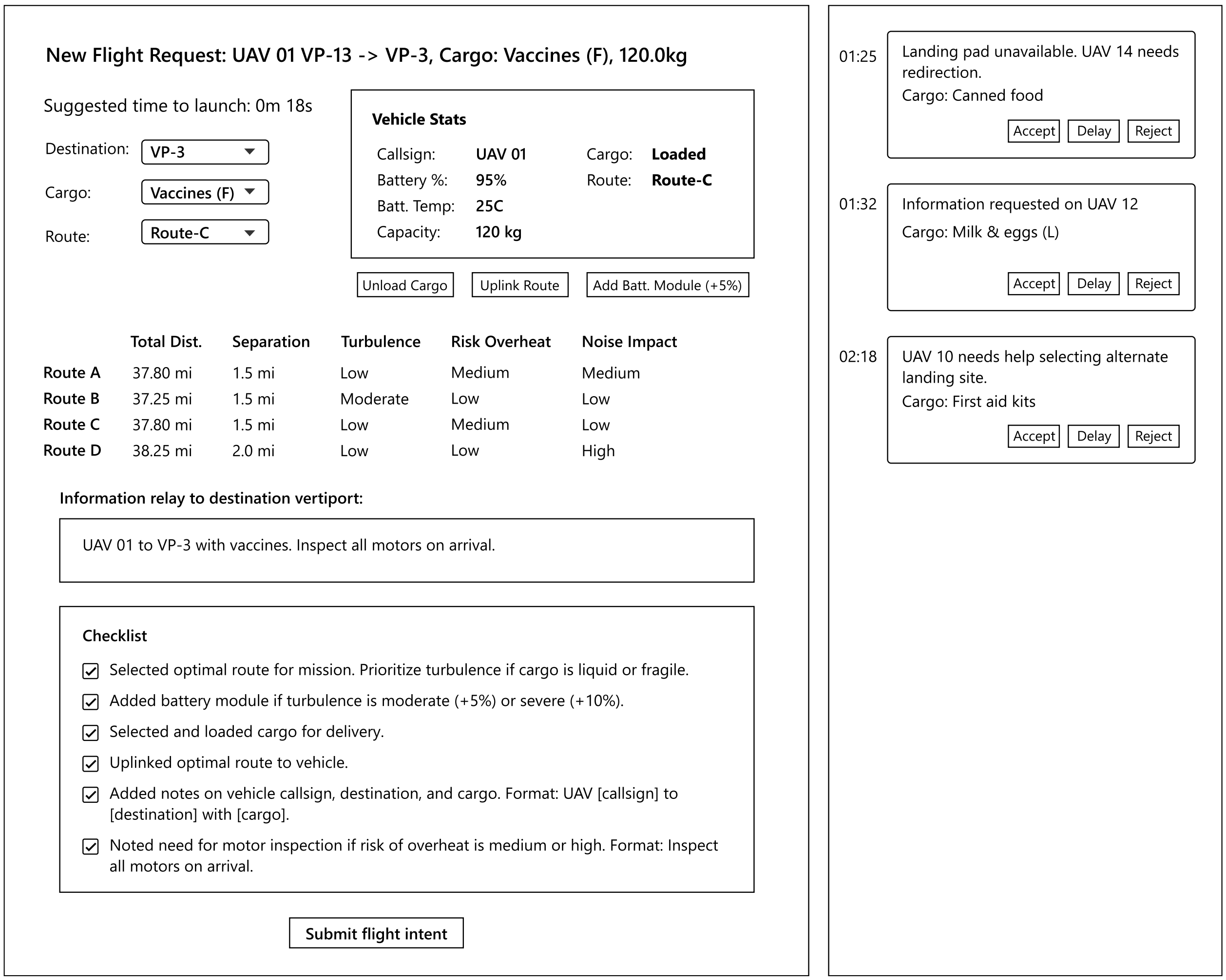

Figure 2.2 Task interface displayed on the left -hand monitor showing the flight request task (left) and notifications for pending interrupting tasks (right). Background colors were removed to improve legibility.

Interrupting Tasks

Four interrupting tasks were used in this experiment—vertiport diversion, alternate landing site task, vehicle authorization, and request for information. To examine whether participants would take longer to resume working on a new flight task when returning from an interrupting task with the same processing code (H4), these tasks required processing of either verbal or spatial information. The alternate landing site task and the vertiport diversion task involved ‘analog-spatial processing’ (Wickens & McCarley, 2019, p. 136). Participants had to compare shapes/patterns and remember the relative location of elements on a map display. The vehicle authorization and information request tasks were verbal in nature. They required ‘categorical-symbolic processing’ (Wickens & McCarley, 2019, p. 136) as they involved making comparisons of letters/numbers, and remembering UAV attributes like fuel and callsign.

Table 2.2 Mapping of task urgency to cargo type and cargo names. ‘L’ and ‘F’ indicate whether the cargo was considered liquid or fragile, respectively.

Task urgency	Cargo type	Cargo name
High	Medical	Insulin pumps, first aid kits, vaccines (F), syringes & vials (F)
Medium	Perishable	Frozen items, takeout orders, milk & eggs (L), fruit juice (L)
Low	Non-perishable	Amazon packages, canned food, clothes, pet supplies

The interrupting tasks were each assigned an urgency level of low, medium, or high depending on the type of cargo the affected UAV was carrying (see **Table 2.2**). On the right side of **Figure 2.2**, for example, the notifications panel shows three pending tasks of low, medium, and high urgency from top to bottom, respectively. Notifications were displayed in the order in which they arrived—that is, newer notifications were displayed below older ones. The interrupting tasks were designed to be similar in difficulty and the amount of time needed for completion (approximately 25-40 seconds) and could be completed independently of the ongoing flight request task.

Vertiport diversion (spatial). In the vertiport diversion task (**Figure 2.3**), participants were asked to redirect vehicles to an alternate vertiport in case the original landing location became unavailable due to, for example, weather or limited vertiport capacity. Participants were given four different route options and asked to select an optimal route based on the level of noise impact and level of turbulence experienced along the route, prioritized based on whether the cargo carried by the affected UAV was liquid or fragile.

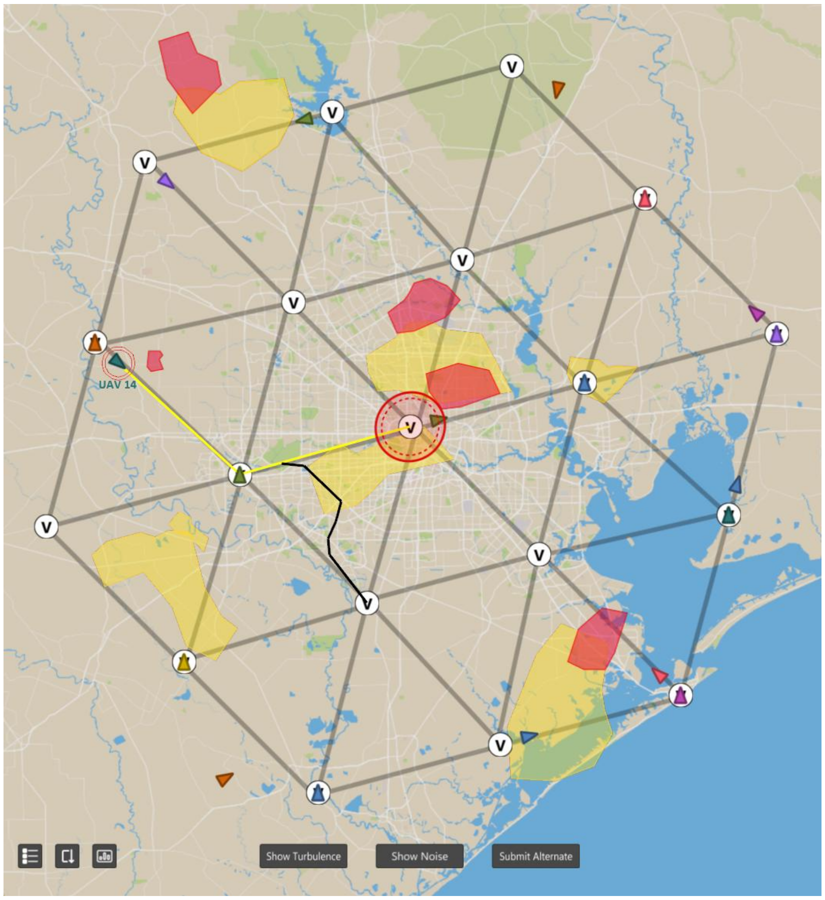

Figure 2.3 Vertiport diversion task interface. Participants cycled through and reviewed four route options, shown one at a time with a black line, to safely divert to an alternate vertiport (i.e., takeoff and landing location shown with a "V" symbol). Route prioritization criteria included minimizing the level of turbulence (high or medium, shown as red and yellow polygons) and level of noise impact (high or medium, not pictured). Vertiport placement and route layout were modeled in the Houston, Texas metropolitan area based on a hexagonal structure proposed by Patterson et al. (2018).

Alternate landing site (spatial). The alternate landing site task (**Figure 2.4**) involved selecting an optimal emergency landing location by comparing the population density, landing site risk, path risk, and landing site type of four different options shown on a map display. In addition, participants were asked to select an emergency dispatch station closest to the selected landing site based on the type of aircraft (i.e., multicopter or tiltrotor).

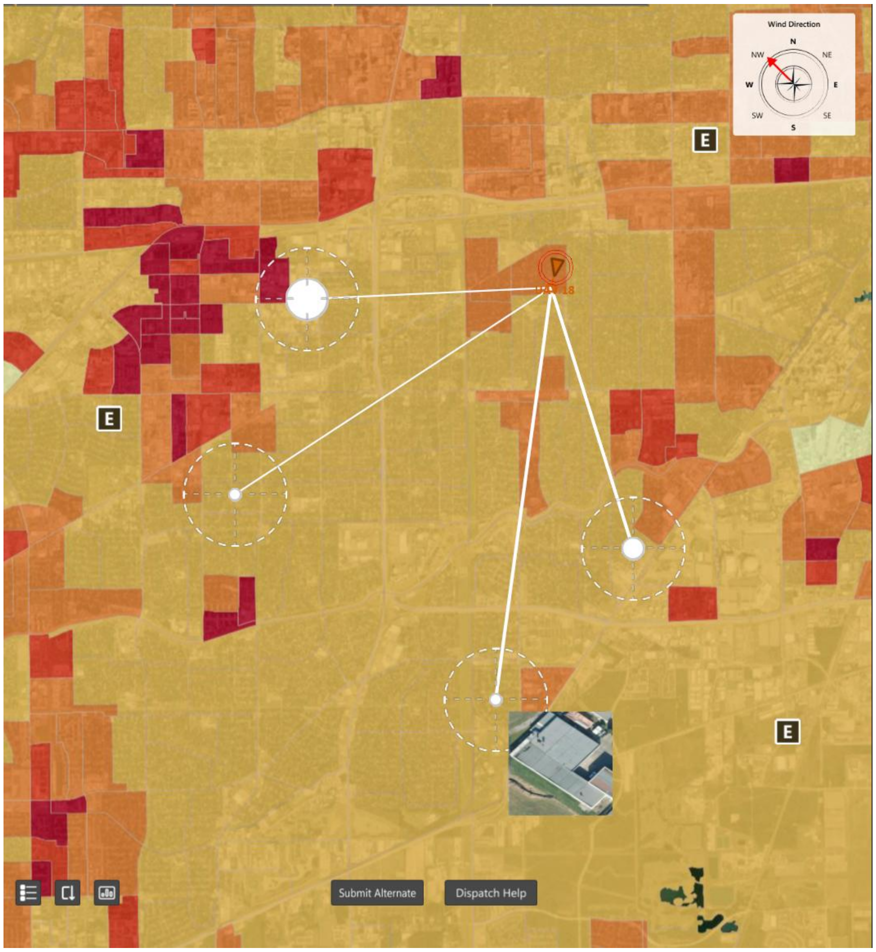

Figure 2.4 Alternate landing site task interface. Participants selected an optimal landing site based on population density (represented by yellow to red color range for low to high population density), landing site risk (size of landing site circle), landing site type (presence of a flat roof or grass area), and path risk (path thickness).

Vehicle authorization (verbal). The vehicle authorization task (**Figure 2.5**) involved categorizing unidentified aircraft as authorized or unauthorized to fly through operator-designated airspace based on aircraft characteristics such as the callsign, altitude, vehicle type, and assigned permits. Participants then typed a brief justification for the assessed authorization status.

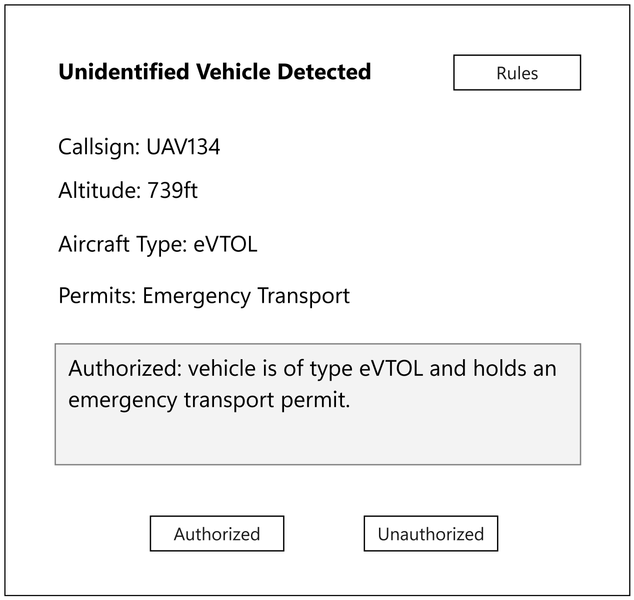

Figure 2.5 Vehicle authorization task interface. Participants determined whether unidentified vehicles were authorized to fly through operator-designated airspace based on vehicle attributes (callsign, altitude, aircraft type, and the permits carried). Background colors were removed to improve legibility).

Request for information (verbal). The request for information task (**Figure 2.6**) asked participants to report on two to three items related to the current vehicle status, such as fuel level, GPS signal strength, communication link strength, level of turbulence, or name of next waypoint. As envisioned in the proposed concept of operations for Urban Air Mobility (FAA, 2020), external service providers may request this information from vehicle operators to ensure compliance with regulatory and operational requirements. To gather the requested information, participants referred to the health information panel of the target UAV (shown in **Figure 2.7**) and typed a response in a chat box displayed in the interrupting task area on the right-hand monitor (see **Figure 2.1**).

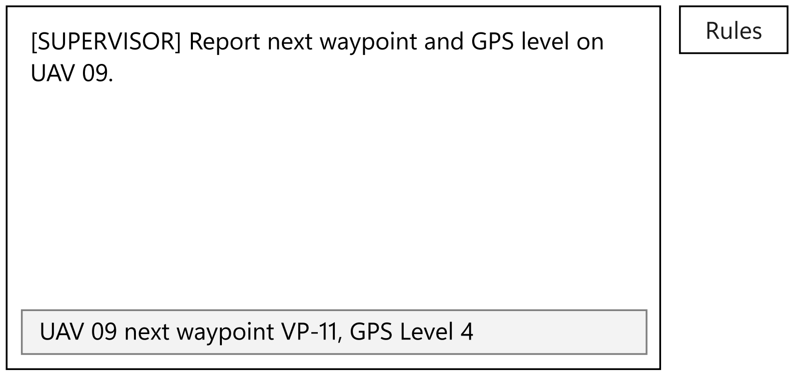

Figure 2.6 Request for information task interface. Participants reviewed requests for information on specified vehicles (e.g., UAV 09 as pictured), gathered the required information based on the vehicle's flight progress and health status (see Figure 2.7), and typed an appropriate response into a chat box (a sample response is shown at the bottom of the figure). Background colors were removed to improve legibility.

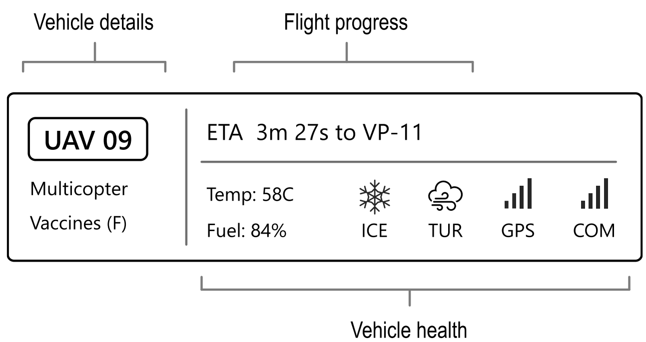

Figure 2.7 One of the 20 UAV health information panels displayed on the right monitor. To the left of the vertical bar are the UAV callsign, vehicle type, and onboard cargo. To the right are the estimated time of arrival to the next waypoint, vehicle temperature, fuel level, icing indicator, turbulence indicator, GPS level, and communications level.

2.2.3 Experiment Design

The study featured a 2 (interruption frequency: low, high) x 3 (nesting level: single, serial, nested) x 2 (processing code similarity: same, different) x 3 (interruption urgency: low, medium, high) within-subjects, fractional factorial design. The processing code of the ongoing flight request task was verbal, and the processing code of the interrupting task either alternated between spatial and verbal (i.e., a verbal task interrupted by a spatial task, followed by a verbal task), or was fixed to be verbal only (i.e., all interrupting tasks required verbal processing). The urgency level was fixed to medium for the ongoing task, low/medium/high for single interruptions, and medium-medium or medium-high for serial and nested interruptions. Low urgency interruptions were not presented in the serial and nested conditions to incentivize participants to switch from the medium urgency flight request task to an incoming medium or high urgency interrupting task. Because participants were told to hand off tasks that were of lower urgency than the ongoing task, inclusion of trials with low urgency serial and nested interruptions a) would not have resulted in meaningful data for those trials, and b) would have resulted in more factor combinations than could reasonably fit within two 30-minute scenarios. Overall, limiting the processing code and urgency levels for serial and nested interruptions allowed for a smaller, more targeted, set of conditions to be tested while retaining sufficient operational validity.

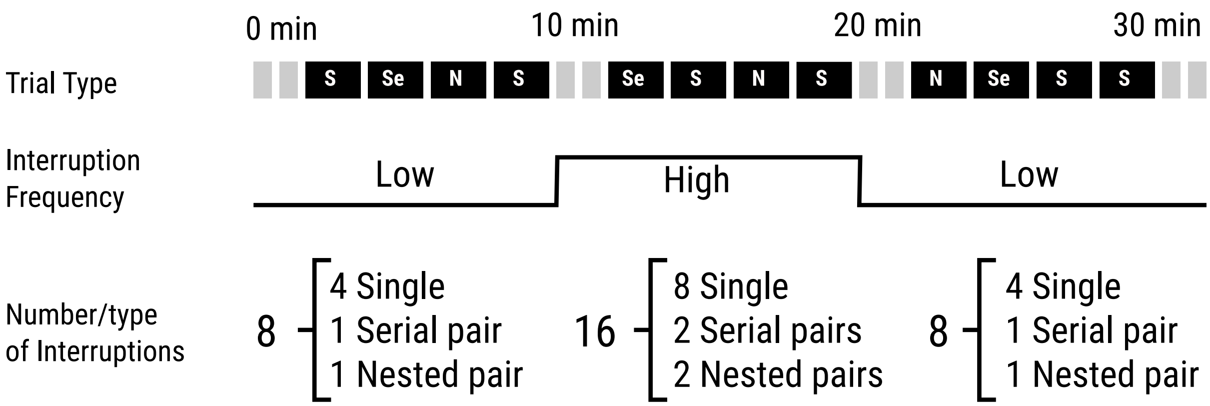

Figure 2.8 Design of one of two experiment scenarios, showing the sequence and number of ongoing tasks. In the first row, light gray bars indicate uninterrupted ongoing task trials, and black bars indicate interrupted trials. ‘S’, ‘Se’, and ‘N’ indicate single, serial, and nested trials, respectively.

Participants completed two scenarios, lasting approximately 30 minutes each. Each scenario included 20 flight request tasks and 32 interruptions (see **Figure 2.8**). Eight of the 20 flight request tasks were uninterrupted—two each at the beginning and end of the scenario, and four at the transitions between low and high interruption frequency. The uninterrupted flight request tasks were added as a buffer to mitigate carry-over effects from one phase to the next, and to calculate baseline performance without interruptions. Participants were not informed about how many and which tasks would be interrupted.

Due to the high complexity of this experiment, factors were manipulated in a pseudo-random order, i.e., the order of flight request tasks and interrupting tasks was randomized at the time of scenario development but was ultimately the same for all participants. The flight request task was designed to not have large variations in the sequence of steps taken. Different flight request tasks involved only small differences such as comparing different values associated with route characteristics, adding two battery modules instead of one, etc. To better account for any correlations between performance and the fixed sequence of tasks, participants performed two different scenarios, each with a different ordering of independent factors, and each with a counterbalanced order of completion. Finally, each scenario included periods of no interruptions at transitions between low and high frequency phases. These uninterrupted periods served as a buffer to mitigate carry-over effects from one interruption frequency phase to the next.

2.2.4 Procedure

Participants attended one three-and-a-half-hour session that involved simulator training, three practice scenarios, and two experiment scenarios. Before beginning the training portion, participants completed a pre-experiment questionnaire (**Appendix A**) on gaming experience and multitasking ability (adopted from Basoglu et al., 2009). After the pre-experiment survey, participants completed an approximately one-hour long PowerPoint-based training session on managing UAV fleet operations for transporting cargo. Immediately after learning how to perform each task, participants were guided by the experimenter to practice it in the UAV-simulator. After training, participants completed a second scenario to again practice each of the five tasks three times, separately and without any interruptions. After a five-minute break, participants completed a third practice scenario with interruptions. Following another five-minute break, each participant completed two experiment scenarios (counterbalanced) with a 10-minute break in-between. The third practice scenario and the two experiment scenarios were completed while wearing the Tobii Pro Glasses 2 eye tracking glasses.

During each scenario, participants completed 20 flight request tasks. Each new flight request task appeared five seconds after completion of the previous one. While participants completed the flight request task, they were notified when an interrupting task arrived. This occurred twice per ongoing flight request task during the low interruption frequency phase, and four times per flight request task during the high frequency phase. Each interruption was announced first by an auditory chime. Participants were asked to acknowledge this interruption signal by pressing the foot pedal. Once the potential interruption had been acknowledged, a visual notification of the pending task appeared in the notifications panel (**Figure 2.2**, on the right). The notification contained information on the callsign of the affected UAV, the type of task to be performed, and the name of the cargo on board. Also contained within each notification box were three buttons to “Accept,” “Delay,” or “Reject” the pending task. Participants were instructed to a) prioritize and accept as soon as possible high-urgency interruptions over the medium-urgency flight request task, b) prioritize the medium urgency flight request tasks over low-urgency interruptions, and c) to use their discretion to schedule interrupting tasks of the same urgency level as the ongoing task.

Interruption notifications stayed in the notifications panel only for a limited period of time. Participants were asked to switch to pending tasks of high, medium, and low urgency within 15, 30, and 60 seconds of the initial notification chime, respectively. Clicking the accept button displayed the interrupting task on the right monitor, and the participant was expected to complete the task right away. Clicking the delay button showed a timer in the top right-hand corner of the associated notification. This timer provided a countdown for the number of seconds remaining until notification expiration. The delay option indicated that the participant intended to complete the task at a later time (e.g., after reaching a breakpoint in the current task). Clicking the reject button immediately removed the notification from the notifications panel as this response indicated that the participant did not intend to complete that task. A task could be rejected, for example, when an ongoing task of medium or high urgency was interrupted by a low urgency task. Participants were told that rejected tasks would be handed off to another operator. Once an interrupting task was accepted, both the interrupting task and the ongoing flight request task were visible at the same time. Participants could switch back and forth as desired. If more than one interrupting task was accepted, only the most recently accepted interrupting task was visible.

The total number of interruptions presented in the low (20-minute) and high (10-minute) frequency periods was the same. 16 interruptions were presented in each phase, composed of eight single interruptions, two serial interruption pairs (four interruptions), and two nested interruption pairs (four interruptions). As shown in **Figure 2.9**, interruptions were presented relative to the participants' progress on the ongoing new flight task—that is, between three to six seconds after making the UAV destination/cargo selection (near the beginning of the first half of the new flight task), and/or between three to six seconds after the start of typing in the notes section (near the beginning of the second half of the new flight task). These points were chosen to coincide with the demanding parts of the new flight task, including the selection of the optimal route and relaying safety-critical information to the destination vertiport. During the high frequency phase, an additional single interruption was presented 10-15 seconds after the end of each single interruption between the destination/cargo selection and the notes section. For the serial condition, the second interruption was presented one second after the end of the previous task in the interruption pair. For the nested condition, the second interrupting task was presented three to six seconds after the start of the previous interrupting task. Nested interruptions were presented at no more than one level of depth, as shown in **Figure 2.9,** but participants were free to choose when and which tasks to switch between.

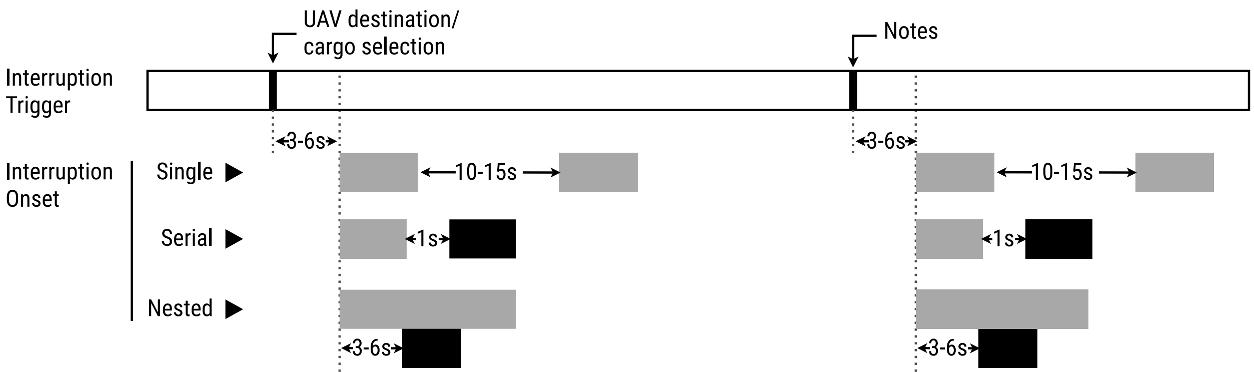

Figure 2.9 Timing and sequence of single, serial, and nested interruptions presented during the high frequency period. Interruptions were triggered by one of two actions (UAV destination/cargo selection or Notes) taken during the ongoing flight request task. Gray bars represent secondary interrupting tasks, and black bars indicate tertiary interrupting tasks in a pair of serial or nested interruptions.

At the end of each scenario, participants completed a NASA Task Load Index (NASA-TLX) survey to assess perceived workload (**Appendix B**). Each scenario was screen-captured for use in a cued debrief at the end of the experiment. The debrief interview was conducted to gain insight into the participants' strategies, challenges, and thought processes in handling interruptions (**Appendix C**). The interviews were voice-recorded using an off-the-shelf handheld voice-recorder and transcribed at the conclusion of the experiment's data collection phase using a third-party transcription service.

2.2.5 Dependent Measures

The dependent measures included task performance, eye tracking metrics, survey responses, and transcripts from post-experiment debrief interviews. Performance metrics consisted of notification detection rate and acknowledgement time, interpretation accuracy, resumption lag, and accuracy on task. Survey data included questions about the participants' self-reported interruption management self-efficacy (IMSE) and multi-tasking computer self-efficacy (MTCSE; borrowed from Basoglu et al., 2009), gaming experience, and perceived workload (NASA-TLX). Eye tracking metrics, collected using the Tobii Pro Glasses 2 eye tracker, comprised visit duration and pupil diameter. The latter was used as a measure of perceived mental workload (Cain, 2007; Longo et al., 2022; Recarte et al., 2008).

2.2.6 Assumptions

The following assumptions were made regarding the experiment setup and during the analysis of the data:

- The operator handles only interruptions from vehicles under their supervision (i.e., one source of interruptions), rather than from both vehicles and other operators working towards a similar goal in a team-based environment.
- The tasks employed in this study are different only in terms of the processing code (spatial or verbal) but are comparable in terms of both difficulty and time needed for completion.
- Participants are trained sufficiently and followed task instructions related to how to interpret notifications and handle interrupting tasks.
- Performance on each flight request task trial is independent of other trials (each trial was presented with a 5-second delay in between) and the two uninterrupted

periods included at the transitions between low and high frequency periods prevented carry-over effects between phases.

- Overt visual attention (e.g., visit duration in the notifications panel) is equivalent to active processing of information in the area of interest.

2.3 Results

Analyses were performed using linear mixed effects models with random effects in the lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017) R-based packages. Participant ID was used as a random effect in all models. Main effects of independent variables were computed using Chi-squared tests to compare the null model (e.g., $Acknowledgement\ Rate \sim 1 + (1 \mid ParticipantID)$) and a model containing the fixed effect of interest (e.g., $Acknowledgement\ Rate \sim Frequency + (1 \mid ParticipantID)$). Alternate models were developed and compared using a similar technique—more than one fixed effect was included in the same model (e.g., $Acknowledgement\ Rate \sim Frequency + Nesting\ Level + (1 \mid ParticipantID)$) if including the fixed effect significantly improved the model fit (i.e., the Chi-squared test yielded a significant p-value less than 0.05). Scenario ID was also included as a random effect *within* each participant (e.g., $Acknowledgement\ Rate \sim Frequency + (ScenarioID \mid ParticipantID)$) if it improved the model fit. All significance levels in figures and tables are reported as *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$.

2.3.1 Notification Acknowledgement

We expected that participants would be less likely and slower to detect interruption notifications during high frequency periods and for nested interruptions, compared to low frequency periods, and serial/single interruptions, respectively (H1). Acknowledgement rate was calculated as the proportion of notifications acknowledged, out of total presented. Acknowledgement time was defined as the time between the onset of the auditory chime, and activation of the foot pedal.

Generalized linear mixed effects models were fitted for acknowledgement rate (binomial distribution with logit link function), and acknowledgement time (gamma distribution with log link function). The selected models included frequency and nesting level as fixed effects, and participant and scenario as random effects (see **Table 2.3**).

Table 2.3 Summary of two generalized linear mixed effects models fitted for notification acknowledgement rate and acknowledgement time.

	Acknowledgement Rate					Acknowledgement Time				
Predictors	*Odds Ratio*	*SE*	*95% CI*	*z*	*p*	*β*	*SE*	*95% CI*	*z*	*p*
Intercept (Low/Single)	351.96	187.31	124.02 – 998.86	11.02	<0.001	1.78	0.18	1.45 – 2.18	5.59	<0.001
Frequency - High	**0.37**	**0.10**	**0.22 – 0.63**	**-3.66**	**<0.001**	1.05	0.03	1.00 – 1.10	1.85	0.065
Nesting Level - Serial	1.10	0.47	0.48 – 2.53	0.24	0.813	**0.72**	**0.03**	**0.67 – 0.78**	**-8.52**	**<0.001**
Nesting Level - Nested	**0.28**	**0.09**	**0.15 – 0.51**	**-4.12**	**<0.001**	**1.09**	**0.04**	**1.01 – 1.18**	**2.26**	**0.024**
Observations	2560					2478				
Marginal / Conditional R^2	0.057 / 0.558					0.010 / 0.277				

There was a main effect of interruption frequency and nesting level on acknowledgement rate (see **Figure 2.10**). Participants were less likely to acknowledge interruption notifications during high frequency periods (M = 95.6%, SE = 1.58; z = -3.66, p < 0.001), compared to low frequency periods (M = 98.0%, SE = 0.88). Nested interruptions (M = 93.1%, SE = 2.37) were less likely to be acknowledged (z = -4.12, p < 0.001), compared to single interruptions (M =

97.3%, *SE* = 1.13). There was no significant difference in the rate of acknowledgement between single and serial interruptions (*M* = 97.5%, *SE* = 1.02; $z = 0.24$, $p = 0.970$). No interaction effects were observed between interruption frequency and nesting level on the rate of acknowledgement.

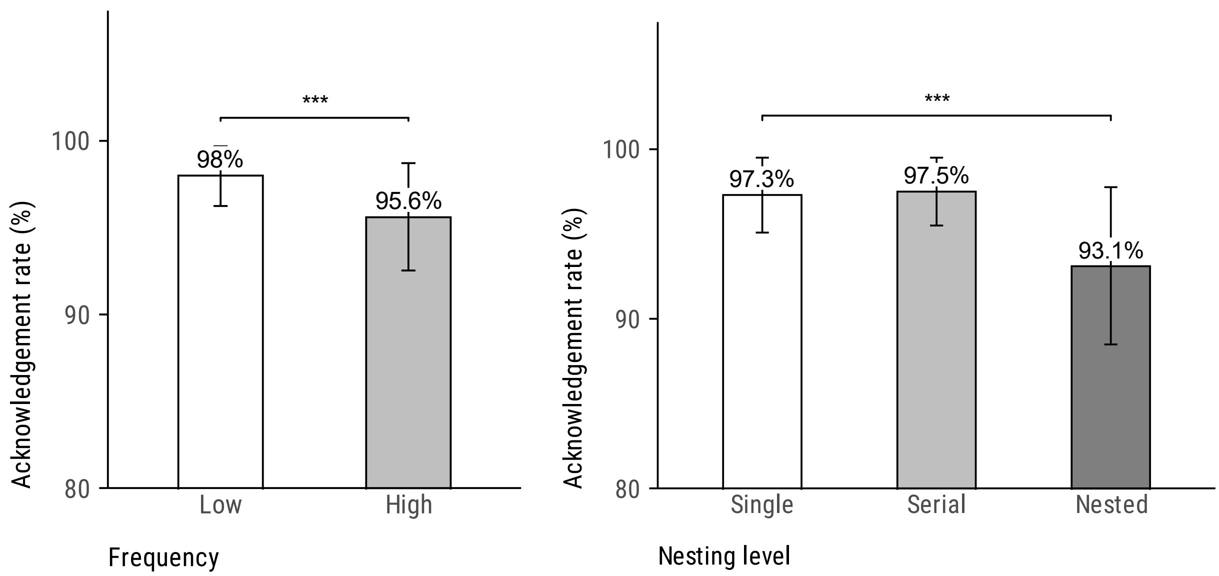

Figure 2.10 Acknowledgement rate shown as a function of interruption frequency (left) and nesting level (right). Error bars show 95% CI.

Average interruption acknowledgement time was not significantly different between the low (*M* = 1.65s, *SE* = 0.17) and high interruption frequency phases (*M* = 1.73s, *SE* = 0.18; $z = 1.82$, $p = 0.065$). However, compared to single interruptions (*M* = 1.82s, *SE* = 0.19), participants were 27% (502ms) *faster* when acknowledging serial interruptions (*M* = 1.32s, *SE* = 0.14; $z = -8.52$, $p < 0.001$), and 11% (168ms) *slower* when acknowledging nested interruptions (*M* = 1.99s, *SE* = 0.215; $z = 2.26$, $p = 0.024$; see **Figure 2.11**). There was no interaction between interruption frequency and nesting level on acknowledgement time.

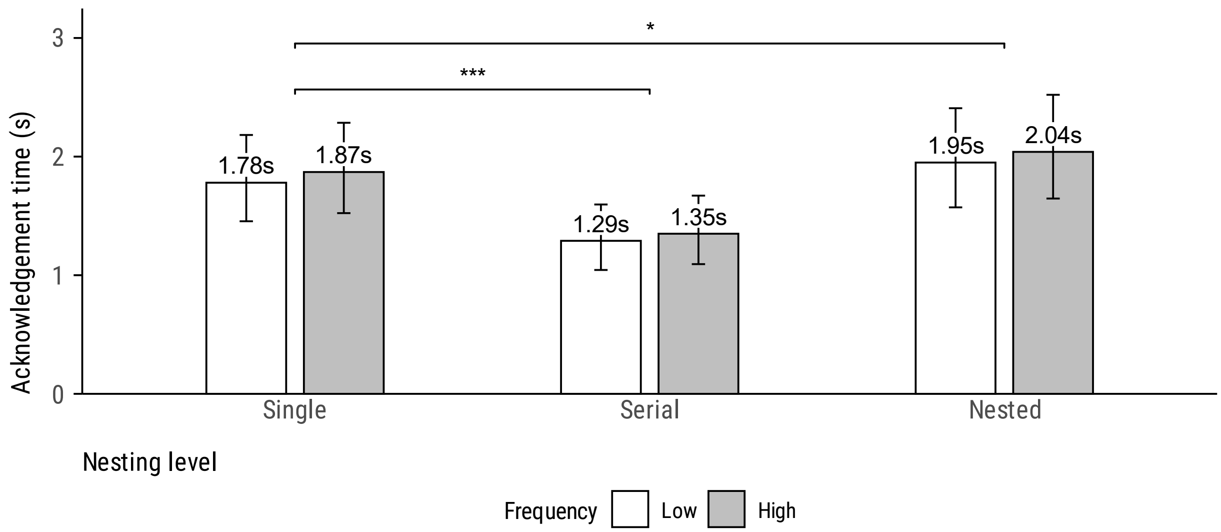

Figure 2.11 Acknowledgement time shown as a function of interruption frequency and nesting level. Error bars show 95% CI.

Lastly, as an exploratory measure, both acknowledgement rate and acknowledgement time were compared as a function of presentation time. Participants were significantly less likely to acknowledge notifications when they were presented towards the beginning of the ongoing flight request task (i.e., relative to the dropdown trigger; M = 98.0%, SE = 0.69; β = 0.3, *95% CI* [0.18, 0.49], z = -4.71, p < 0.001), compared to when they were presented towards the end of the task (i.e., relative to the start of the notes; M = 94.8%, SE = 2.23). A similar pattern was observed with acknowledgement time. Participants were slower to acknowledge interruption notifications when they were presented towards the end of the flight request task (M = 1.91s, SE = 0.20; β = 1.14, *95% CI* [1.09, 1.2], z = 5.06, p < 0.001), compared to the beginning (M = 1.67s, SE = 0.18).

2.3.2 Notification Interpretation

Interpretation accuracy was approximated based on whether high urgency interrupting tasks were accepted before expiration. Only high urgency interruptions were considered in the evaluation of interpretation accuracy because a) participants were given explicit instructions to switch to high urgency interrupting tasks as soon as possible (unlike medium urgency interruptions where participants used their discretion to decide whether to switch to an interrupting task), and b) to allow comparisons between all three nesting levels, where the serial and nested conditions did not include any low urgency interruptions. A generalized mixed linear effects model with nesting level as a fixed effect was used to compare interpretation accuracy.

We expected that participants would be less accurate when interpreting interruption notifications during high frequency periods, and when interruptions are nested, compared to low frequency and single/serial interruptions, respectively (H2). There was a main effect of nesting level ($\chi^2(2) = 25.74$, $p < 0.001$), but not frequency ($\chi^2(1) = 2.47$, $p = 0.116$). Participants were significantly more accurate when interpreting serial interruptions ($M = 94.2\%$, $SE = 1.87$), compared to both single ($M = 84.0\%$, $SE = 3.00$; *Odds Ratio (OR)* = 0.23, 95% *CI* [0.09, 0.57], $z = -3.19$, $p = 0.002$) and nested ($M = 76.9\%$, $SE = 3.54$; $OR = 0.14$, 95% *CI* [0.06, 0.34], $z = -4.36$, $p < 0.001$) interruptions. There was no difference in interpretation accuracy between single and nested interruptions ($OR = 0.60$, 95% *CI* [0.76, 3.72], $z = 1.53$, $p = 0.279$). Interpretation accuracy was lower for nested interruptions despite the fact that participants spent more time reviewing the associated notifications (**Figure 2.12**). Total visit duration in the notifications panel was significantly longer during only the high interruption frequency period for nested interruptions ($M = 7.92$s, $SE = 0.53$), compared to both single ($M = 5.50$s, $SE = 0.53$; $\beta = -2.42$, 95% *CI* [-3.29, -1.54], $z = -5.45$, $p < 0.001$) and serial ($M = 5.56$s, $SE = 0.53$; $\beta = -2.36$, 95% *CI* [-3.23, -1.48], $z = -5.31$, $p < 0.001$) interruptions.

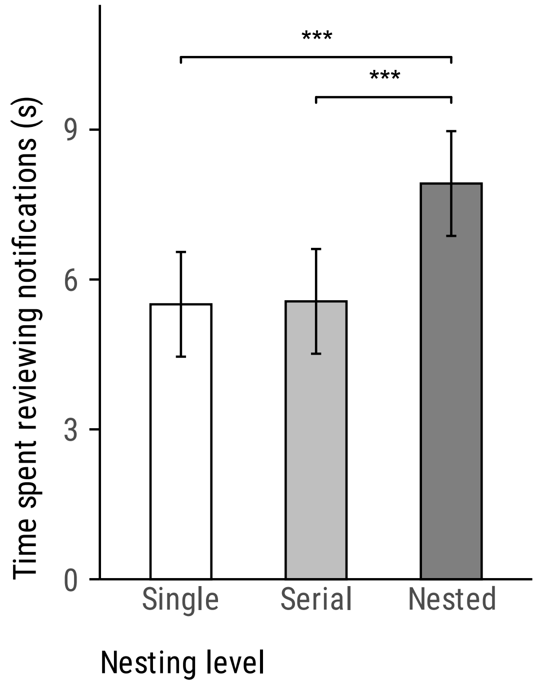

Figure 2.12 Time spent reviewing interruption notifications per flight request task for single, serial, and nested interruptions during the high interruption frequency phase. Error bars show 95% CI.

2.3.3 Task Accuracy as a Function of Interruption Frequency and Nesting Level

Because participants were free to choose how to respond to interrupting tasks, the actual (as opposed to planned) nesting level of interruptions had to be calculated for each participant using the start and end time of interrupting tasks relative to the flight request task. Total number of tasks completed, across all participants, were 714 (73% of planned) for single, 258 (30%) for serial, and 116 (15%) for nested. Task accuracy was defined as the number of steps completed accurately, out of the total number of steps involved in a task. For the flight request task, there were 8 steps: selecting the correct cargo, destination, and optimal route, loading the correct number of battery modules, loading the cargo onto the vehicle, uplinking the route, and completing the checklist. Baseline accuracy on the flight request task was calculated from trials with no interruptions. For interrupting tasks, the number of steps involved ranged from one to four.

Our expectation was that participants would perform worse on the interrupted task during high frequency periods, and that this effect would be more pronounced when interruptions were nested (H3). Accuracy on the flight request task was indeed worse during high frequency periods (*M* = 91.8%, *SE* = 0.89; β = -2.08, 95% *CI* [-3.80, -0.36], *z* = -2.37, *p* = 0.018), compared to low frequency phases (*M* = 93.8%, *SE* = 0.80). Participants scored higher on the flight request task in the serial condition (*M* = 95.0%, *SE* = 1.09; **Figure 2.13**), compared to the single (*M* = 92.4%, *SE* = 0.64; β = -2.67, 95% *CI* [-4.89, -0.44], *z* = -2.35, *p* = 0.019) and nested trials (*M* = 91.0%, *SE* = 1.51; β = -4.02, 95% *CI* [-7.47, -0.57], *z* = -2.29, *p* = 0.022), irrespective of interruption frequency. There were no interaction effects between interruption frequency and nesting level.

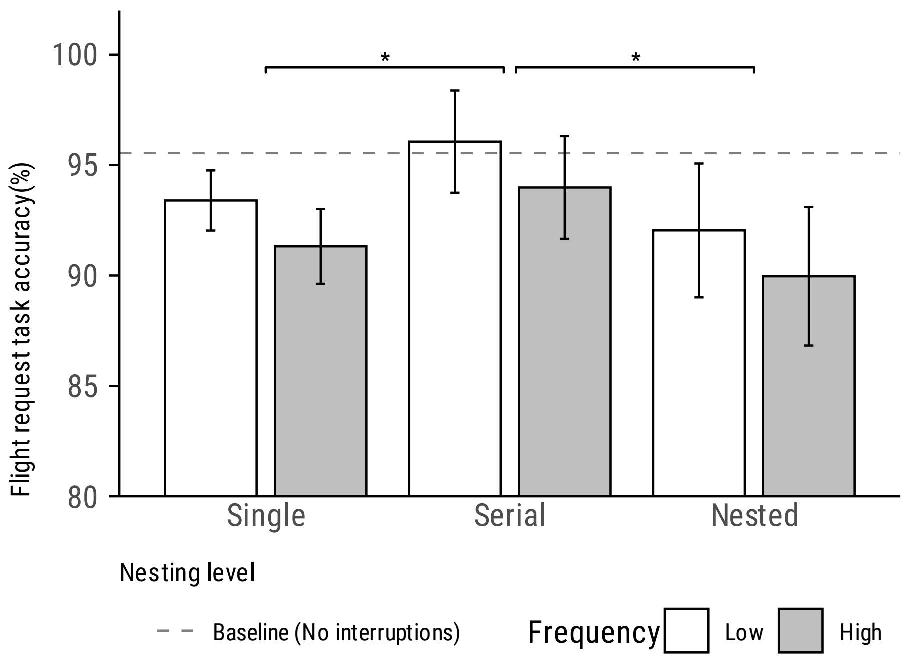

Figure 2.13 Accuracy on the flight request task during low and high interruption frequency periods for single, serial, and nested interruptions. Error bars show 95% CI.

Types of errors on the flight request task included failing to select the optimal route, failing to add a sufficient number of battery modules if the turbulence along route was high, or failing to complete appropriate items in the checklist before submitting the flight intent. Incorrect selection of the optimal route was generally a result of either prioritizing route parameters in the wrong order (e.g., due to forgetting to check the vehicle and/or cargo type), or due to incorrect sorting of routes based on an incorrect comparison of distance and separation values. The latter seemed to occur because some participants developed a strategy to compare only the digits after the decimal point, rather than comparing both the whole number and the values after the decimal (e.g., considering a value of 34.75 miles to be a longer than 35.25 miles).

For interrupting tasks, we expected that performance would be worse during high frequency periods, compared to low frequency periods (H3). Accuracy on interrupting tasks was indeed significantly lower during high frequency periods (*M* = 84.1%, *SE* = 1.54; β = -3.97, 95% *CI* [-6.81, -1.13], *z* = -2.74, *p* = 0.006), compared to low frequency (*M* = 88.0%, *SE* = 1.44). The data were analyzed further to determine whether accuracy on the secondary (a single interrupting task, or the first of a pair of interrupting tasks in the serial and nested conditions) and tertiary tasks (the second interrupting task presented during the serial and nested conditions) differed between the single, serial, and nested conditions. There was no difference in accuracy on the secondary interrupting task. On the tertiary task, however, participants performed worse on nested interruptions (*M* = 82.1%, *SE* = 3.28; β = -9.94, 95% *CI* [-17.33, -2.55], *z* = -2.64, *p* = 0.008), compared to serial interruptions (*M* = 92.0%, *SE* = 2.30).

Common errors included forgetting to prioritize turbulence over noise and vice versa based on cargo type (vertiport diversion task), forgetting to verify that the selected alternate landing site has a compatible roof type/landing area (alternate landing site task), including only a partial explanation for one of the two possible rules for why an aircraft was authorized to be in the operator's airspace (vehicle authorization task), and not including all attributes of flight progress requested by the supervisor (request for information task).

2.3.4 Task Resumption on Flight Request Tasks

Task resumption on flight request tasks was examined using resumption lag which is defined as the time taken to perform the first action on the interrupted task after returning from an interrupting task. We expected that participants would take longer to resume working on the flight request task when returning from nested interruptions, and from interruptions with the same processing code as the interrupted task (H4). There was a main effect of processing code similarity (i.e., whether or not the interrupting tasks required switching between verbal and spatial tasks), but not nesting level (**Figure 2.14**). Post-hoc Tukey tests revealed that, in the case of single interruptions only, participants resumed more quickly in the same processing code condition (M = 5.99s, SE = 0.34; β = -1.19, 95% CI [-1.96, -0.42], t(767) = 3.03, p = 0.031), compared to when the processing code was different (M = 7.18s, SE = 0.43).

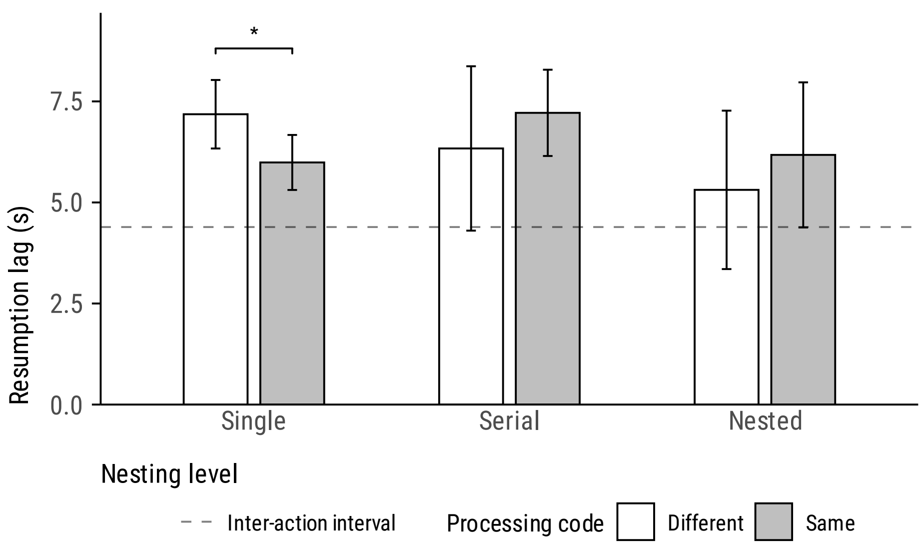

Figure 2.14 Resumption lag, in seconds, when returning from single, serial, and nested interruptions to the new flight task for both the processing code switch and no switch trials. Error bars show 95% CI.

To explore potential differences in perceived mental workload when returning to the flight request task from spatial versus verbal interrupting tasks, participants' pupil diameter was compared during the five-second period following the return. An increase in pupil diameter, compared to each participant's baseline pupil diameter without interruptions, indicated a rise in perceived mental workload while a decrease suggested a reduction in mental workload. Changes in pupil diameter were significantly larger in the same processing code condition (M = 0.145mm, SE = 0.02; β = 0.31, 95% CI [0.27, 0.34], z = 17.51, p < 0.001), compared to when the processing code was different (M = -0.141mm, SE = 0.02; see **Figure 2.15**). Post-hoc comparisons were conducted to explore interactions between processing code similarity and nesting level. Compared to serial interruptions, the drop in pupil diameter was significantly larger during the same processing code condition for both single ($t(731)$ = -17.50, p < 0.001) and nested interruptions ($t(737)$ = -6.35, p < 0.001).

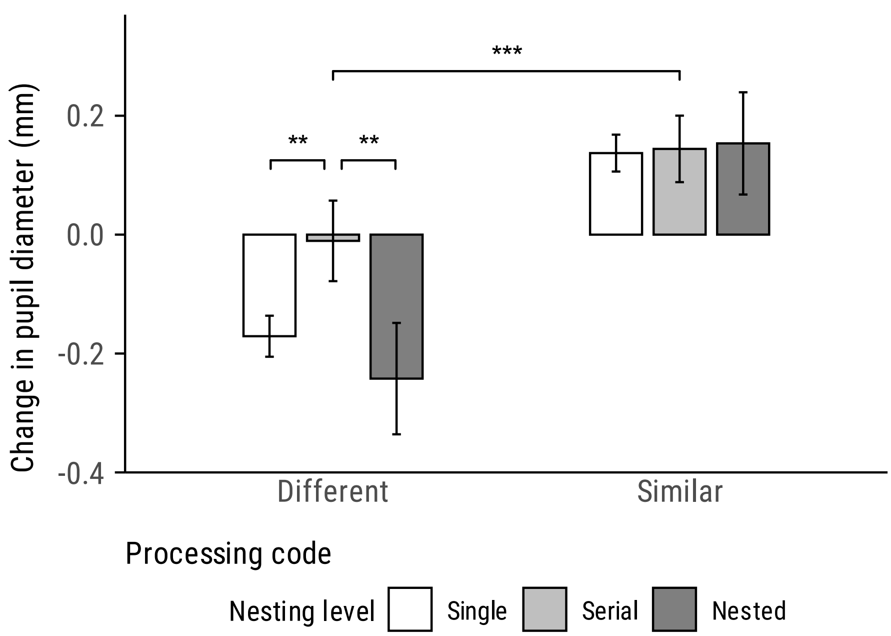

Figure 2.15 Change in pupil diameter when returning from single, serial, and nested interruptions, shown as a function of processing code similarity. Error bars show 95% CI.

2.3.5 Survey and Debrief Findings

We expected that participants would cope with the increase in workload from frequent and nested interruptions by simplifying their strategy to switch to interrupting tasks. During post-experiment debrief interviews, half (20) of all participants reported changing their strategies and behavior during high frequency periods and for nested interruptions such that they a) focused on completing the ongoing task before switching (6), b) only switched to pending tasks if they were of higher urgency than the current and ignored equal and lower urgency tasks (7), or c) only switched at specific breakpoints during the flight request task (e.g., after completing route selection; after loading battery modules, uplinking the selected route, and loading cargo item; and after completing the notes), regardless of interrupting task urgency (7).

We further compared the strategies of participants with overall performance (defined as the combined accuracy on all completed tasks, weighted by the time taken to complete them) in the top and bottom quartiles. Participants with poor overall performance rejected tasks more frequently than the top-performers who rarely rejected tasks and reliably used the delay feature to keep track of the notifications' expiration time. One participant in the worst-performing group mentioned that he stopped delaying tasks because he believed that not using the timer allowed for more time to accept the interrupting task. In reality, the amount of time a notification stayed in the notification panel was the same regardless of whether the task was delayed or not. Delaying a task only made visible a countdown timer that indicated the number of seconds until the notification expired. Additionally, the worst performers were three times more likely to refer to task instructions (which could be accessed in each task at any time by pressing a button in the UAV simulator), compared to the best-performing participants who made a deliberate effort to memorize the task rules and prioritization scheme in order to minimize the need to refer back to them frequently, and, in turn, minimize unnecessary interruptions of task flow.

Participants' overall task performance was compared also as a function of gaming experience and multitasking ability. Participants who reported playing games more frequently did not perform better than those who reported to rarely or never play games. Similarly, the IMSE and MTCSE scores showed no significant effect on overall performance. To explore possible other reasons explaining the largest observed differences in performance, we next compared only the five best and five worst performing participants in terms of IMSE and MTCSE scores, gaming experience, and interruption management strategies. The bottom five performers, on average, had higher IMSE (M = 0.74) and MTCSE (M = 0.77) scores compared to the best performers ($M_{IMSE} =$ 0.68, $M_{MTCSE} =$ 0.71), indicating that the participants overestimated their interruption management and multitasking abilities.

Table 2.4 NASA-TLX workload ratings (with standard error) measured on a 10-point Likert item response scale ranging from low (1) to high (10).

NASA-TLX	First Scenario	Second Scenario
Frustration Level	4.632 (±0.531)	4.733 (±0.342)
Task Difficulty	7.000 (±0.359)	7.348 (±0.202)
Mental Effort	7.474 (±0.353)	7.122 (±0.244)
Physical Effort	4.368 (±0.563)	4.275 (±0.387)
Time pressure	7.473 (±0.414)	6.697 (±0.236)
Performance	4.737 (±0.621)	4.238 (±0.383)

During the debrief interviews, participants frequently reported performing better on the second scenario than the first. To examine possible effects of learning on perceived workload and task accuracy, NASA-TLX ratings and task scores were compared between the first and second scenarios. NASA-TLX ratings (summarized in **Table 2.4**) were not different between the first and second scenarios. Accuracy on the flight request task, however, was significantly better in the second scenario, compared to the first ($t(867) = 2.17$, $p = 0.030$; see **Figure 2.16**). Post-hoc Tukey HSD comparisons revealed that accuracy on the flight request task was significantly better only for nested interruptions in the second scenario ($M = 92.6\%$, $SE = 0.86$; $\beta = 6.03$, *95% CI* [1.47, 10.59], $t(939) = -3.78$, $p = 0.002$), compared to the first ($M = 86.6\%$, $SE = 1.43$).

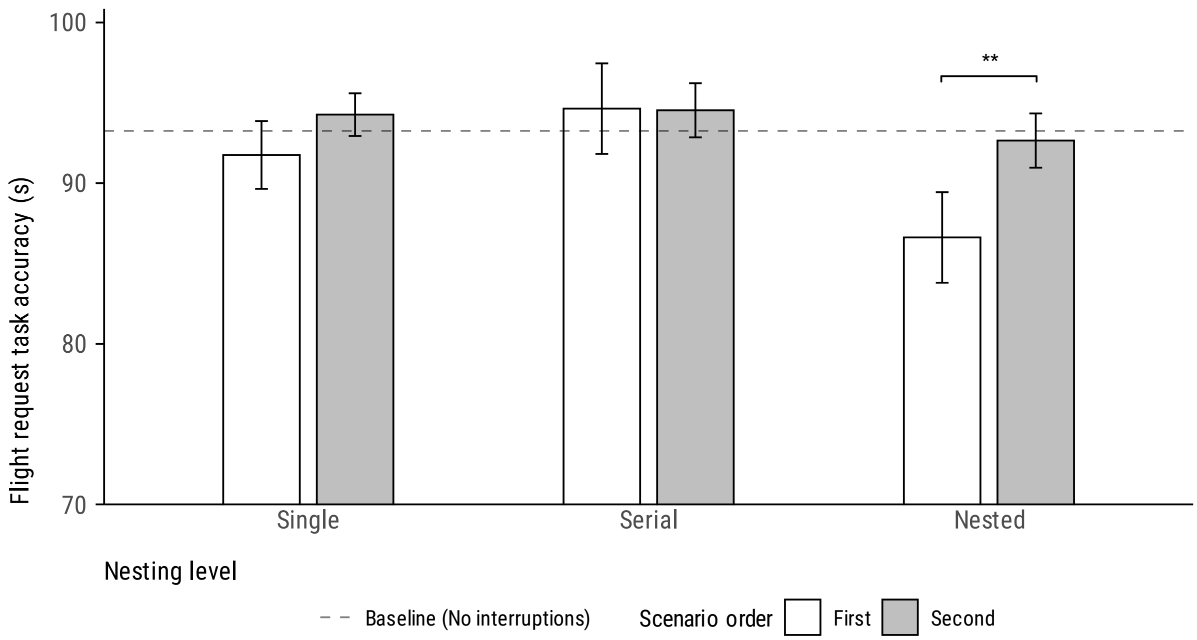

Figure 2.16 Accuracy on the new flight task in the first and second scenario for the single, serial, and nested interruption conditions. Error bars show 95% CI.

Table 2.5 Summary of study expectations and results.

Expectation	**Results**
H1. Participants will be less likely and slower to acknowledge interruptions during high frequency periods and interruptions that are nested.	Frequent and nested interruption notifications were significantly less likely to be acknowledged, compared to both single and serial notifications. Acknowledgement time was not significantly different between the high and low interruption frequency phases. Compared to single interruptions, participants acknowledged serial notifications more quickly, and nested interruptions more slowly.
H2. Participants will be less accurate at interpreting interruption notifications during high frequency periods and when interruptions are nested.	There was no main effect of interruption frequency on interpretation accuracy. Interpretation accuracy was lower for single and nested interruptions, compared to serial interruptions. Participants spent more time reviewing nested interruption notifications, compared to single and serial notifications.
H3. Participants will perform worse on the ongoing and interrupting tasks during high frequency	Accuracy on the ongoing flight request task was lower during high frequency periods compared to low frequency, and for both single and

interruption periods. This effect will be more pronounced when interruptions are nested.	nested interruptions, compared to serial interruptions. Accuracy on interrupting tasks was lower during high frequency interruptions, and during nested conditions for the tertiary task (the second interrupting task in a pair of serial or nested interruptions).
H4. Participants will take longer to resume the ongoing task when returning from nested interruptions and from interruptions with the same processing code as the interrupted task.	Resumption lag was not observed to be longer when returning from nested interruptions. For single interruptions only, resumption lag was *shorter* when returning from interrupting tasks with the same processing code.
H5. Participants will cope with more frequent and nested interruptions by simplifying their strategy to switch to interrupting tasks.	Participants reported simplifying their strategies and behavior during high frequency periods and for nested interruptions such that they a) focused on completing the ongoing task, b) only switched to higher urgency tasks, or c) only switched at certain breakpoints.

2.4 Discussion

This experiment examined the effects of frequent and nested interruptions on all three stages of interruption management: detection, interpretation, and integration. As expected, frequent and nested interruptions were significantly less likely to be acknowledged (though they may have been noticed), compared to less frequent, and both single and serial notifications. A likely explanation for this finding is that, as posited by MFG, new goals (in this case, responding to a signal indicating a potential interruption of an ongoing task) must surpass a given interference level to be activated reliably and direct the operator's behavior (Altmann & Trafton, 2002). Given a higher level of "mental clutter" (Altmann & Trafton, 2002) due to residual goals, activation of the appropriate goal is more likely to fail. In the case of frequent interruptions, the increased interference likely resulted from the increased number of both recently completed and pending tasks. In the case of nested interruptions, the increased interference was likely a result of the need to switch back to the ongoing task and the first interrupting task. However, it is also possible that the lower rate of acknowledgement resulted from a deliberate decision to skip notification acknowledgment, rather than a limitation of attentional resources. Acknowledgment rate was discovered to be lower for notifications presented towards the end of the ongoing flight request task, compared to the beginning of the flight request task. In other words, participants may have deliberately suppressed incoming stimuli when nearing the end of an ongoing task and when preoccupied with an interrupting task (for a discussion on effects of interruption timing, see Bailey & Konstan, 2006; Czerwinski et al., 2000). This suggests that higher-level executive control had at least some influence over participants' acknowledgement behavior. As such, the lower rate of acknowledgement for high frequency and nested interruptions could result from a combination of limitations of attention and working memory and a deliberate coping mechanism to avoid the need to handle more interruptions.

Acknowledgement time was expected to be longer for nested interruptions and during the high interruption frequency period, as a result of increased interference from procedural and declarative memory. The results do not confirm the latter expectation—acknowledgement time did not differ between high and low interruption periods. Compared to single interruptions, participants were *faster* to acknowledge serial interruption notifications and *slower* to acknowledge nested ones. In the case of serial interruptions, participants had disengaged from an earlier interruption and were still in the process of returning to the interrupted task when the notification arrived. This lack of task engagement may have been responsible for their fast responses. Regarding nested interruptions, a high level of procedural and working memory interference (complete current task, return to interrupted task) may explain the slower acknowledgement times. However, like acknowledgement rate, participants delayed acknowledgement of interruptions by approximately 230ms when they were presented towards the end of the ongoing flight request task. Some participants noted a similar behavior during post-experiment debrief interviews. One participant, for example, mentioned that even after hearing the notification sound, they delayed acknowledgement until after they were done typing a word mid-sentence: "…when I was in the middle of writing the note and I would find that I would hear the noise and even before like hitting the pedal, I would finish a word or something." This again suggests that in addition to, or instead of interference from working memory, acknowledgement times may have been affected by top-down executive control.

Interpretation accuracy was expected to be lower during high frequency periods and for nested interruptions. This expectation was partially supported. Both single and nested interruptions had lower interpretation accuracy compared to serial interruptions, and participants spent more time reviewing nested interruptions, compared to single and serial. The higher interpretation accuracy for serial interruptions may again be explained by the fact that these interruptions were presented within 1-2 seconds of completing the previous interruption. During this period, participants were likely still in the process of switching back to (rather than having started) the flight request task and, as a result, a) had an easier time judging the urgency of the incoming task as it did not require comparing it to the urgency of the current task (unlike the nested interruption case), and b) were less resistant to switching to the incoming task rather than delaying it and risking expiration (unlike both the single and nested interruptions). In 68% of the cases where an interruption was misinterpreted, the associated notification expired (failure to hit "Accept" in time). In the remaining 32% of cases, the interruption was incorrectly rejected. Because participants were not asked to report the classification of urgency for each interruption notification, it remains unclear what portion of the 68% of expired notifications were interpreted correctly but not switched to in time. It seems that participants either a) struggled to classify and remember the relative urgency of the incoming interrupting task (i.e., problem with interpretation stage), or b) resisted switching away immediately from the current task and then failed to switch to it before it expired. The latter suggests a failure of integration, not interpretation. In those cases where the notification expired, participants tended to stay on the current task, especially if they were close to completing it. This behavior resembles the tendency observed by Salvucci and Bogunovich (2010) whose participants chose to "monotask" and stay on the current task when interruptions could be deferred. A follow-up study by Bogunovich and Salvucci (2011), as well as several earlier studies (e.g., Adamczyk et al., 2005; Iqbal & Bailey, 2005), show that even in the presence of time constraints, users prefer to continue with the current task and/or not be interrupted until a breakpoint (e.g., the end of a subtask) is reached.

For the integration stage, the data show that performance on the interrupted and interrupting tasks degraded slightly during high frequency interruptions. Results from past experiments on the effects of interruption frequency on the interrupted task are mixed. While some studies have found that increased interruption frequency degrades performance on the interrupted task (Lee & Duffy, 2015; Westbrook et al., 2010), others have found no difference or even the opposite effect (Drews et al., 2019; Monk, 2004; Speier et al., 1999). These apparent contradictions are addressed by Baethge et al. (2015) who note that a higher frequency of interruptions likely results in task performance that follows an inverted U curve (Yerkes & Dodson, 1908), where the cumulative increase in workload, while initially beneficial (e.g., due to increased motivation, exerted effort), eventually leads to performance breakdowns. Lee and Duffy (2015), for example, observed an approximately 14% drop in task accuracy during high frequency phases with three interruptions per ongoing task, compared to one interruption during low frequency phases. In contrast, participants in this study experienced only a 2% drop in accuracy when the ongoing task was interrupted four times versus twice. We did not observe complete breakdowns in primary task accuracy in this study likely because participants had some control over when to switch to an interrupting task, unlike the forced and immediate interruptions used by Lee and Duffy (2015) and most earlier work. This ability to choose a breakpoint (i.e., switch during periods of lower workload), while being aware of the impending interruption, has been shown to preserve performance on interrupted tasks and likely allowed participants to compensate for the high frequency of interruptions (Andrews et al., 2009; Bogunovich & Salvucci, 2011; Labonté et al., 2019; McFarlane, 1997; Monsell et al., 2003; Morgan et al., 2013; Rogers & Monsell, 1995). This conjecture is supported by debrief interviews where participants indicated that when interruptions became more frequent, they switched to interrupting tasks only at specific breakpoints, such as after making a route selection, regardless of task priority.

We observed better accuracy on the interrupted task for serial interruptions, compared to both single and nested. One possible explanation for this finding is the availability of more attentional resources to process interruption-related information due to the fewer number of goals participants had to keep in working memory for serial interruptions (i.e., complete current task, return to flight request task), compared to single (i.e., complete current task, return to new flight task, remember to interpret and respond to pending interruptions while working on the new flight task) and nested interruptions (i.e., complete current task, interpret and respond to nested interruption, return to secondary task, return to new flight task).

Accuracy on the interrupting tasks dropped significantly on the tertiary task during high frequency nested interruptions. This finding is different from that reported by Sasangohar et al. (2017) who observed no performance differences on the tertiary task between serial and nested interruptions for a medication entry task. This apparent contradiction may be, again, due to the manual task-switching paradigm employed in the current study wherein participants had control over when to switch between pending tasks. In contrast, participants in the Sasangohar et al.

(2017) study were transitioned back to the interrupted task automatically and only after a fixed amount of time had passed. The need to return to the interrupted secondary and ongoing flight request tasks in this study could have induced a sense of urgency in participants to quickly finish the current task, particularly during high interruption frequency periods. Participants seemed to confirm this sentiment during the post-experiment debrief. One participant, for example, stated that “just knowing that I had to do something else and having a timer as well, was an additional thing that I had to consider.”

Our data does not provide support for a negative effect of nesting level on the resumption of the interrupted flight request task. This result also contradicts findings by Sasangohar et al. (2017), who observed longer resumption lags on a medication entry task when returning from nested interruptions. It is possible that the long duration of interrupting tasks used in this study made rehearsal of problem state more difficult and inhibited its retrieval upon returning to the interrupted task, leading participants to instead *reconstruct* (Gray & Fu, 2004) the task context by gathering information directly from the task interface (which could not be done in the Sasangohar et al. study). If participants indeed preferred problem state reconstruction over reactivation, the resumption lag would not differ based on nesting level.

Contrary to our expectation, processing code similarity resulted in shorter, rather than longer, resumption lag, but only in the case of single interruptions. This result differs from previous findings indicating that the resumption of interrupted tasks suffers when interrupting tasks have the same processing code (see Brudzinski et al., 2007; Ratwani et al., 2007, 2008; Ratwani & Trafton, 2004, 2008; Shen & Jiang, 2006). It again suggests that participants preferred reconstruction of task state information, rather than memory-based recall. Because reconstruction requires returning to and gathering information from relevant parts of the task interface, participants were likely able to do so more quickly and easily when returning to the interrupted task from a verbal interruption rather than a spatial one—while the latter required participants to compare shapes and relative locations of elements like routes and turbulence zones (which are spatial processes that would interfere with spatial memory; see also Brüning et al., 2020), the former did not. We suspect that this effect was limited to single interruptions due to the shorter intervals between switching to and returning from the single interrupting tasks, which at least allowed for the possibility that participants would remember where they left off, compared to serial and nested interruptions that were performed as pairs.

An alternate explanation for the lack of benefit of processing code dissimilarity may be that the different processing code condition did not include trials where a pair of two spatial tasks interrupted the ongoing verbal task. Instead, the conditions that were tested alternated processing code similarity based on the most recently interrupted task (e.g., a verbal OT—spatial IT—verbal IT task set). Thus, interference from the tertiary task likely overshadowed potential performance benefits in resumption lag that could be achieved with a spatial-only interrupting task set. While no differences in resumption lag were observed, the perceived workload (as measured by pupil diameter) when returning to the interrupted flight request task was significantly higher in the same processing code condition. This increased (perceived) workload may have been countered by increased effort, resulting in no differences in resumption lag. It seems surprising that in the different processing code condition, the pupil diameter was higher for serial interruptions, compared to single and nested. However, this finding is likely due to unintended differences in the illumination of interface elements when returning from a verbal task (only applicable to serial conditions), which uses darker colors, as opposed to when returning from a spatial task (applicable to nested and, sometimes, single conditions), which uses brighter colors to display the map interface. Therefore, the pupil diameter, which can be sensitive to ambient light conditions, was probably larger when returning from serial interruptions in order to adjust to the dark verbal task interface (Recarte et al., 2008).

Lastly, we observed significant improvements in accuracy on the ongoing flight request task due to learning from experience between the first and second scenario. During debrief interviews, participants indicated that they improved in their ability to handle and prioritize interruptions. For example, one participant recalled that "…the second session I did better because I was learning to prioritize the tasks." One possible explanation for this learning effect could be a lack of training on task. However, given that flight request task accuracy was above 90% in both scenarios for single and serial interruptions but not for nested, this is an unlikely reason. What is more likely is that as participants' experience with the flight request task increased, they were able to retrieve more quickly from declarative memory the associated rules and procedures for handling interrupting tasks (e.g., mapping of cargo name to urgency level, selecting action, etc.). This reduced interference could have contributed, at least in part, to an improvement in the participants' ability to interleave and better recover from nested interruptions. In fact, the creation of strong memory associations in primary task elements, using concepts like associative cueing, has been proposed as a countermeasure to the disruptive effects of interruptions (Li et al., 2012).

In summary, this experiment shows that 1) frequent and nested interruption notifications are significantly less likely and slower to be acknowledged, compared to both single and serial notifications, 2) interpretation accuracy is worse for single and nested interruptions, compared to serial interruptions, and 3) accuracy on the ongoing task is worse during high frequency phases, and for single and nested interruptions, compared to serial. One limitation of this study was that the experiment setup did not allow for a clear separation of performance at the interpretation and integration stages. Specifically, in regard to the second finding that interpretation accuracy was lower in the case of single and nested interruptions, it remains unclear whether participants failed to reliably switch to incoming tasks of higher priority due to incorrect classification of task urgency during interpretation, or due to difficulty with timely switching to incoming tasks during integration. We investigate this issue next in **Chapter 3**.

Chapter 3 – Identifying Breakdowns in the Interpretation and Scheduling of Interruptions

3.1 Introduction

In **Chapter 2,** we evaluated how interruption frequency, nesting level, urgency, and processing code similarity affected operator performance at the detection, interpretation, and integration stages of interruption management. At the detection stage, frequent and nested interruption notifications were less likely to be acknowledged, compared to less frequent and non-nested notifications. Still, the overall rate of acknowledgement was above 95% and is therefore not a significant concern.

Participants struggled more with appropriate task switching. Specifically, participants switched to high-urgency single and nested interruptions in only 84% and 77% of the cases, respectively, compared to 94% in the case of serial interruptions. Because participants were not asked to indicate explicitly their assessment of the urgency level for each incoming notification, it is not clear whether these findings are the result of difficulties at the interpretation or integration stage. In other words, it is not clear whether participants misinterpreted task urgency, failed to determine the appropriate action based on relative urgency, or struggled to switch to interrupting tasks in a timely fashion (before the task expired).

To answer this question, the task interface and procedures for the second experiment were modified in two ways. First, we asked participants in this study to explicitly state their evaluation of the absolute urgency level of the ongoing and the incoming task, as well as the appropriate action to take based on the relative urgency of the ongoing and incoming tasks (i.e., accept, delay, or handoff). As shown in **Figure 3.1**, this was accomplished by adding three sets of checkboxes that a) allowed participants to mark the urgency of ongoing and incoming tasks as low (L), medium, or high (H), and b) indicate the appropriate action as accept as soon as possible (ASAP), delay, or handoff to another agent. An alternative approach for measuring interpretation accuracy may be to instruct participants to verbally state the urgency level and action for each incoming task notification, similar to the think-aloud protocol (Fonteyn et al., 1993). This method was not adopted due to its potentially disruptive nature (e.g., van den Haak et al., 2003), and due to the difficulty of recording verbal responses to frequently-presented interruptions.

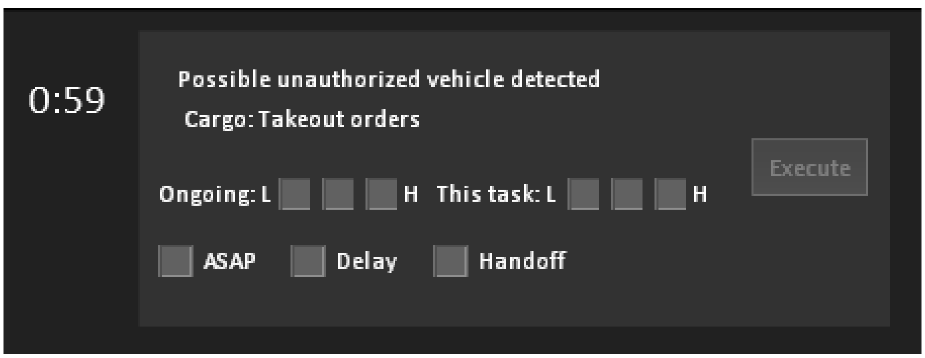

Figure 3.1 Checkboxes in task notification to indicate urgency level and action.

The second change concerned the expiration of tasks. Task notifications in the present study did not expire after a fixed period of time to allow us to determine whether participants were aware of urgency levels and the correct action to take but simply switched late. Pending tasks stayed in the notifications panel until responded to by the participant.

These two changes allowed us to assess how frequent and nested interruptions affect the participants' ability to: 1) correctly classify the urgency of incoming interruption notifications, 2) activate the appropriate rule associated with the level of relative urgency of the notification, and 3) switch between ongoing and pending tasks accurately and in a timely manner. Regarding aim (1), we expected that more frequent and nested interruptions would lead to greater working-memory interference and, in turn, less accurate mapping of cargo names to level of urgency. At this step, there is a possibility that participants incorrectly recall the urgency level of the ongoing task and/or incorrectly remember the rules for mapping cargo names to urgency level. When activating the appropriate rule (aim 2), participants may compare the urgency levels of the ongoing and incoming tasks incorrectly and/or map the relative level of urgency (i.e., higher, equal, or lower) to an incorrect task action. Lastly, regarding aim (3), more frequent and nested interruptions increase the difficulty of integrating incoming tasks into ongoing workflow. Multiple pending notifications require multiple comparisons of relative urgency, as well as a comparison of the time at which the notifications arrived. The operator may delay an interrupting task momentarily (until reaching a breakpoint in the ongoing task, for example) but then forget to switch to it in time. Or they might mistakenly switch to a more recent or less urgent pending task, while an older task of equal or higher urgency is postponed for longer. In line with these aims, we expected the following relationships at the interpretation and integration stages:

H1: *Participants will be less accurate and take longer to interpret urgency of ongoing and incoming tasks during high frequency periods and in the case of nested interruptions, compared to low frequency periods and single/serial interruptions, respectively*.

H2: *Participants will be less accurate at activating the correct rule during high frequency periods and nested conditions, compared to low frequency and single/serial interruptions, respectively.*

H3: *Participants will take longer to switch to pending tasks of higher urgency during high frequency periods, and when interruptions are nested, compared to low frequency and single/serial trials, respectively.*

3.2 Method

3.2.1 Participants

41 students (23 male, 18 female) from the University of Michigan participated in the study. Ages ranged between 18 and 28 years (M = 22, SD = 2.7). Participant eligibility was limited to ages 18-30 years, which is comparable to the age range of UAV operators. Participants were paid $17/hour for completing one three-and-a-half-hour experiment session, and a $25 performance bonus was awarded to participants with performance scores in the top quartile. This research was conducted in compliance with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at the University of Michigan (IRB # HUM00221232). Informed consent was obtained from each participant before the experiment.

3.2.2 Experiment Design

The study employed a 2 (interruption frequency: low, high) x 3 (nesting level: single, serial, nested) x 3 (interruption urgency: low, medium, high) within-subjects, fractional factorial design. The urgency level varied between low-medium-high for the flight request task and single interruptions, and between medium-high for both serial and nested interruptions. Low urgency interruptions were not presented in the serial and nested conditions to incentivize participants to switch from the medium urgency new flight task to an incoming medium or high urgency interrupting task. Similar to the study detailed in **Chapter 2**, participants completed two scenarios, lasting approximately 30 minutes each. Each scenario included 20 flight request tasks and 32 interruptions (see **Figure 2.8**). Eight of the 20 flight request tasks were uninterrupted. Participants were not informed about how many and which tasks would be interrupted.

3.2.3 Experiment Apparatus and Tasks

Participants were tasked with supervising a set of autonomous drones delivering cargo to commercial locations in the Houston, Texas metropolitan area. Participants performed the same set of five tasks as those detailed in **Chapter 2**. These included approving flight requests, selecting alternate landing sites, diverting to alternate vertiports, detecting unauthorized aircraft, and responding to requests for vehicle status information. Two minor changes were made to the flight request task to encourage more effortful processing of task-related information, rather than completing the task as a fixed, memorized procedure.

Participants in this experiment were required to always load the extra battery module(s) before loading the cargo item, and to check off items in the completion checklist if they were applicable to the task (i.e., check off ‘added extra battery module’ only if it was needed for the flight). The experiment was conducted using a multi-UAV simulator developed by the THInC lab based on the Air Force Vigilant Spirit ground control station (Feitshans et al., 2008). It consists of two 24” monitors running at a resolution of 1920 x 1200 each. The left-hand monitor displayed the task area and a notification area (see **Figure 3.2**). Only one task was visible at any given time in the task area. Participants switched between ongoing and interrupting tasks using a tab interface, located at the top of the task area on the left monitor. The right-hand monitor displayed vehicle health and mission progress information for each of the 20 simulated aircraft.

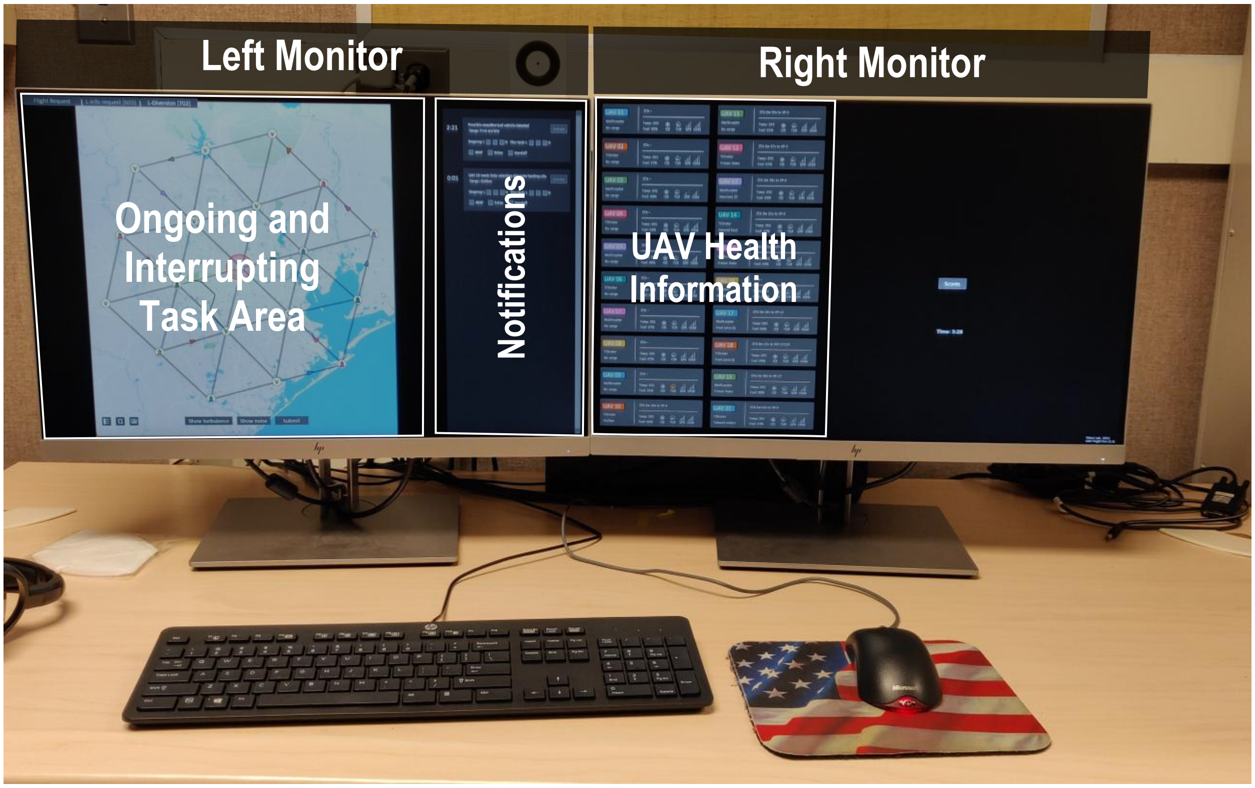

Figure 3.2 Experiment setup showing the multi-UAV simulator running on two monitors. The left-monitor displays the task area and notification area. The right-monitor displays the UAV health information.

3.2.4 Procedure

Participants were asked to attend one three-and-a-half-hour session that involved training, three practice scenarios, and two experiment scenarios. Before beginning the training, participants completed a questionnaire on gaming experience and multitasking ability (**Appendix D**). After the pre-experiment survey, participants completed an approximately one-hour long PowerPoint-based training session on managing UAV fleet operations for transporting cargo. After learning about each of the five tasks (one ongoing, four interrupting) in the training slides, participants were guided by the experimenter to practice the task in the context of the UAV-simulator. Participants were told to prioritize tasks based on the relative urgency between the ongoing and incoming tasks. They were asked to 1) always hand off lower urgency tasks, 2) always delay incoming tasks with the same urgency level as the ongoing task, and 3) always accept tasks of higher urgency as soon as possible. Participants were informed that tasks will be scored based on the level of urgency and completion time (see **Table 3.1**). High urgency interrupting tasks, for example, were worth six points if completed within 1 minute of being notified, three points if completed within 2 minutes, and *negative* two points after two minutes since notification onset. There were two exceptions to the scoring rules. First, tasks that were correctly handed off were worth half points for medium urgency, and 0 points in the case of low urgency. Second, the flight request task was worth partial points (3 for high, 2 for medium, and 0 for low urgency) if the flight was delayed, which always occurred after two minutes, regardless of the level of urgency.

Table 3.1 Scores awarded for high, medium, and low urgency interrupting tasks based on time to completion after being notified with the auditory chime.

Task urgency	**Max score**	**Partial Score**	**Penalty**
High	6 (< 1 min)	3 (< 2 min)	-2 (> 2 min)
Medium	4 (< 2 min)	2 (< 4 min)	-1 (> 4 min)
Low	2 (< 2 min)	0 (> 2 min)	No penalty

After completing the training module and the first practice scenario, participants completed a short questionnaire on their understanding of task management and scoring (**Appendix E**). This questionnaire included questions to verify that participants were able to correctly map cargo name to urgency level, understood which action to select for incoming tasks of higher, equal, or lower urgency, and understood that tasks were scored based on urgency level and completion time. The experimenter asked participants to explain their response when they answered incorrectly, and when they were unsure of the correct response. A perfect score was required before proceeding to the second practice scenario.

In the second practice scenario, participants again completed each of the five tasks three times, separately, and without any interruptions. After a five-minute break, participants completed a third practice scenario with interruptions. Following another five-minute break, each participant completed two experiment scenarios (counterbalanced), lasting approximately 30 minutes each, with a 10-minute break in-between. The scenario layout, moment of interruption presentation, and interruption timing were the same as reported in **Chapter 2**. The third practice scenario and the two experiment scenarios were completed while wearing the Tobii Pro Glasses 2 eye tracking glasses.

After each experiment scenario, participants completed a NASA Task Load Index survey to assess perceived workload (**Appendix B**). Finally, at the end of the experiment, participants completed an open-ended questionnaire to share their strategies and challenges faced when handling interruptions (**Appendix F**).

3.2.5 Dependent Measures

The dependent measures included task performance, eye tracking metrics, and survey responses. Performance metrics comprised notification acknowledgement rate, notification acknowledgement time, interpretation accuracy, interpretation time, task accuracy, time on task, and resumption lag. Survey data included questions about the participants' self-reported interruption management self-efficacy (IMSE) and multi-tasking computer self-efficacy (MTCSE; borrowed from Basoglu et al., 2009), frequency of gaming, perceived workload (NASA-TLX), and challenges experienced with handling interruptions. Eye tracking data was collected using the Tobii Pro Glasses 2 eye tracker. Eye tracking metrics included number of fixations, fixation duration, visit duration, and pupil diameter. Pupil diameter was used as a measure of perceived mental workload (Cain, 2007; Longo et al., 2022; Recarte et al., 2008).

3.3 Results

Analyses were performed using generalized linear mixed effects models with fixed and random effects. The models were developed and tested in the R programming language using the *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017) packages. Participant ID and scenario order were used as random effects if they improved the model fit. Main effects were evaluated using Chi-squared tests between a null model and another containing the variable of interest as a fixed effect. All significance levels in figures and tables are reported as *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$.

3.3.1 Notification Acknowledgement

Notification acknowledgement rate and time were evaluated using a generalized linear mixed effects model. Acknowledgement rate did not differ between the low (M = 98.6%, SE = 0.10) and high (M = 97.8%, SE = 0.10) frequency interruptions ($\chi^2(1) = 2.64$, $p = 0.104$). Nested interruptions (M = 95.0%, SE = 2.10) had a lower rate of acknowledgement, compared to both single (M = 98.5%, SE = 0.44; *Odds Ratio* = 1.43, *95% CI* [0.75, 2.11], $z = 4.11$, p = < 0.001) and serial interruptions (M = 99.4%, SE = 0.61; *Odds Ratio* = 2.36, *95% CI* [0.83, 3.89], $z = 3.03$, $p = 0.002$; see **Figure 3.3**). There was no significant difference in the rate of acknowledgement between serial and single interruptions (*Odds Ratio* = 0.39, $z = -1.24$, $p = 0.428$).

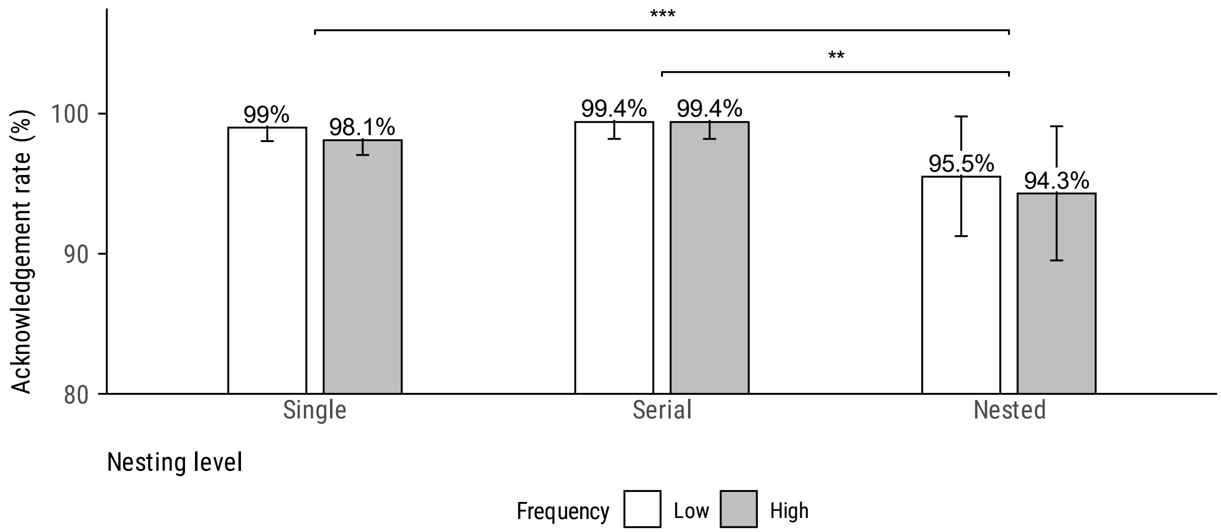

Figure 3.3 Acknowledgement rate shown as a function of nesting level and interruption frequency. Error bars show 95% CI.

Notification acknowledgement time was not different between the low (M = 2.87s, SE = 0.44) and high (M = 2.66s, SE = 0.44) interruption frequency phases (χ^2 (1) = 0.03, p = 0.874). Compared to single interruptions (M = 2.51s, SE = 0.42), participants were, on average, 1.2 seconds (48%) *faster* when acknowledging serial interruptions (M = 1.31s, SE = 0.46; β = -1.2, *95% CI* [-1.64, -0.77], t(2380) = -5.4, p = < 0.001), and 59% *slower* when acknowledging nested interruptions (M = 3.99s, SE = 0.47; β = 1.48, *95% CI* [-1.64, -0.77], t(2384) = 6.45, p = < 0.001; see **Figure 3.4**).

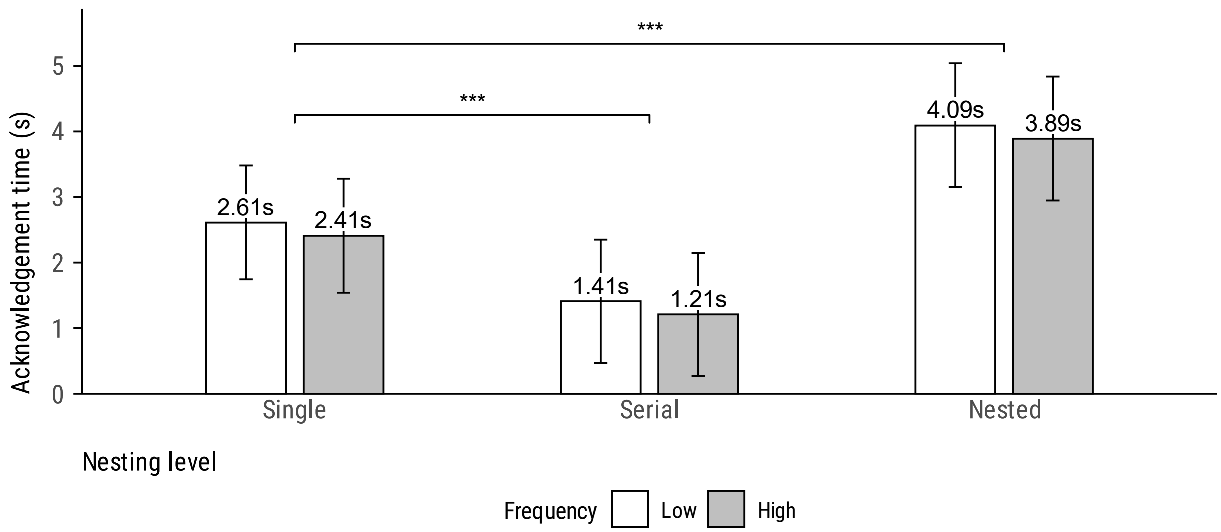

Figure 3.4 Acknowledgement time shown as a function of nesting level and interruption frequency. Error bars show 95% CI.

3.3.2 Notification Interpretation

Interpretation accuracy was evaluated separately for urgency of, and chosen action in, response to an interruption. The latter was calculated only for cases where participants correctly identified the urgency of both the incoming and ongoing task. Some participants misclassified the urgency level of cargo items in more than 50% of the cases, indicating a lack of understanding of task instructions. Those cargo items for which participants reached an accuracy level no higher than chance—108 trials, approximately 4% of data—were excluded from the analysis.

We expected that participants would be less accurate at interpreting the urgency and selecting the appropriate action during high frequency periods and in the case of nested interruptions, compared to low frequency periods and single/serial interruptions, respectively. For accuracy of urgency, our expectation was partially confirmed—there was a main effect of interruption frequency (χ^2 (1) = 5.02, p = 0.025), but not nesting level (χ^2 (2) = 3.05, p = 0.218). Accuracy of urgency was lower during high frequency periods (M = 91.3%, SE = 1.22; β = -

2.31, *95% CI* [-4.30, -0.31], *t*(2324) = -2.27, *p* = 0.023), compared to low frequency (*M* = 93.6%, *SE* = 1.22). Accuracy of urgency did not differ significantly between the single (*M* = 92.0%, *SE* = 1.15), serial (*M* = 92.8%, *SE* = 1.72; β = 0.83, *95% CI* [-2.17, 3.83], *t*(2317) = 1.87, *p* = 0.588), and nested conditions (*M* = 94.7%, *SE* = 1.75; β = 2.69, *95% CI* [-0.37, 5.74], *t*(2321) = 1.73, *p* = 0.085). For accuracy of chosen action, accuracy was very high (*M* = 96.6%, *SE* = 0.38), and there was no main effect of interruption frequency (χ^2 (1) = 2.82, *p* = 0.093) or nesting level (χ^2 (2) = 3.83, *p* = 0.147).

We reviewed more closely those cases in which accuracy of urgency and accuracy of action selection were less than perfect. In 80% (116) of cases with an incorrect classification of urgency, participants overestimated low urgency items as medium urgency, and underestimated medium urgency items as low urgency (see **Figure 3.5**). There was a similar pattern among the trials in which an incorrect action was selected despite accurate urgency classification—participants chose to incorrectly delay tasks, rather than handoff or accept, 68% (51 cases) of the time (**Figure 3.6**).

Interpretation time was defined as the interval between the first and last checkbox completed for each notification in the notifications panel. We expected that participants would take longer to interpret interruption notifications during higher frequency and nested interruptions. The first expectation was partially supported—participants took longer to interpret interruption notifications during high frequency periods (*M* = 2.61s, *SE* = 0.22; β = 0.59, *95% CI* [0.17, 1.52], *t*(2350) = 2.72, *p* = 0.007), compared to low frequency periods (*M* = 2.02s, *SE* = 0.22). However, there was no main effect of nesting level on interpretation time (χ^2(2) = 2.438, *p* = 0.296).

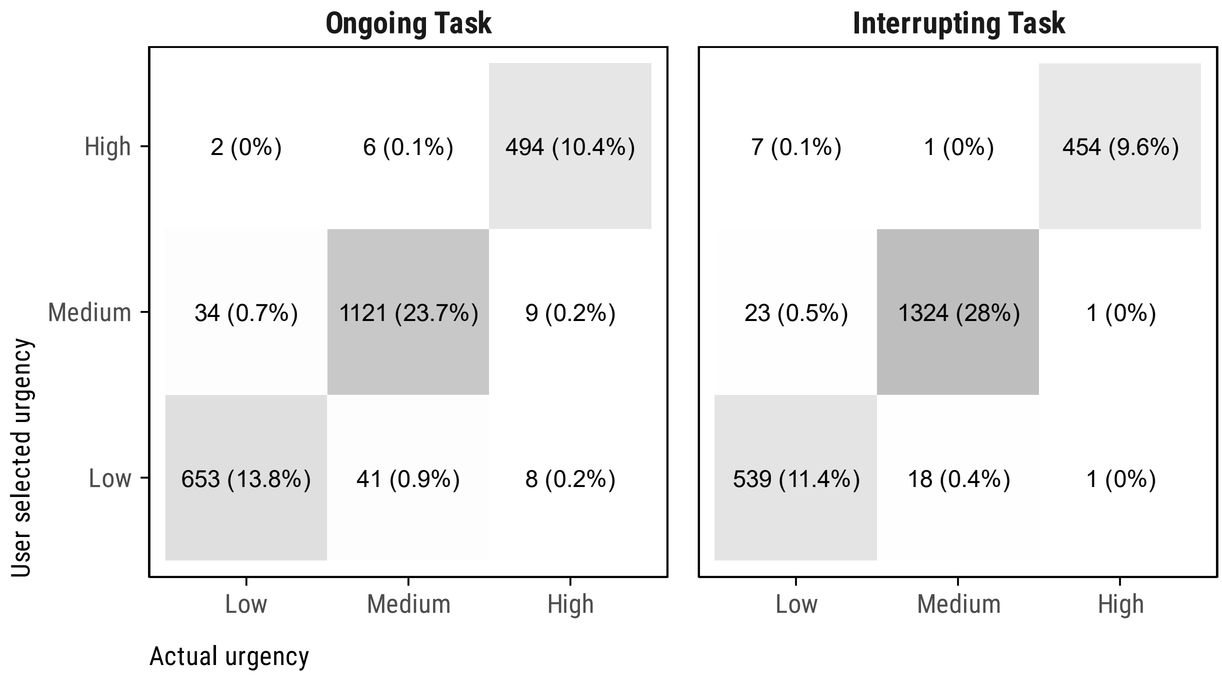

Figure 3.5 Confusion matrices showing the distribution of user selected urgency and the actual urgency of ongoing and interrupting tasks.

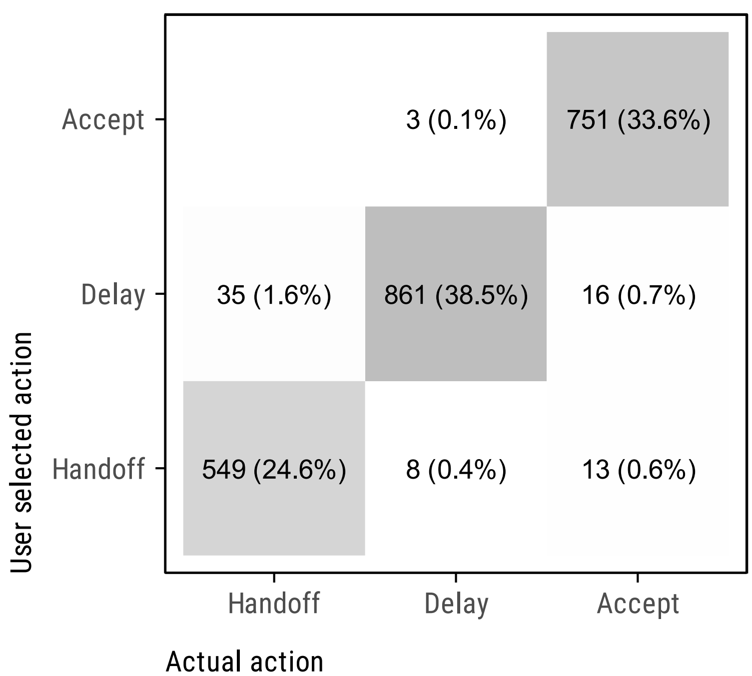

Figure 3.6 Confusion matrix showing the distribution of user selected action and the action that should have been taken based on the relative levels of urgency for the ongoing and interrupting tasks. Excludes cases in which ongoing or interrupting task urgency was indicated incorrectly.

3.3.3 Interruption Lag

Interruption lag was defined as the time taken, after notification onset, to switch to an incoming task of higher priority. A linear mixed effects model was fitted with interruption frequency and nesting level as fixed effects with interaction (see **Table 3.2**). There was a main effect of interruption frequency ($\chi^2(1) = 4.47$, $p = 0.035$) and nesting level ($\chi^2(2) = 19.39$, $p < 0.001$). There was also a significant interaction between interruption frequency and nesting level ($\chi^2(2) = 16.06$, $p < 0.001$). Interruption lag was longer during high frequency periods (*M* = 17.00s, *SE* = 1.87; $\beta = 4.41$, *95% CI* [0.53, 8.72], $t(371) = 2.22$, $p = 0.027$), compared to low frequency (*M* = 12.38s , *SE* = 1.75). Interruption lag was longer also for nested interruptions (*M* = 21.05s, *SE* = 2.07), compared to both single (*M* = 13.75s, *SE* = 2.19; $\beta = -7.29$, *95% CI* [-13.47, -1.12], $t(374) = -2.78$, $p = 0.016$) and serial (*M* = 9.26s, *SE* = 2.00; $\beta = -11.79$, *95% CI* [-17.58, -5.99], $t(370) = -4.79$, $p < 0.001$) interruptions (see **Figure 3.7**). Post-hoc tests using the Tukey method revealed that for nested interruptions only, interruption lag was longer during the high frequency period (*M* = 29.18s, *SE* = 2.82; $\beta = 16.27$, *95% CI* [6.105, 26.49], $t(372) = 4.560$, $p < 0.001$), compared to the low frequency period (*M* = 12.91s, *SE* = 2.64).

Table 3.2 Summary of model fitted for interruption lag, with frequency and nesting level included as fixed effects with interaction.

	Interruption Lag				
Predictors	*Estimates*	*SE*	*CI*	*Z*	*p*
(Intercept)	15.49	2.63	10.33 – 20.65	5.90	**<0.001**
Frequency – High	-3.47	3.85	-11.03 – 4.09	-0.90	0.367
Nesting Level – Serial	-6.76	3.42	-13.47 – 0.05	-1.98	**0.048**
Nesting Level – Nested	-2.58	3.42	-9.30 – 4.14	-0.76	0.451
Frequency [High] × Nesting Level [Serial]	4.54	5.13	-5.55 – 14.63	0.88	0.377
Frequency [High] × Nesting Level [Nested]	19.75	5.25	9.43 – 30.07	3.76	**<0.001**

Observations	412
Marginal R^2 / Conditional R^2	0.086 / 0.173

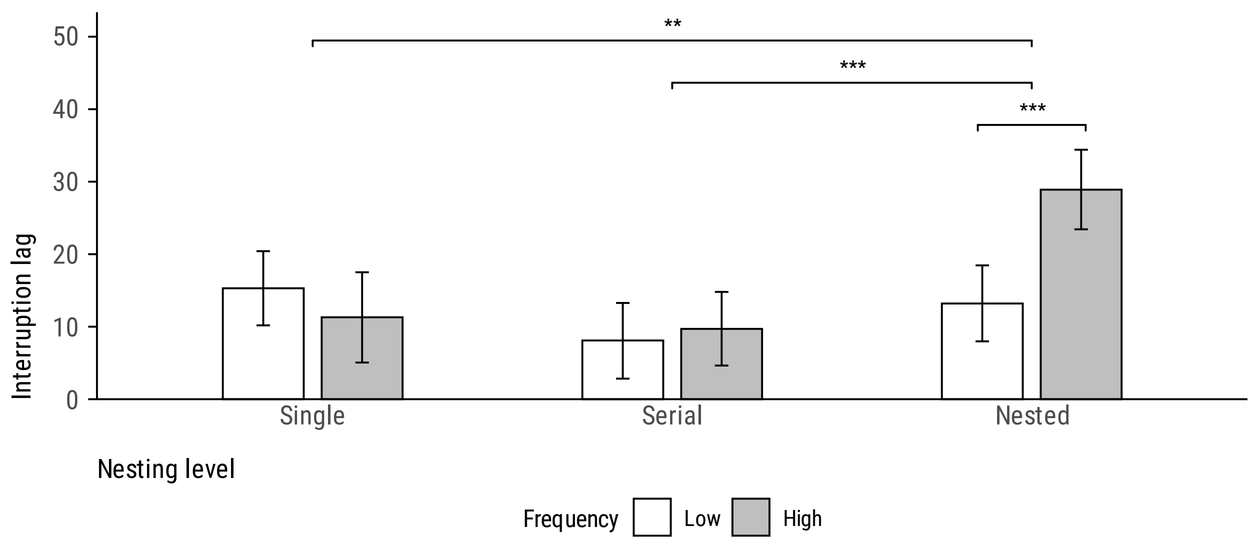

Figure 3.7. Interruption lag as a function of interruption frequency (left) and nesting level (right). Error bars show 95% CI.

A closer inspection of eye tracking data showed that time to first fixation in the notifications panel (**Figure 3.8**, left), after notification onset, was longer for nested interruptions (M = 9.14s, SE = 0.84), compared to both single (M = 4.76s, SE = 0.67; β = -4.38, *95% CI* [-6.11, -2.64], $t(778) = -5.93$, $p < 0.001$) and serial (M = 2.38, SE = 0.82; β = -6.76, *95% CI* [-8.81, -4.7], $t(776) = -7.72$, p = < 0.001). This was again particularly pronounced for high frequency nested interruptions (M = 12.12s, SE = 1.06), compared to low frequency (M = 6.15s, SE = 1.05; β = - 5.97, *95% CI* [-9.59, -2.35], $t(775) = -4.71$, $p < 0.001$). The number of visits also was higher, for nested interruptions only, in the high frequency period (M = 1.80, SE = 0.10; β = 0.39, *95% CI* [0.09, 0.70], $t(181) = 3.692$, $p = 0.004$), compared to low frequency (M = 1.50, SE = 0.09).

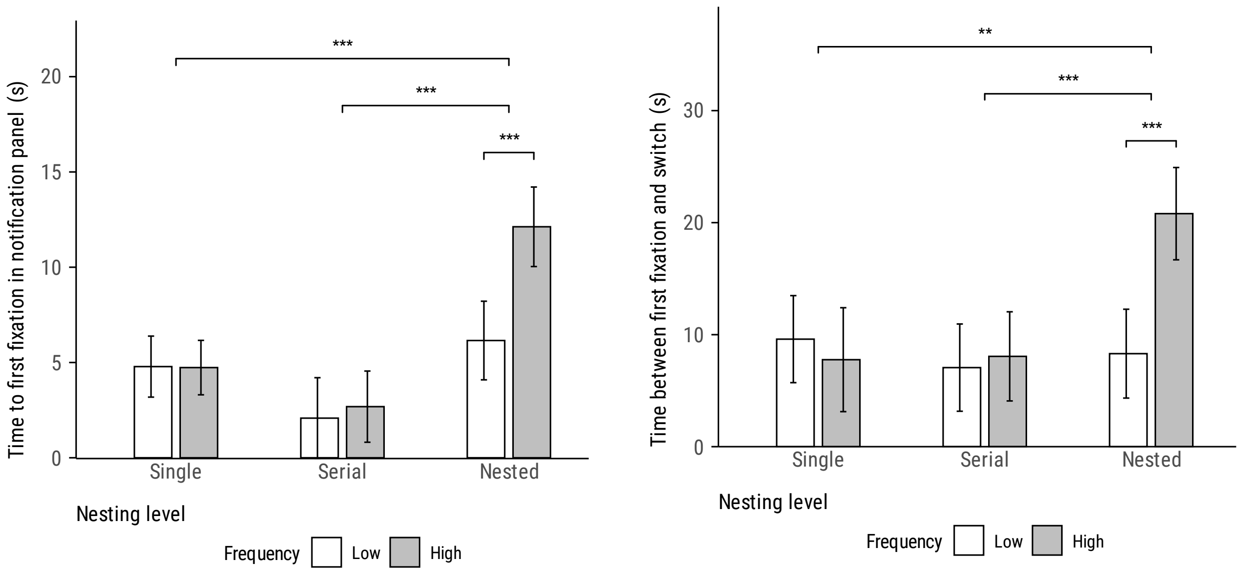

Figure 3.8. Time to first fixation in the notifications panel after notification onset(left) and time between first fixation and switch to incoming task of higher priority (right) shown as a function of nesting level. Error bars show 95% CI.

Time between the first fixation and the switch to the incoming task was also compared across nesting level (**Figure 3.8**, right). Even after viewing the notification panel, participants took longer to switch to the incoming task for nested interruptions (*M* = 14.55s, *SE* = 1.58), compared to single (*M* = 8.68s, *SE* = 1.65; *β* = -5.87, *95% CI* [-10.42, -1.32], *t*(346) = -3.03, *p* = 0.007) and serial (*M* = 7.56, *SE* = 1.54; *β* = -6.99, *95% CI* [-11.3, -2.68], *t*(341) = -3.82, *p* < 0.001) interruptions.

3.3.4 Task Accuracy

There was a main effect of interruption frequency ($\chi^2(1) = 6.82$, *p* = 0.009) and nesting level ($\chi^2(1) = 18.59$, *p* < 0.001) on interrupted flight request task accuracy. Post-hoc tests revealed that, only for single interruptions, accuracy on the flight request task was significantly worse during high frequency periods (*M* = 90.5%, *SE* = 1.21; *β* = -5.36, *95% CI* [0.66, 10.06], *t*(236) = 3.28, *p* = 0.015; see **Figure 3.9**), compared to low frequency (*M* = 95.9%, *SE* = 1.18). Flight request task accuracy was worse also for nested interruptions (*M* = 84.2%, *SE* = 2.38; *β* =

-8.98, 95% CI [-14.9, -3.07], $t(245) = 3.58$, $p = 0.001$), compared to single ($M = 93.2\%$, $SE = 0.87$) interruptions. There was no significant difference in performance on the flight request task between single and serial interruptions ($M = 90.6\%$, $SE = 2.28$; $\beta = -2.54$, 95% CI [-8.23, 3.15], $t(249) = 1.05$, $p = 0.545$).

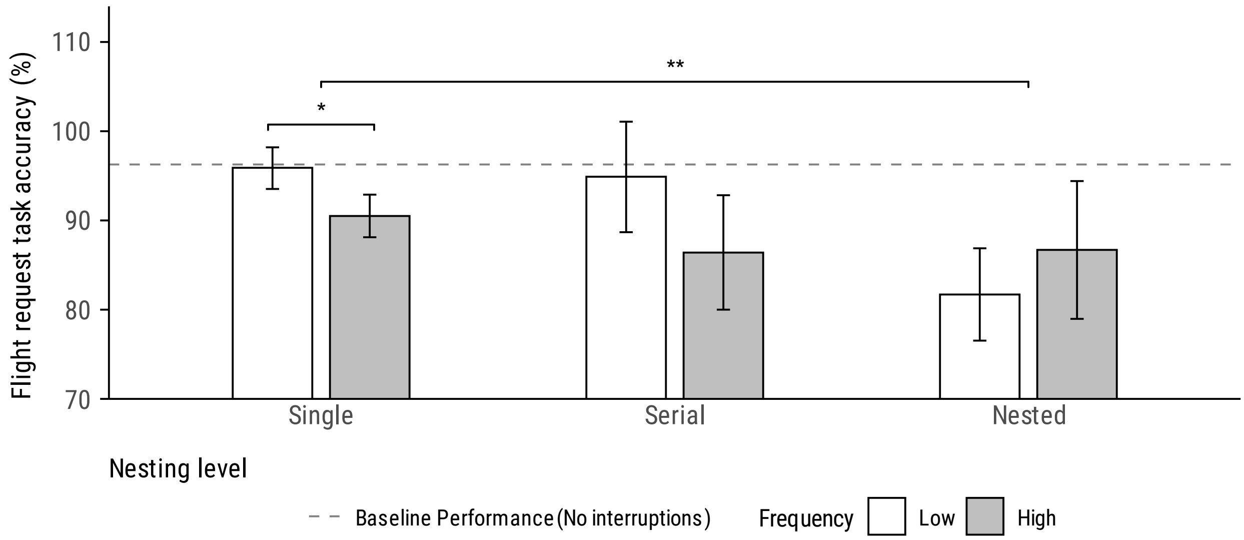

Figure 3.9. Flight request task accuracy as a function of interruption frequency and nesting level. Error bars show 95% CI.

For accuracy on interrupting tasks, there were no significant main effects of interruption frequency ($M_{low} = 92.2\%$, $SE_{low} = 0.77$, $M_{high} = 91.9\%$, $SE_{high} = 0.77$; $\chi^2(1) = 0.12$, $p = 0.710$) or nesting level ($M_{single} = 93.0\%$, $SE_{single} = 1.49$, $M_{serial} = 90.2\%$, $SE_{serial} = 2.83$, $M_{nested} = 95.6\%$, $SE_{nested} = 2.95$; $\chi^2(2) = 1.85$, $p = 0.397$).

3.3.5 Task Resumption

There was no main effect of interruption frequency on resumption lag for the primary flight request task ($\chi^2(1) = 1.27$, $p = 0.260$). Nesting level had a marginal effect on resumption lag ($\chi^2(2) = 6.32$, $p = 0.042$). Post-hoc tests showed that only during low frequency periods (**Figure 3.10)**, participants resumed the flight request task more slowly when returning from single interruptions ($M = 8.62$s, $SE = 0.57$; $\beta = 3.74$, 95% CI [0.5, 6.98], $t(215) = 3.32$, $p =$

0.013), compared to nested interruptions (*M* = 4.88s, *SE* = 1.09) but not serial (*M* = 6.14, *SE* = 1.29; β = 2.48, *95% CI* [-1.37, 6.32], *t*(224) = 1.85, *p* = 0.435). Differences in resumption lag were not significantly different during the high frequency period between single, serial, and nested trials.

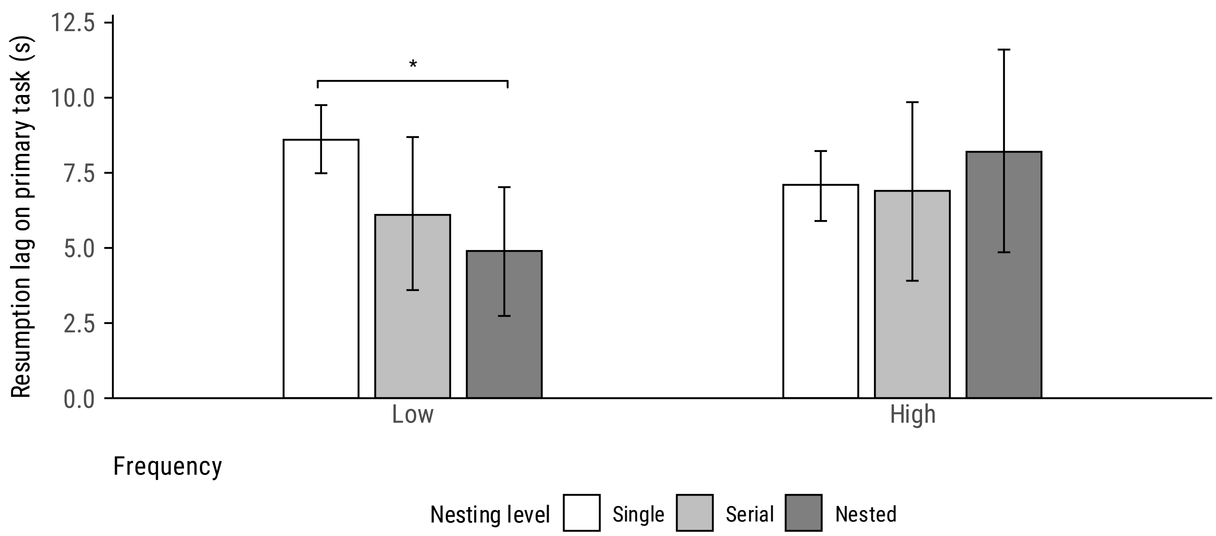

Figure 3.10. Resumption lag on the primary flight request task, shown as a function of interruption frequency and nesting level. Error bars show 95% CI.

3.3.6 Survey and Debrief Findings

Participants' overall task performance was compared for different levels of gaming experience and multitasking ability. Participants who reported playing games more frequently did not perform better than those who reported playing games rarely. Similarly, the IMSE and MTCSE scores showed no significant relation to task performance. NASA-TLX survey ratings were compared between the first and second scenarios. Only ratings for the difficulty dimension differed significantly (see **Figure 3.11**). Participants reported that they had to work less hard to accomplish their level of performance in the second scenario (*M* = 7.23, *SE* = 0.20, *95% CI* [6.84, 7.61]; *t*(38) = -2.449, *p* = 0.019), compared to the first scenario (*M* = 7.73, *SE* = 0.20, *95% CI* [7.33, 8.12]).

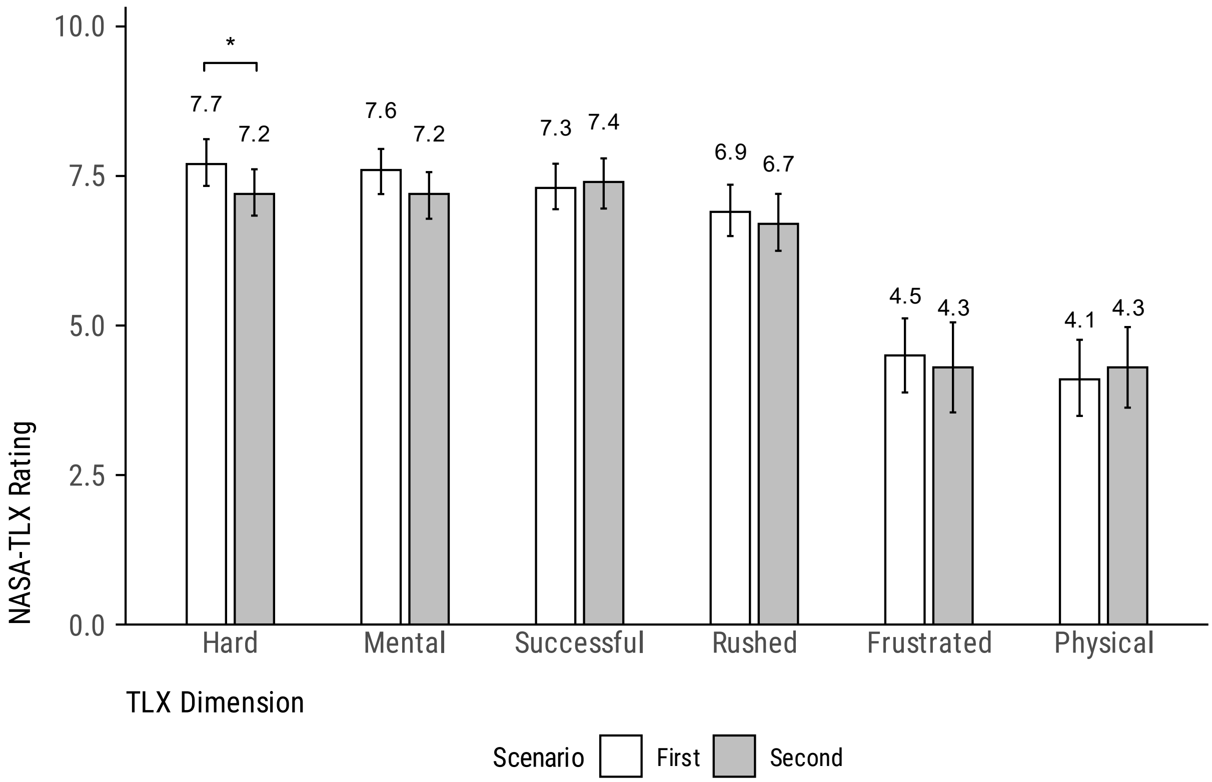

Figure 3.11. NASA Task Load Index survey ratings for the first and second scenarios. Error bars show 95% CI.

Table 3.3 Summary of study expectations and results.

Expectation	Results
H1: Participants will be less accurate and take longer to interpret urgency of ongoing and incoming tasks during high frequency periods and in the case of nested interruptions, compared to low frequency periods and single/serial interruptions, respectively.	Interpretation time and accuracy of urgency were worse during high interruption frequency phases but did not differ as a function of nesting level.
H2: Participants will be less accurate at activating the correct rule during high frequency periods and nested conditions, compared to low frequency and single/serial interruptions, respectively.	Accuracy of action selection was affected neither by interruption frequency, nor by nesting level.
H3: Participants will take longer to switch to pending tasks of higher urgency during high frequency periods, and when interruptions are nested, compared to low frequency and single/serial trials, respectively.	There was an interaction effect between interruption frequency and nesting level. Participants took significantly longer to switch to nested interruptions during high frequency periods, compared to both single and serial interruptions.

3.4 Discussion

This study analyzed whether observed difficulties with task switching in the first experiment resulted from breakdowns at the interpretation and/or integration stage of interruption management. Specifically, it assessed how frequent and nested interruptions affect people's ability to a) correctly classify the urgency of incoming interruption notifications, b) activate the appropriate rule associated with the relative urgency levels of current and incoming tasks, and c) switch between tasks in a timely manner.

We expected participants to be less accurate and slower when determining the urgency level of ongoing and incoming tasks during high frequency periods, and for nested interruptions. This expectation is partially supported by the data. Accuracy of urgency dropped during high interruption frequency periods, compared to low frequency, but did not differ with respect to nesting level. Interpretation time also was longer during high interruption frequency periods, by approximately 350ms, but did not differ across nesting levels. It appears that frequent interruptions, which are assumed to result in higher working memory interference from current and residual goals, disrupted both the probability and latency of recall from declarative memory. This can be explained by the adaptive control of thought–rational (ACT-R) computation model (Anderson et al., 2004), which forms the basis for many of the assumptions on working memory performance made by the MFG model. According to ACT-R, activation of an appropriate chunk in memory is less than perfect and depends on factors such as the strength of association between a concept and an associated fact, the recency of activation, and relevance to the current task. Activation is also affected by noise in working memory. Distractions, such as the intent to process an interruption signal is one source of noise. Other sources include intentions to integrate pending tasks, and residual thoughts related to recently completed tasks. Similar to how a visual cue is less likely to be noticed when surrounded by many non-target stimuli, successful activation of the appropriate chunks in working memory is both less probable and takes longer to reach a sufficiently high level of activation when the interference from noise is high, such as during high frequency interruptions.

In contrast to interruption frequency, nested interruptions did not incur a significant cost to the accuracy or latency of recall from declarative memory. This is surprising, given that we expect nested interruptions to increase the load on working memory—participants need to additionally remember to return to the secondary task after responding to the nested notification for the tertiary task, and remember to resume the interrupted primary task after completing the secondary task. There are two possible explanations for this finding. First, participants correctly assessed the urgency level in more than 92% of cases, suggesting a ceiling effect. Second, the current study did not require participants to remember the urgency level of the current task when presented with a nested interrupting task notification. This was done to make the nested condition comparable to single and serial trials, both of which were presented relative to the flight request task that provided access to cargo type. It is possible that if participants had been asked to completely rely on working memory to compare the urgency level of the current and incoming tasks, interpretation performance (i.e., accuracy and time) might have suffered more in the case of nested interruptions.

Participants did not seem to struggle with mapping the relative urgency level to appropriate action (i.e., accept, delay, or handoff). Accuracy of action selection was affected neither by interruption frequency, nor by nesting level. This may be explained by the strength-of-association factor included in the ACT-R model (Anderson, 1974). The model posits an inverse relationship between the strength of association between a concept and a fact, and the number of facts the concept is associated with. In the current experiment setup, the three urgency levels mapped to one of 12 cargo names. But the three possible actions—accept, delay, handoff—mapped to only three levels of relative urgency—higher, equal, or lower. Thus, the strength of association between urgency level and required action was higher, compared to recalling the level of urgency based on cargo name. As a result, accuracy of action selection was less susceptible to interference during high frequency periods.

In 68% of the 75 trials in which an incorrect action was selected, participants chose to delay, rather than handoff or accept the incoming task, even when they had correctly assessed the urgency of both tasks. Furthermore, all but one of the tasks that were delayed incorrectly involved urgency levels of ongoing and interrupting tasks that differed by only one level, such as a high urgency ongoing task interrupted by a medium urgency task rather than a high urgency ongoing task interrupted by a low urgency one. This suggests that a lower degree of separation between urgency levels may have created an increased level of uncertainty about which action to take. In these moments of uncertainty, participants likely avoided taking a decisive action and instead chose a “safer” option of dealing with the task at a later time. Another contributing factor may also be a tendency to conserve the effort (Wickens, 2014) involved in subsequently integrating the task into the ongoing workflow. By delaying a task, participants avoided completing the task immediately (accept) or not being able to complete it at a later time (handoff).

One of the main goals of this experiment was to determine how nested interruptions affect various stages of interruption management. When viewed separately, the detection and interpretation stages were not affected significantly. However, when viewed together, small performance decrements at either stage (i.e., longer notification acknowledgement time and delayed interpretation) add up and contribute to overall poorer performance. Specifically, participants take longer to switch to incoming tasks of higher urgency (i.e., longer interruption lag) for nested interruptions, compared to both single and serial interruptions. A closer look revealed that following notification onset, the number of visits in the notifications panel was higher for nested interruptions, particularly during high frequency periods, compared to single and serial interruptions. This suggests that participants tended to glance at the notification area to perform a preliminary interpretation of the urgency of the incoming notification, noted that it was of higher urgency, but stayed on the current task with the intention to respond to the notification and/or switch to the incoming task at a later time (e.g., after reaching a breakpoint or completing the current task). This reliance on prospective memory (McDaniel et al., 2004) seems to have resulted in delayed or failed (i.e., notification expiration in experiment 1) switching to high priority tasks. It follows that the difficulties observed in the previous study (detailed in **Chapter 2**) were not a result of misinterpretation of notifications, but rather problems with appropriate task integration.

Challenges associated with prospective memory have been documented in the context of interruption management by several researchers (e.g., Dismukes, 2012; Dismukes & Nowinski, 2007; Dodhia & Dismukes, 2009; McDaniel et al., 2004). Einstein et al. (2003), for example, showed that even for short 5-second intervals, the intent to perform a task in the future is ‘fragile’. Failures of prospective memory occur easily in the absence of deliberate rehearsal, and especially in the presence of other distractions, interruptions, and high task demands (McDaniel et al., 2004). Similarly, according to the MFG model, the activation level of a goal in memory rapidly decays below the interference threshold if not rehearsed. By delaying the switch away from the current task (for example, until a breakpoint is reached, or the current task is completed), it appears that participants struggled to remember to switch to the pending task on time. Interestingly, the presence of the visual notifications as a persistent reminder to switch to the incoming task was not sufficient to overcome working memory limitations. This would indicate that, in contrast to what has been suggested by previous research (McDaniel et al., 2004), a simple visual marker on the task interface is not a sufficiently salient reminder to prevent degradation of prospective memory and activate, in a timely manner, the goal of switching to pending high priority tasks. We explore this issue further in the next chapter.

One limitation of the current study was that in order to measure interpretation accuracy without relying on implicit metrics like switching behavior, we asked participants to explicitly classify the urgency level and intended action for each incoming notification. It is possible that this intervention led participants to be more deliberate in assessing task urgency, and as a result, inflated the observed levels of interpretation accuracy, compared to what might be seen in an operational setting.

Chapter 4 – Summary of Challenges and Mitigations

In the last two chapters, we have reported findings from two empirical studies that examined the effects of frequent and nested interruptions on operator performance across the three stages of detection and interpretation of notifications and the integration of interrupting tasks. In the first experiment, 40 participants completed one ongoing and four interrupting tasks in a UAV cargo delivery simulator during one 3½-hour session. Interruption frequency, nesting level, urgency, and processing code similarity were varied to examine performance breakdowns when handling single, serial, and nested interruptions. In the second experiment, 44 participants completed the same set of cargo delivery tasks as in the first experiment during a 3½-hour session, and experienced interrupting tasks varying in frequency, nesting level, and urgency. Processing code similarity was not included as a factor in this study since it did not result in any significant performance differences in Experiment 1. The main goal of the second study was to examine more closely whether breakdowns in handling frequent and nested interruptions observed in Experiment 1 occurred primarily at the interpretation or the integration stage. In both experiments, task performance measures, eye-tracking data, and subjective data from semi-structured interviews and surveys were collected.

The results from the first study indicated that frequent and nested interruption notifications were less likely to be acknowledged than less frequent and non-nested ones. Still, the overall rate of acknowledgement was high (i.e., above 95%), and participants experienced at most a 2-4% drop in the rate of acknowledgement when dealing with frequent or nested interruptions. Similarly, differences in acknowledgement time which ranged between 150-500ms were statistically significant but, from an operational perspective, delays of less than one second may not be considered a major concern. However, because missing even a single notification may jeopardize safety in high risk domains, mitigation of such risks has received substantial attention in past research (e.g., Maltz & Meyer, 2001; Rios et al., 2023; Smith et al., 2009; Wan, 2019; Woods, 1995). This line of research therefore focuses on other stages of interruption management.

Interpretation accuracy—inferred from the percentage of high urgency tasks switched to on time—suffered more significantly than detection for single (84%) and nested (77%) interruptions. However, it remained unclear whether these performance decrements resulted from problems at the interpretation or the integration stage. Experiment 2 investigated this issue.

Specifically, the second study aimed to determine whether frequent and nested interruptions affected the participants' ability to a) correctly classify the urgency of interruption notifications, b) activate the appropriate rule associated with the relative urgencies of ongoing and interrupting tasks, and c) switch between tasks in a timely manner. To this end, a set of checkboxes were added to each incoming notification window, and participants were asked to explicitly indicate the urgency of the ongoing and incoming tasks, as well as the appropriate action to take based on relative urgency.

The results from the second experiment indicated that participants did not struggle to accurately interpret the urgency level of current and incoming interruptions. Participants were also able to reliably indicate the appropriate action to take based on the relative urgency between ongoing and incoming tasks. However, participants did struggle with the appropriate scheduling of incoming tasks, and took longer to switch (i.e., longer interruption lag) to nested interrupting tasks of higher urgency, compared to both single and serial interruptions. The longer interruption lag compounds from earlier stages due to delayed acknowledgement and interpretation of notifications. Following notification interpretation, participants resisted switching away from the current task even for highly urgent interrupting tasks. A summary of results from the first two studies is provided in **Table 4.1**.

Table 4.1 Summary of challenges with managing frequent and nested interruptions.

Metric	Frequency	Nesting Level		
	High vs. low	Single	Serial	Nested
Detection				
Percent acknowledged	↓†	•	~	↓
Acknowledgement time	~	•	↑	↓
Interpretation*				
Accuracy	↓	•	~	~
Time to start interpretation	↓	•	↑	~
Interpretation time	↓	•	↑	~
Integration				
Interruption lag	↓	•	↑	↓
Accuracy on primary task	↓	•	~	~
Time delay on primary task	↓	•	↓	~
Accuracy on interrupting task	~	~secondary	•secondary •tertiary	~ secondary ↓tertiary†
Resumption lag	~	↓††	•	~

'•' symbol indicates comparison baseline, '~' indicates no significant difference relative to the baseline, '↑' indicates an improvement in the associated performance metric, and '↓' indicates a performance decrement.

* Interpretation results reported in this table are based only on data collected in Experiment 2
† Relationship holds true only for Experiment 1
†† Relationship holds true only for Experiment 2

One interesting finding across Experiments 1 and 2 was that, at the detection stage, not all participants responded the same way to an increase in interruption frequency. Although there was no overall effect of interruption frequency on acknowledgement time, one group of participants acknowledged notifications *faster* when interruptions became more frequent, while the other group acknowledged them more *slowly*. Given that participants in the slow group took approximately 300ms longer to acknowledge notifications during high frequency periods, it is unlikely that the slower acknowledgement was the result of procedural interference (which would be expected to lead to a delay close to 50ms; Salvucci & Taatgen, 2008) or saturation of attentional resources (multiple participants confirmed that hitting the pedal could be done "without much focus"). Rather, it seems that participants exercised higher-level executive control and deliberately delayed notification acknowledgement in order to postpone the (mentally effortful) interpretation and scheduling of the incoming task. Participants' subjective accounts seem to confirm this assumption. Some of them indicated that when interruptions became more frequent, they preferred to focus on completing the current subtask before shifting their attention to another one (i.e., acknowledging the notification). For example, one participant explained that "sometimes I would have the prioritization rules memorized or something to type, but would forget when distracted by a notification. In these instances, I found it best to finish my train of thought". This strategy also aligns with participants' behavior at the interpretation stage. During high frequency interruption periods, participants took longer to start notification interpretation after acknowledgement, and took more time to complete the interpretation. One participant mentioned: "I found it easy to push the pedal when I heard the notification noise, but addressing it and knowing if/when to switch tabs in order to optimize my score/performance was difficult to do on the spot." This suggests that anticipated difficulties at later stages of interruption management have the potential to affect strategies at earlier stages. In other words, interruption notifications that are perceived or expected to be difficult to interpret may prolong their acknowledgement even before interpretation is started.

Delays at the detection and interpretation stages, particularly during high frequency periods, further contributed to the poor switching performance (i.e., interruption lag) at the integration stage. This took the form of expiring task notifications in the first experiment, particularly for nested interruptions, and longer interruption lag in the second experiment (again, particularly for nested interruptions). These performance decrements can be accounted for, in part, by the limitations of prospective memory. Even if the intention is to defer a task for only a few seconds, failures of prospective memory occur easily in the presence of distractions, interruptions, and high ongoing task demands (Brewer et al., 2020; Einstein et al., 2000, 2003; Marsh et al., 2006; McDaniel et al., 2004; Schaper & Grundgeiger, 2018).

Past research suggests several methods (for a review, see Guo et al., 2021) for overcoming failures in prospective memory. These include: a) the use of cognitive strategies to rehearse the goal until it is executed (e.g., mental rehearsal, saying it out loud, etc.; see McDaniel et al., 1998; Morgan et al., 2013; Penningroth & Scott, 2013; Trafton et al., 2003), b) association of activities in the ongoing task with those in the delayed task in order to, for instance, trigger the activation of the deferred goal (e.g., Cook et al., 2005; Dewitt et al., 2012; Rummel et al., 2017), and c) presentation of external cues/feedback such as a visual indicator, a reminder, etc. (e.g., McDaniel et al., 2004; Wilson et al., 2020). Strategy-based mitigations, such as being instructed to actively rehearse the delayed intent, are seldom effective when ongoing task demands are high because they, in turn, create another prospective memory task (Einstein et al., 2003; Monk et al., 2008). In fact, such strategies have been shown to negatively affect performance on the ongoing task because of the sharing of limited attentional resources between performance on the ongoing task and rehearsal of the delayed intent (Einstein et al., 2003). The second proposed method—associating ongoing task activities with those in the deferred task—is not always feasible in event-driven environments (such as where the timing of tasks is unpredictable), and, like strategy-based methods, is also susceptible to interference from peripheral tasks (Dewitt et al., 2012).

There is limited evidence that the third category of design-based methods, such as external cues, can improve prospective memory recall without interfering with ongoing activities. McDaniel and colleagues (2004), for example, showed that providing a small blue dot as an external reminder significantly improved prospective memory recall after completing an interrupting task. However, the results from our first and second study suggest that such simple indicators may be insufficient for helping participants switch between ongoing and pending tasks in a timely manner. Particularly during high frequency periods, when the notifications panel may contain multiple pending tasks, such simple notifications likely become ineffective because they are not noticed or become indistinguishable from the rest of the rather busy task interface, and because they fail to provide interruption-related information such as the *relative* levels of priority of pending tasks.

A better way to aid prospective memory and, in turn, the appropriate and timely scheduling of interrupting tasks, may be the use of preattentive features, such as changing the color, location, and shape of visual elements on the interface, to support the processing—in peripheral vision—of the relative priority of current and incoming tasks in parallel with working on the ongoing task (e.g., to finish a subtask). Hicks and colleagues (2005), for example, showed that prospective memory recall can be improved by increasing the salience of cues in peripheral vision (e.g., as result of changes to color, size, etc.). The authors assert that manipulations of cue salience are particularly effective in directing the operator's attention towards cues that are not routinely processed as a part of the ongoing task (i.e., interruption notifications).

Candidate solutions to better manage interruptions were suggested by participants during post-experiment debrief interviews and surveys in Experiments 1 and 2, respectively (see **Table 4.2**). We reviewed the most frequently suggested changes and selected two potential methods for mitigating prospective memory failures. Suggestions that called for changing the nature of tasks—an approach rarely possible in real-world operations—and suggestions that did not directly address challenges associated with the timely switching between tasks were excluded from selection as candidate solutions (see crossed-out items in **Table 4.2**).

Table 4.2 Changes suggested by study participants for improving the task interface and the interruption management process, sorted by frequency of suggestion. Items shown in bold were selected as candidate mitigation methods for implementation in Experiment 3.

Suggestion	Experiment	Frequency
~~Provide easier access to UAV information (e.g., vehicle type) for flight request task~~	1,2	15
Color-code tasks by urgency level	1,2	10
~~Provide quick access to frequently used phrases (e.g., checkboxes, menu of words, etc.)~~	1,2	9
Automatically sort interruption notifications by urgency level	1,2	8
Automatically fill-out known flight information for the flight request task (e.g., requested cargo)	1,2	7
Automatically select the next task based on urgency level	1,2	6
Show countdown of remaining time to complete interrupting tasks	1,2	6
Provide easier access to task rules (e.g., printed copy, always visible on interface, etc.)	1,2	5
Add keyboard shortcuts for task actions (i.e., copy/paste, load cargo, uplink route)	1	4
Make checklist completion automatic/optional for the flight request task	1,2	3
Show a warning when tasks are about to expire	1	3
Show multiple tasks on the interface (e.g., split screen)	2	3
Divide work between two operators (new flight and interruptions)	1	2
Make the name of onboard cargo visible next to UAV on the map display	1	2
Save progress on the interrupted task	2	2
Add labels for selected task action (i.e., accepted, delayed, or handed off)	2	1
Allow turbulence and noise overlays to be visible at the same time	1	1
Auto-correct typing mistakes	2	1
Automatically populate notification panel, rather than requiring acknowledgement	1	1
Automatically sort route suggestion columns in order of priority (e.g., based on vehicle type)	1	1
Automatically switch back to interrupted task	2	1
Change authorization task to a checklist, rather than requiring typing	2	1
Change color of text when the cargo is loaded, and the route is uplinked	1	1
Change notification chime based on urgency level	2	1
Highlight vehicle destination for the flight request task	1	1
Improve the quality of images showing the roof type	2	1
Increase the font size of UAV health information display	2	1
Keep accepted tasks in notifications panel until completion	1	1
Provide ability to manually highlight parts of the task interface for easier return	2	1

Provide reference table for determining relative task urgency	1	1
Show suggested route options in a table with gridlines	1	1
Show type of dispatch station without hovering in the select alternate task	1	1
Sort cargo dropdown alphabetically (rather than by urgency level)	1	1

Participants' second most frequently-suggested solution was the color coding of notifications based on the level of urgency—and more generally, the level of priority—of pending tasks. Color coding may be an effective means of distinguishing between the various levels of urgency (e.g., Alqahtani & Histon, 2012; Friedrich & Vollrath, 2022). Friedrich and Vollrath (2022), for example, show that urgency-based manipulations of color are an effective means of increasing the relative salience of icons in displays supporting supervisory control operations. The authors found that such a mapping of salience-based color to urgency allowed operators to search for and react more quickly to critical alerts, and, at the same time, prevented less urgent system states from unnecessarily drawing visual attention.

The implementation of color-based encoding of urgency in the present research requires an automated decision aid that is capable of mapping cargo name to low, medium, or high level of urgency. Recall that vehicles carrying medical cargo like vaccines are considered highly urgent, those carrying perishable items such as takeout orders are categorized as medium urgency, and vehicles with non-perishable items like pet supplies are marked as low urgency. As proposed by participants, the color-based indication of urgency is likely limited in supporting prospective memory and task switching. It may help better distinguish the urgency level of incoming tasks but it does not convey the incoming task's priority level *relative* to the ongoing task (i.e., whether an incoming task has a priority level higher, lower, or equal to the ongoing task). In the next chapter, we will therefore develop and describe in more detail an enhanced visualization of task urgency that combines color and location.

An automated decision aid is needed also for the fourth most frequently suggested change in **Table 4.2**—automatic sorting of pending tasks by level of urgency. Automatic sorting is different from color-based mapping of urgency in that color-based encoding relies on manipulation of hue to differentiate urgency level of task notifications whereas automatic sorting relies on the reordering of notifications such that higher urgency items are displayed above ones with a lower urgency. We suspect that because automatic sorting still requires notifications to be processed in focal vision to determine the actual level of urgency (unlike color which can be processed to a certain extent through peripheral vision), it will not be much more effective than the original interface in preventing failures of prospective memory. Still, automatic sorting is being included in the third experiment as it helps determine whether any observed differences in interruption management performance are simply the result of effort reduction through automation or whether more effective visualizations that support preattentive reference are needed.

Chapter 5 – Improving Switching Performance

5.1 Introduction

In **Chapter 2** and **Chapter 3**, we found that participants struggled with the appropriate scheduling of incoming tasks. They took longer to switch to nested interruptions of higher urgency, compared to both single and serial interruptions. Part of this delay was a result of delayed acknowledgement and delayed interpretation at the early stages of interruption management; another contributor was breakdowns in prospective memory, i.e., forgetting to switch to pending tasks of higher urgency. Based on suggestions from study participants, we selected two candidate methods for mitigating these issues: (1) automatic sorting and (2) color-coding of task notifications based on level of urgency. Automatic sorting involves reordering notifications for the participant so that incoming higher urgency tasks are displayed above ones with a lower urgency. Color-coding involves showing a red, yellow, or green border around notifications to represent high, medium, or low urgency tasks, respectively.

By themselves, these techniques may not be sufficient, however. Recall that appropriate switching between tasks involves several steps: 1) map cargo name to level of urgency for the interrupting task(s), 2) compare the level of urgency of the interrupting task to that of the ongoing task, 3) determine the appropriate action to take based on the relative level of urgency, and 4) switch to the pending task when appropriate. In the form described above, both automatic sorting and a color-based urgency visualization would provide support only for the first step of mapping of cargo name to urgency level and then sorting the overall set of pending tasks accordingly. However, results from the first two experiments show that despite the presence of notifications as a persistent cue, participants failed to reliably switch to incoming tasks of higher urgency, possibly due to the notifications ‘fading’ into the background over time. Neither automatic sorting, nor color-based urgency mapping address this issue. A more effective mitigation method might be one that a) supports not only automatic interpretation of urgency level, but also the comparison of *relative* levels of urgency between interrupting and ongoing tasks, and b) selectively guides visual attention to only the most relevant parts of the task interface.

The color-based visualization was modified to accomplish those goals. Specifically, color coding was combined with another preattentive feature that allows for the parallel processing of task urgency of both the ongoing and interrupting tasks, without the need to shift focal attention (Woods, 1995). According to Barrera-Leon et al. (2023), there are three preattentive attributes in addition to color: form (e.g., orientation, size), motion, and spatial position. In the context of the interruption management paradigm employed in this study, spatial grouping and location may be particularly effective in conveying the relative levels of task urgency. As shown in **Figure 5.1,** this can be done by dividing the task notification panel into three “bins” that categorize incoming notifications as high (top bin), medium (middle bin), and low (bottom bin) urgency. To represent the urgency level of the ongoing task, two vertical color- and location-coded bars were also added to the left and right of the task area. These vertical bars changed color and location between top/red, middle/yellow, and bottom/green to indicate an ongoing task with a high, medium, or low level of urgency. A green, yellow, and red color-palette was used in this experiment due to its natural and intuitive mapping to increasing levels of urgency. Notably, other hue combinations may be used instead to improve the accessibility of the task interface, such as to accommodate persons with color deficiencies (e.g., Chaparro & Chaparro, 2017). Regardless of the choice of specific hues, the redundant encoding of color and location results in a display where relative urgency becomes an easily observed emergent feature that indicates whether an incoming task is of equal, lower, or higher level of urgency than the current task.

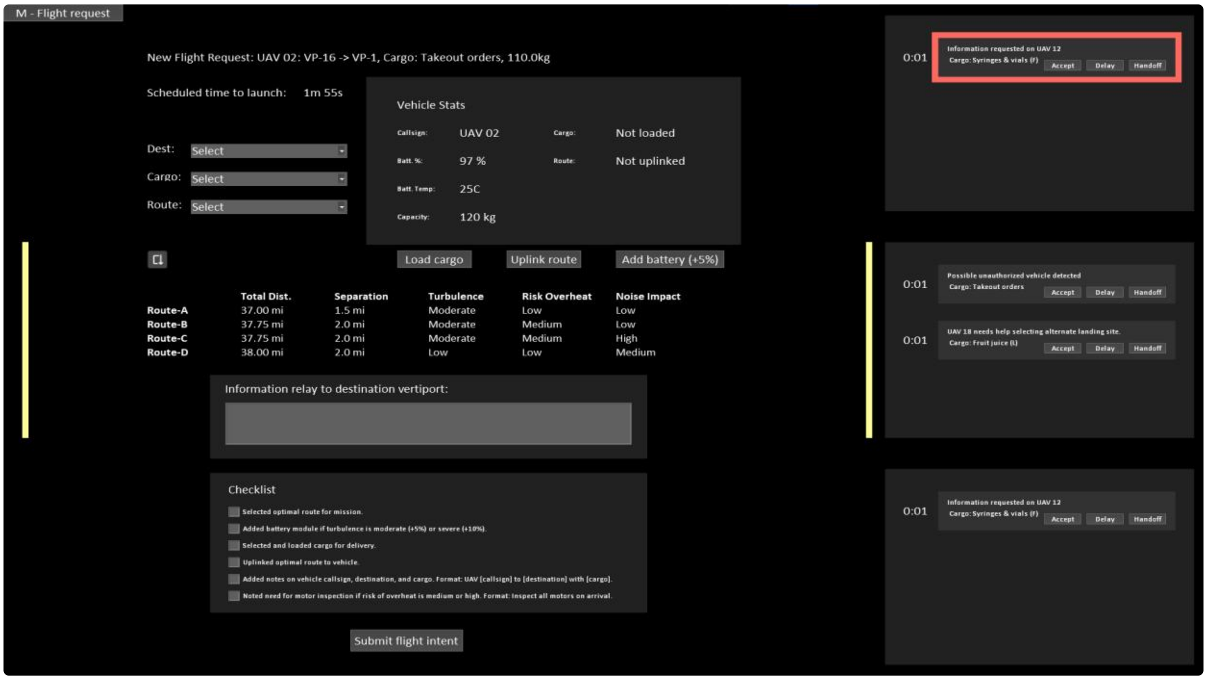

Figure 5.1. Task interface for the Visual group. The color and location of vertical bars indicates the urgency level of the ongoing/current task. Location of notifications (right) indicates urgency level of pending tasks. Red, yellow, green colors, and top, middle, bottom locations map to high, medium, and low urgencies, respectively. A colored border around a notification is only displayed when an incoming interruption notification has a higher urgency than the ongoing task. Pictured is a medium urgency flight request task on the left and a pending notification of high urgency with a red border in the top right.

The failure of notifications to serve as reminders due to ‘fading’ was addressed by limiting which notifications are color-coded. Because the dimension of color can be processed to a certain extent through peripheral vision (Wolfe & Horowitz, 2004), we expect that a color-based visualization would be more effective than automatic sorting in serving as a reminder to switch to a pending task. However, color is effective only if it is used sparsely. Therefore, rather than showing the border color of all pending task notifications, the most relevant ones can be highlighted conditionally, based on whether a pending task has a level of urgency higher than the ongoing task. Reducing the number of elements that are color-coded and thus guiding visual attention more effectively has been shown to activate intentions in support of prospective memory (e.g., Hicks et al., 2005).

It should be noted that one downside of introducing a visualization based on automatic interpretation of relative task urgency is the possibility of over-reliance on automation and a failure to catch automation failures (Parasuraman & Riley, 1997; Schuler & Yang, 2022). In the context of this study, an automation failure can occur, for example, when a high urgency task of carrying vaccines is incorrectly marked as low due to the cargo name being temporarily unavailable at time of notification onset. To explore this risk, we include in this study rare automation failures due to incorrect interpretation.

In summary, the main goals of this experiment were to 1) evaluate to what extent automating the assessment and sorting of notifications improves the participants' ability to appropriately prioritize and switch to incoming interruptions, and 2) determine whether additionally introducing visualizations of relative task urgency leads to significant further performance improvements. With this in mind, we propose the following hypotheses:

H1: *Participants in the Visual group will perform better overall, compared to the Sort and baseline groups. Participants in the Sort group will outperform participants in the Baseline group, but not the Visual group.*

H2: *Participants in the Visual group will acknowledge notifications more quickly, compared to the Baseline and Sort groups. Participants in the Sort group will acknowledge notifications more quickly than the Baseline group.*

This is based on the finding from previous experiments that participants delayed acknowledgement of notifications to deliberately postpone the mentally effortful interpretation of pending tasks. Because we expect the parallel processing of relative urgency level in the case of the visual aid to be less mentally effortful, we expect that participants may be less prone to deliberately delaying notification acknowledgement. Performance in the Sort group is expected to be better than Baseline due to the need to review and compare the urgency level of fewer notifications.

H3: *Participants in the Visual group will be more accurate and quicker when deciding whether and when to accept incoming tasks, compared to the Baseline and Sort groups. Decision accuracy and speed will not be different between the Baseline and Sort groups. In the case of automation failures, we expect that both the Baseline and Sort groups will outperform participants in the Visual group.*

Improvements in decision speed and accuracy are expected to be a result of not having to keep track of or obtain information on the urgency of the current task. Because there is a risk that participants will rely solely on the color of the notification to determine the urgency, we expect that participants in the visual group may be more likely to miss cases where the task urgency is misclassified by the automated decision aid.

H4: *Participants in the Visual group will attend to higher urgency incoming tasks more quickly, compared to the Sort and Baseline groups. This will be true particularly for nested interruptions, compared to single.*

H5: *Compared to Baseline and Sort, participants in the Visual group will be better at integrating and prioritizing pending tasks. They will be more accurate at prioritizing pending tasks, have a lower delay on the flight request task, and be more accurate and faster when switching back to interrupted tasks.*

5.2 Method

5.2.1 Participants

57 students (30 female, 26 male, 1 non-binary) from the University of Michigan participated in the study. Ages ranged between 18 and 30 years (M = 22.5, SD = 3.24). Participant eligibility was limited to this age range which is comparable to that of UAV operators. Eligibility was also limited to participants with normal or corrected-to-normal vision without color-deficiencies (i.e., color blindness). Participants were paid $17/hour for completing one three-and-a-half-hour experiment session, and a $25 performance bonus was awarded to participants with performance scores in the top quartile. This research was conducted in compliance with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at the University of Michigan (IRB # HUM00230170). Informed consent was obtained from each participant before the experiment.

5.2.2 Experiment Design

The study featured a 3 (aid type: baseline, sort, visual) x 2 (nesting level: single, nested) x 3 (interruption urgency: low, medium, high), fractional factorial design. Each participant was assigned to the Baseline, Sort, or Visual group using a random number generator. The Baseline group was not given any visual indications of task urgency. For the Sort group, all interruption notifications in the notification area were automatically sorted by their level of urgency. Higher urgency notifications were displayed above lower urgency ones. For notifications of equal level of urgency, older notifications took precedence and were displayed above more recent ones. In addition to automatic sorting, participants in the Visual group were provided with color- and location-based indications of relative task urgency (urgency of the interrupting task relative to that of the ongoing task; see **Figure 5.1**).

Nesting level (either single or nested) and task urgency were varied within-subject. Task urgency varied between low-medium-high for the flight request task and single interruptions, and between medium and high for nested interruptions. Each experiment scenario involved a total of 28 interruptions presented across 12 ongoing task trials and took approximately 25 minutes to complete. The layout of Scenario A is shown in **Figure 5.2**. Note that in the second to last trial of each scenario, a single interrupting task notification was intentionally sorted incorrectly or was presented in the incorrect bin. This event simulated an automation failure to examine whether participants would tend to overrely on the system.

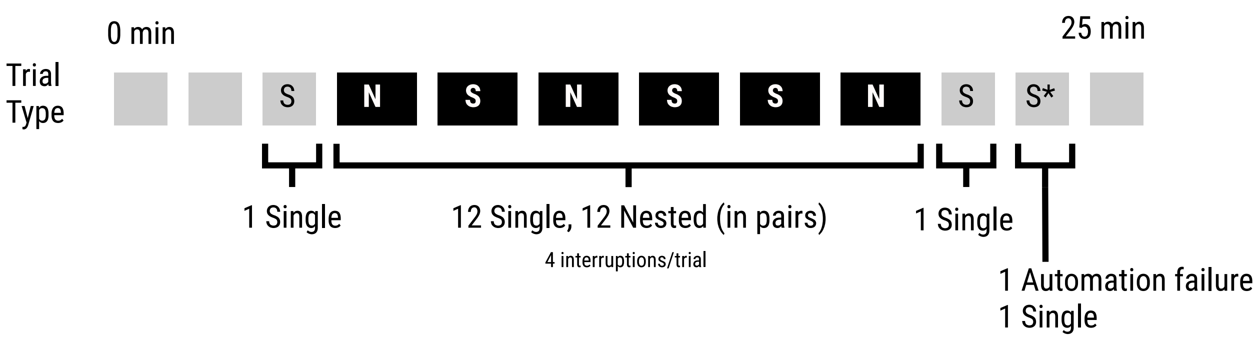

Figure 5.2 Design of one of two experiment scenarios, showing the sequence and number of ongoing and interrupting tasks. ‘S’ indicates a single interruption trial, and ‘N’ indicates a trial with nested interruptions. Blank gray squares indicate uninterrupted trials.

5.2.3 Experiment Apparatus and Tasks

Participants were tasked with supervising a set of autonomous drones delivering cargo to commercial locations in the Houston, Texas metropolitan area. Participants performed the same set of five tasks as those detailed in **Chapter 2**. These included approving flight requests, selecting alternate landing sites, diverting to alternate vertiports, detecting unauthorized aircraft, and responding to requests for vehicle status information. Participants switched between ongoing and interrupting tasks by clicking on the desired task in the notifications panel, which was located on the right side of the left-hand monitor. As in the previous study, the task area was displayed immediately to the left of the notifications panel. Only one task was visible and interactive at any given time during the scenario. The right-hand monitor displayed vehicle health and mission progress information for each of the 20 simulated aircraft. Like the previous experiments, a foot pedal was used to log the acknowledgement of interruption notifications.

5.2.4 Procedure

Participants were asked to attend one three-and-a-half-hour session that involved training, three practice scenarios, and two experiment scenarios. Before beginning the training, participants completed a questionnaire on gaming experience and multitasking ability (**Appendix D**). After the pre-experiment survey, participants completed an approximately one-hour long PowerPoint-based training session on managing UAV fleet operations for transporting cargo. Compared to the training provided to participants in the Baseline group, Sort and Visual group participants received one additional slide on how to interpret the sorted notifications panel or the visual aid. Participants in the Sort group were instructed that notifications were sorted by urgency level and time of arrival, and participants in the Visual group were instructed on how to interpret urgency of incoming notifications and ongoing tasks based on their color and location on the task interface. After learning about each of the five tasks (one ongoing, four interrupting) in the training slides, participants were guided by the experimenter to practice the task in the context of the UAV-simulator. Participants were told to prioritize tasks based on the relative urgency between the ongoing and incoming tasks. They were asked to 1) always hand off lower urgency tasks, 2) always delay incoming tasks with the same urgency level as the ongoing task, and 3) always accept tasks of higher urgency as soon as possible. Participants were informed that tasks would be scored based on the level of urgency and completion time. Points were awarded using the same scheme as in the second study (see **Table 3.1**), with one exception—flight request tasks were considered delayed if not completed within *two-and-a-half*, rather than two, minutes after onset. This change was made to create a balanced tradeoff between the time available to complete the ongoing and interrupting tasks during the prolonged high-interruption frequency period in this study.

After completing the training module and the first practice scenario, participants completed a short questionnaire on their understanding of task management and scoring (**Appendix E**). This questionnaire was the same as the one used in Experiment 2. It included questions to verify that participants were able to correctly map cargo name to urgency level, understood which action to select for incoming tasks of higher, equal, or lower urgency, and understood that tasks were scored based on urgency level and completion time.

In the second practice scenario, participants again completed each of the five tasks three times, separately, and without any interruptions. After a five-minute break, participants completed a third practice scenario with single and nested interruptions of low and high frequency, similar to what would be experienced in the experiment scenarios. Following another five-minute break, each participant completed two experiment scenarios (counterbalanced), lasting approximately 25 minutes each, with a 10-minute break between. Participants completed the third practice scenario and the two experiment scenarios with the Tobii Pro Glasses 2 eye tracker.

Following each scenario, participants completed a NASA-TLX survey to assess perceived workload (**Appendix B**). At the end of the experiment, participants filled out an open-ended questionnaire to share the challenges faced when handling interruptions (**Appendix G**).

Participants in the Sort and Visual groups were additionally asked to rate the helpfulness of the task interface and provide an explanation for their answer.

5.2.5 Dependent Measures

The dependent measures included task performance, eye tracking metrics, and survey responses. The performance and eye tracking metrics are listed and described in **Table 5.2**. Eye tracking data was collected using the Tobii Pro Glasses 2 eye tracker. Pupil diameter was used as a measure of perceived mental workload (Cain, 2007; Longo et al., 2022; Recarte et al., 2008). Survey data included questions about the participants' perceived workload (NASA-TLX), challenges experienced with handling interruptions, and subjective ratings of the task display on a Likert item response scale. One participant was assigned to an incorrect aid type condition; data from this participant was excluded from the analysis.

Table 5.1 Summary of performance and eye tracking metrics.

Metric	Definition
Detection	
Acknowledgement time	Time between the onset of the auditory interruption notification and activation of the foot pedal.
Interpretation	
Time to start interpretation	Time between notification acknowledgement and the first visit to the notifications panel.
Integration	
Decision accuracy	Proportion of tasks correctly accepted, delayed, or handed off, out of the total number of notifications acknowledged.
Decision time	Time between start of notification interpretation, and the acceptance, delay, or handoff of an incoming notification. The start of interpretation was estimated using the time of last visit to notifications panel after detection, and before an action was taken.
Interruption lag	Time between notification acknowledgement and switch to task.
Accuracy on primary task	Proportion of subtasks completed accurately, out of the total number of subtasks. This was calculated on an eight-point scale composed of selecting the correct cargo, destination, and optimal route, loading the correct number of battery modules, loading the cargo onto the vehicle, uplinking the route, and completing the checklist.
Time delay on primary task	Amount of time, in seconds, by which the flight request task is delayed beyond its scheduled time to launch.

Switchback accuracy	Proportion of tasks returned to on first attempt after completing an interrupting task, given that the interrupted task has an urgency level equal to or lower than other pending tasks.
Switchback time	Time taken, in seconds, to switch back to the interrupted task, given that the interrupted task has an urgency level equal to or lower than other pending tasks.
Task prioritization accuracy	Proportion of higher urgency tasks and older tasks of equal urgency completed before newer and/or lower urgency tasks. Additionally, accuracy accounted for the amount of time remaining and already spent on task for making tradeoff decisions for whether to complete the current task or switch to another one to earn more points.
Overall Performance and Eyetracking Data	
Overall score	Sum of overall scores achieved in each experiment scenario, weighted based on decision accuracy, task accuracy, and time taken to complete tasks. Automation failures were excluded from this calculation.
Mental workload	Cumulative change in pupil diameter over the course of an experiment scenario; supplemented by a subjective measure of perceived workload obtained through the mental dimension of the NASA-TLX questionnaire.
Duration. Mean fixation duration	Mean duration of all fixations in a scenario. A greater mean fixation duration is indicative of slower scanning and/or greater difficulty of processing information, whereas lower fixation duration would suggest rapid sampling and trying to quickly piece together the status of the system.
Directness. Scanpath length (pixels/second)	The sum of saccade amplitudes per second. A larger scanpath length per second indicates faster movement across the display, suggesting the user is trying to gather information from far-reaching areas.
Dispersion. Nearest neighbor index (NNI)	NNI is a measure of spatial dispersion of fixation points. It is based on the "distance from an individual to its nearest neighbor, irrespective of direction." (Duchowski, 2017, p. 174). Higher dispersion indicates sampling of a larger area of the display, and the sampling of potentially irrelevant information.

5.3 Results

To verify that the aid type groups were balanced in terms of individual differences, their age, IMSE scores, MTCSE scores, and gaming frequency were compared. Participants in the Baseline, Sort, and Visual groups did not significantly differ with respect to any of these factors.

5.3.1 Overall Score

Our first expectation was that participants in the Visual group would perform better overall, and experience lower mental workload, compared to the Sort and Baseline groups, and that participants in the Sort group will outperform participants in the Baseline group, but not the Visual group (H1). There was no significant main effect of aid type ($\chi^2(1) = 2.01$, $p = 0.145$).

However, overall scores did differ significantly as a function of participants' age ($\chi^2(1) = 6.76$, $p = 0.012$). Younger participants between ages of 18-23 performed better overall ($M = 227.63$, $SE = 7.44$, *95% CI* [212.71, 242.56]), compared to those between ages of 24-30 ($M = 194.75$, $SE = 9.25$, *95% CI* [176.19, 213.31]; $\beta = -32.88$, *95% CI* [-56.71, -9.06], $t(52) = -2.77$, $p = 0.008$). To account for the variance from age, another linear model was fitted with both aid type and age group as fixed effects. The updated model revealed that overall score (**Figure 5.3**) was significantly better in the Visual group ($M = 228.94$, $SE = 9.99$, *95% CI* [208.90, 248.99]), compared to the Baseline ($M = 198.54$, $SE = 10.07$, *95% CI* [178.34, 218.74]; $\beta = -30.4$, *95% CI* [-58.65, -2.15], $t(52) = -2.16$, $p = 0.035$), but not Sort ($M = 206.09$, $SE = 10.30$, *95% CI* [185.42, 226.77]; $\beta = -22.85$, *95% CI* [-51.48, 5.77], $t(52) = -1.60$, $p = 0.115$). There was no difference in overall score between the Baseline and Sort groups ($\beta = 7.55$, $t(52) = 0.53$, $p = 0.857$).

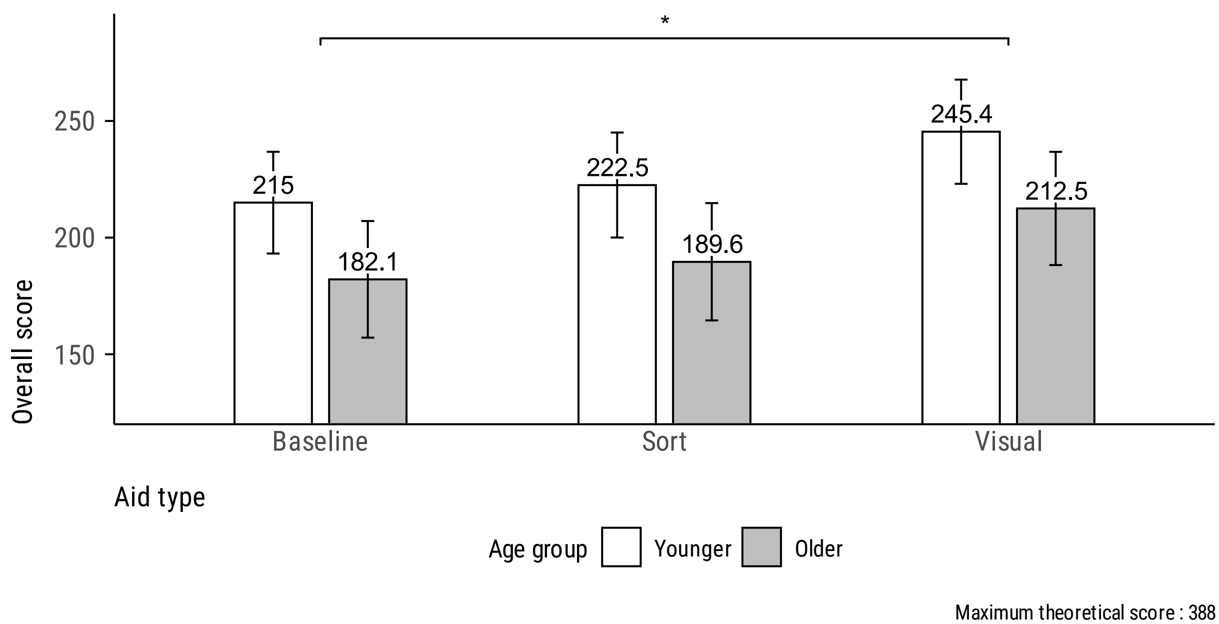

Figure 5.3 Overall score shown as a function of aid type and age group. Error bars show 95% CI.

Mental workload, as measured by the change in pupil diameter and the NASA-TLX mental score, was not different between the three aid type groups. However, older participants reported higher NASA-TLX perceived physical effort (*M* = 5.95, *SE* = 0.38, *95% CI* [5.20, 6.71]; β = 2.29, *95% CI* [1.32, 3.26], *t*(109) = 4.7, *p* = < 0.001), compared to younger participants (*M* = 3.66, *SE* = 0.30, *95% CI* [3.06, 4.26]). Mental workload, as measured through change in pupil diameter, was not different between the two groups (Younger: *M* = 0.01mm, *SE* = 0.05, *95% CI* [-0.10, 0.11]; Older: *M* = 0.14mm, *SE* = 0.07, *95% CI* [0.01, 0.27]; β = 0.13, *95% CI* [-0.03, 0.3], *t*(50) = 1.57, *p* = 0.122). Older participants also did not significantly differ in the NASA-TLX level of combined mental and physical effort exerted to accomplish their level of performance (the 'Hard' dimension; *M* = 7.81, *SE* = 0.21, *95% CI* [7.41, 8.22]; β = 0.27, *95% CI* [-0.25, 0.79], *t*(109) = 1.02, *p* = 0.308), compared to younger participants (*M* = 7.54, *SE* = 0.16, *95% CI* [7.22, 7.87]).

5.3.2 Notification Acknowledgement

At the detection stage, we expected that participants in the Visual group would acknowledge notifications more quickly than the Baseline and Sort groups (H2). This was not supported by the results. Acknowledgement time was not different between the Baseline (*M* = 1.81s, *SE* = 0.35, *95% CI* [1.12, 2.50]), Sort (*M* = 2.41s, *SE* = 0.37, *95% CI* [1.68, 3.14]; β = 0.60, *95% CI* [-0.40, 1.60], *t*(49) = 1.17, *p* = 0.247), and Visual (*M* = 1.64s, *SE* = 0.36, *95% CI* [0.93, 2.34]; β = -0.17, *95% CI* [-1.16, 0.82], *t*(49) = -0.34, *p* = 0.738) groups.

5.3.3 Decision Accuracy and Time

In H3, we predicted that participants in the Visual group would be more accurate and quicker to decide on the appropriate action for incoming tasks, compared to both the Baseline and Sort groups. We expected the opposite outcome for automation failure cases. In other words, when an incoming task was incorrectly sorted or displayed in an incorrect urgency bin, we anticipated that participants in the Baseline and Sort groups would outperform those in the Visual group. For cases where the automation performed reliably, the results partially confirm our expectation for decision accuracy, but not decision time. In case of automation failures, participants in the Visual group were indeed significantly worse at catching incorrectly mapped urgencies.

Reliable automation. Decision accuracy for incoming tasks with a correctly identified level of urgency was higher for the Visual group (M = 94.4%, SE = 1.91, *95% CI* [89.25, 97.16]), compared to Sort (M = 83.6%, SE = 4.77, *95% CI* [72.07, 91.02]; β = -1.19, *95% CI* [0.11, 0.81], z = -2.39, p = 0.017), but not compared to the Baseline group (M = 88.1%, SE = 3.58, *95% CI* [79.15, 93.55]; β = -0.82, *95% CI* [0.17, 1.16], z = -1.66, p = 0.096; see **Figure 5.3**, left). As expected, decision accuracy was the same for the Baseline and Sort groups (β = 1.45, *95% CI* [0.46, 4.54], z = 0.76, p = 0.725).

Regarding decision time (i.e., time taken to make a decision to accept, delay, or reject a task), participants in the Visual group (M = 2.82s, SE = 0.55, *95% CI* [1.71, 3.92]; β = -1.12, *95% CI* [-2.67, 0.42], $t(50)$ = -1.42, p = 0.160) did not make decisions more quickly than those in the Sort group (M = 3.94, SE = 0.56, *95% CI* [2.81, 5.07]). Surprisingly, participants in the Baseline group were faster in selecting an action (M = 2.18s, SE = 0.55, *95% CI* [1.07, 3.29]; β = -1.76, *95% CI* [-3.31, -0.22], $t(50)$ = -2.23, p = 0.030), compared to the Sort group (**Figure 5.3**, right).

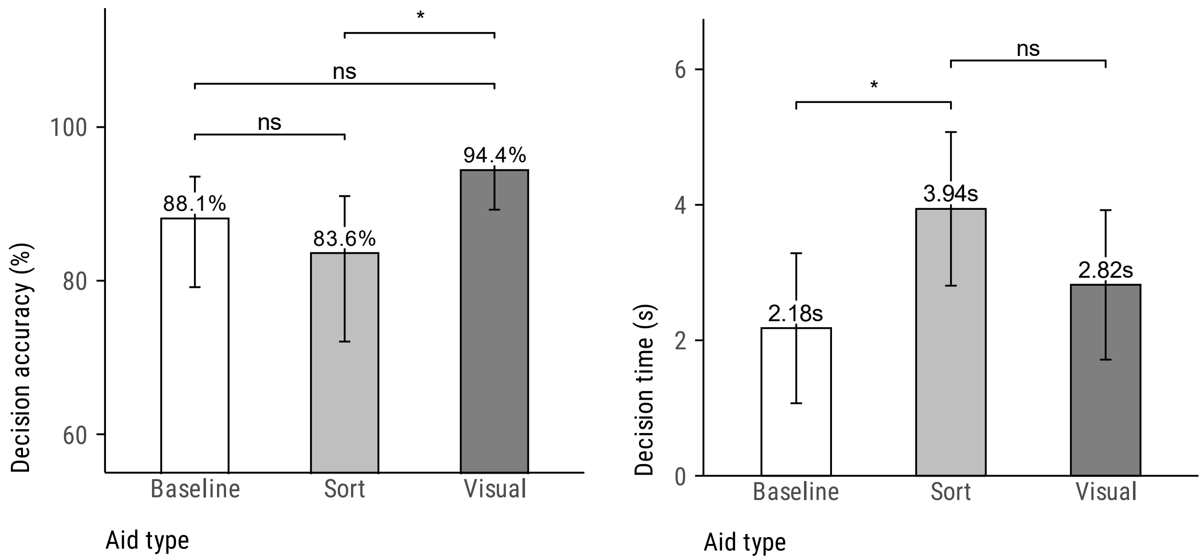

Figure 5.3 Decision accuracy (left) and decision time (right) for pending tasks across the Baseline, Sort, and Visual groups. Error bars show 95% CI.

Automation failure. Recall that decision accuracy was calculated based on participants' actual decision for each task (i.e., accept, delay, or handoff), compared to the prescribed action to take given the relative level of urgency between the ongoing and interrupting tasks. This metric was also used to determine whether participants detected automation failures—i.e., whether they correctly accepted the task regardless of a false indication of urgency level (successful detection of automation failure), rather than incorrectly handing it off (failure to detect automation failure). Decision accuracy for incoming tasks with an incorrectly identified urgency level was significantly lower in the Visual group (M = 0.0%, SE = 5.24, *95% CI* [-10.52, 10.52]), compared to both Baseline (M = 92.1%, SE = 5.20, *95% CI* [81.68, 102.53]; β = 92.11, *95% CI* [77.64, 106.57], $t(54) = 12.48$, p = < 0.001) and Sort (M = 80.6%, SE = 5.34, *95% CI* [69.85, 91.26]; β = 80.56, *95% CI* [65.89, 95.22], $t(54) = 10.77$, p = < 0.001; see **Figure 5.4**). Note that participants in the Visual group did not detect a single automation failure. There was no difference in decision accuracy between the Baseline and Sort groups (β = -11.5, *95% CI* [-29.52, 6.42], $t(52) = -1.551$, p = 0.276).

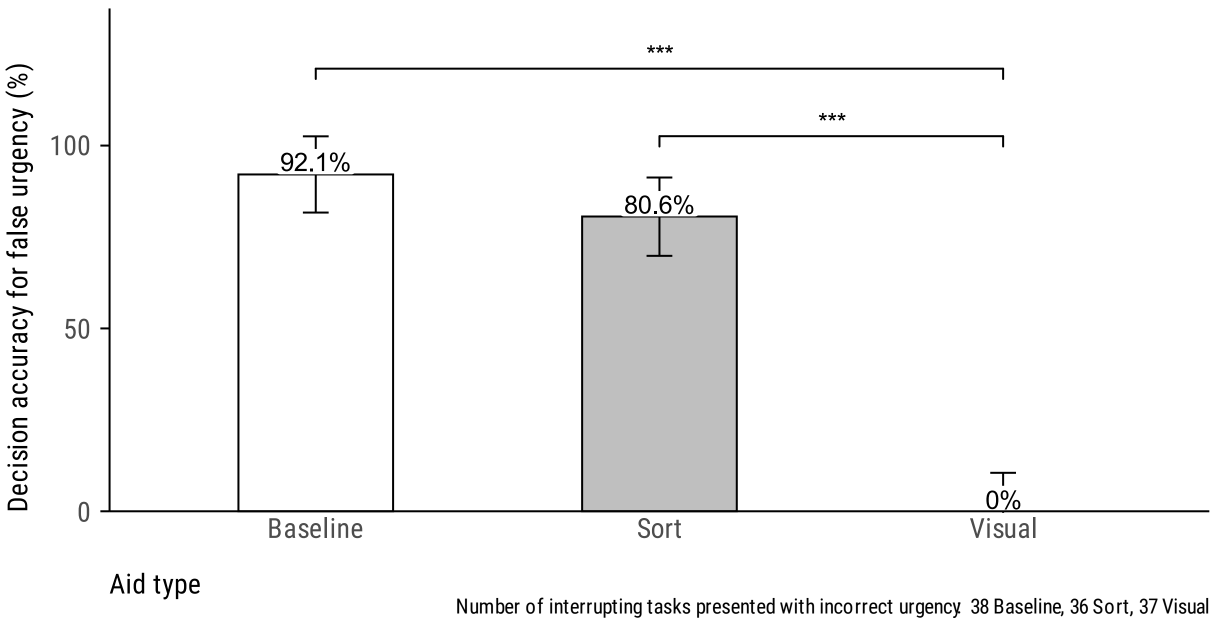

Figure 5.4 Decision accuracy for unreliable urgency indication. Error bars show 95% CI.

Eye tracking data was analyzed to compare the duration, directness, and dispersion of visual attention and scanning behavior in the notifications panel across the Baseline, Sort, and Visual groups. Described in **Table 5.1,** three metrics of mean fixation duration, scanpath length, and nearest neighbor index were calculated from the raw data. Mean fixation duration in the notifications panel was lower in the Visual group (M = 0.53, SE = 0.03, 95% CI [0.46, 0.59]), compared to both Baseline (M = 0.62, SE = 0.03, *95% CI* [0.56, 0.69]; β = 0.10, *95% CI* [0.01, 0.19], $t(53) = 2.08$, $p = 0.043$) and Sort (M = 0.64, SE = 0.03, *95% CI* [0.57, 0.70]; β = 0.11, *95% CI* [0.02, 0.2], $t(52) = 2.34$, $p = 0.023$). Scanpath length and nearest neighbor index in the notifications panel were both significantly higher in the Visual group (see **Figure 5.5)**, compared to Baseline (Scanpath length: β = -21.48, *95% CI* [-33.77, -9.2], $t(52) = -3.43$, p = 0.001; Nearest neighbor index: β = -0.22, *95% CI* [-0.35, -0.08], $t(53) = -3.11$, $p = 0.003$) and

Sort (Scanpath length: β = -17.73, *95% CI* [-30.01, -5.45], $t(52) = -2.83$, $p = 0.007$; Nearest neighbor index: β = -0.30, *95% CI* [-0.43, -0.16], $t(53) = -4.25$, p = < 0.001). For regions outside the notifications panel, such as the task area and UAV health information panel, these metrics were not significantly different across the Baseline, Sort, and Visual groups.

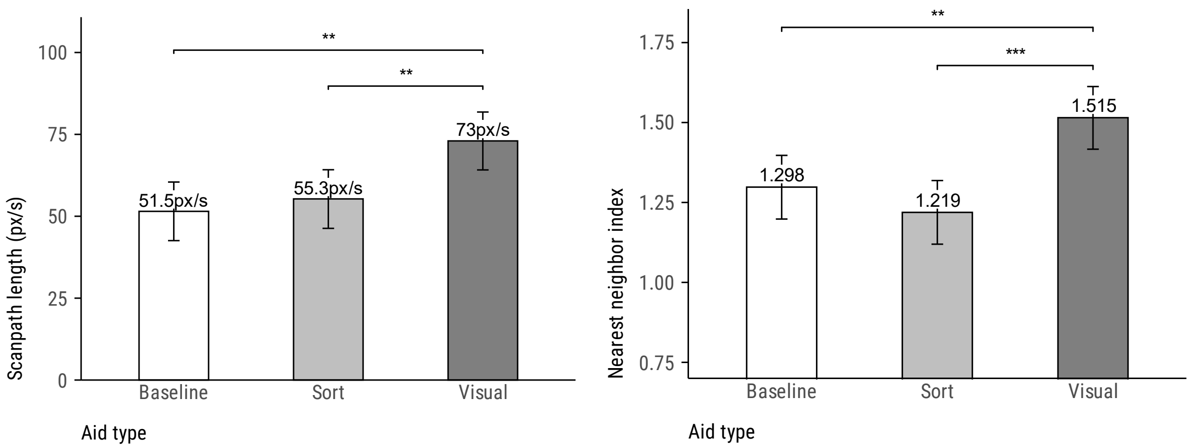

Figure 5.5 Scanpath length per second (left) and nearest neighbor index (right) in the notifications panel. Error bars show 95% CI.

5.3.4 Switching to Interrupting Tasks

We expected that participants in the Visual group would switch to pending tasks of higher urgency more quickly, compared to the Sort and Baseline groups, and that this effect would be stronger for nested interruptions, compared to single. Interruption lag, shown in **Figure 5.6**, was not different between the Baseline (M = 16.17s, SE = 3.07, *95% CI* [10.01, 22.34]), Sort (M = 14.49s, SE = 3.24, *95% CI* [7.99, 20.98]; β = -1.68, *95% CI* [-10.26, 6.9], $t(46) = -0.38$, $p = 0.702$) and Visual (M = 9.26s, SE = 3.07, *95% CI* [3.09, 15.44]; β = -6.91, *95% CI* [-15.26, 1.44], $t(45) = -1.62$, $p = 0.112$) conditions. Similarly, for nesting level, there was no main effect ($\chi^2(1) = 2.88$, $p = 0.089$) or interaction effect on interruption lag.

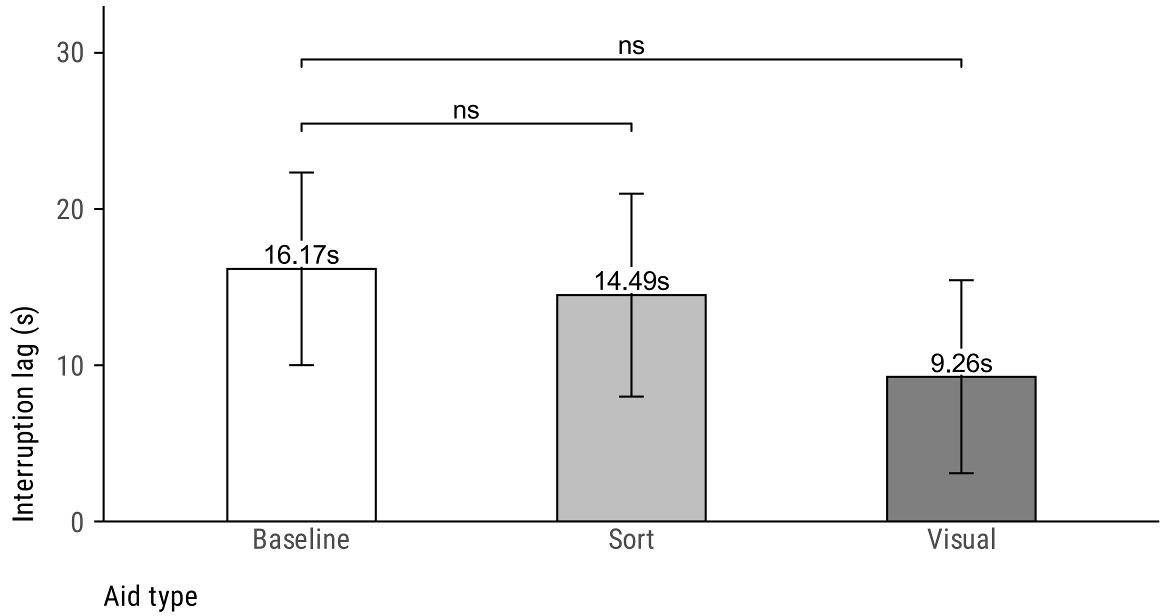

Figure 5.6 Interruption lag for incoming tasks of higher urgency than the current task. Error bars show 95% CI.

5.3.5 Task Prioritization, Performance, and Return

H5 predicted that participants in the Visual group would be better at integrating and prioritizing pending tasks, compared to those in the Baseline and Sort group. This expectation was supported by the data in terms of prioritization accuracy and accuracy and speed of switching back to the correct interrupted task, but not for accuracy and time delay on the ongoing flight request task.

Prioritization accuracy (**Figure 5.7**) was better in the Visual (*M* = 65.0%, *SE* = 2.72, *95% CI* [59.56, 70.48]) group, compared to Baseline (*M* = 56.1%, *SE* = 2.72, *95% CI* [50.68, 61.60]; β = -8.88, *95% CI* [-16.42, -1.33], $t(53) = -2.31$, $p = 0.025$), but not Sort (*M* = 60.3%, *SE* = 2.80, *95% CI* [54.69, 65.91]; β = -4.72, *95% CI* [-12.37, 2.93], $t(53) = -1.21$, $p = 0.232$). There was no difference in prioritization accuracy between the Baseline and Sort groups (β = -4.16, *95% CI* [-13.57, 5.25], $t(53) = -1.07$, $p = 0.540$).

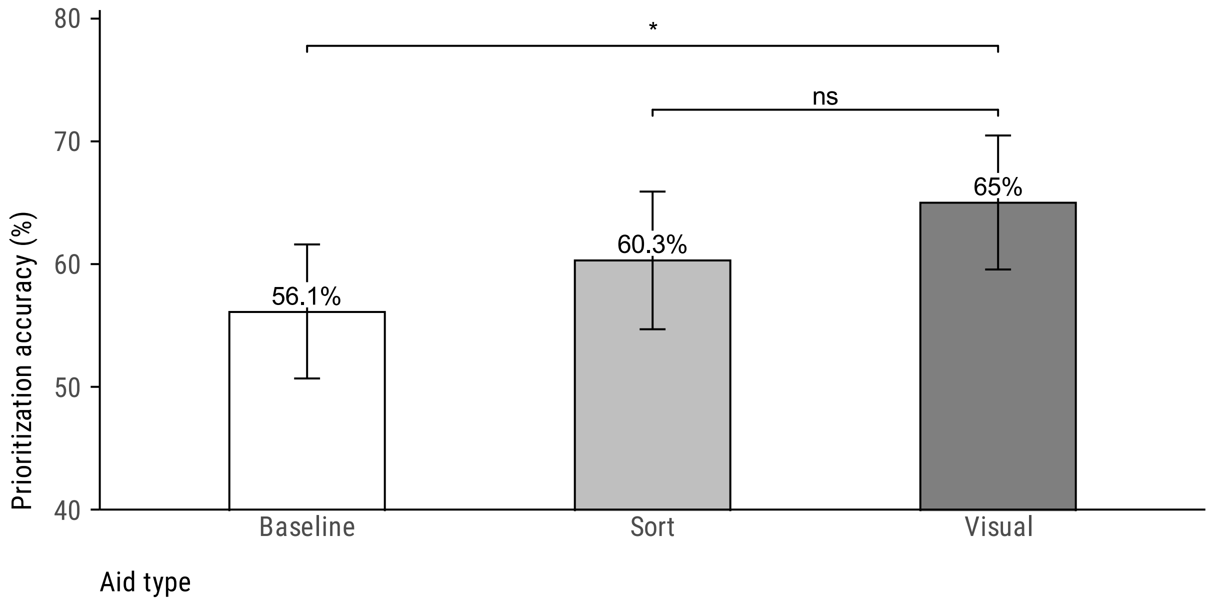

Figure 5.7 Accuracy of prioritization and integration of pending and incoming tasks. Error bars show 95% CI.

Our expectation that participants would be more accurate and experience fewer delays on the primary flight request task in the Visual group, compared to the Baseline and Sort groups, was not confirmed. A summary of these results is shown in **Table 5.2**.

Table 5.2 Accuracy and time delay on the flight request task.

	Flight Request Task Accuracy					**Flight Request Task Delay**				
Predictors	*β*	*SE*	*95% CI*	*Z*	*p*	*β*	*SE*	*95% CI*	*Z*	*p*
(Intercept)	95.28	0.84	[93.63, 96.93]	113.37	<0.001	53.62	8.69	[36.56, 70.68]	6.17	<0.001
Aid Type- Sort	-0.70	1.21	[-3.06, 1.67]	-0.58	0.565	-1.46	13.08	[-27.15, 24.23]	-0.11	0.911
Aid Type - Visual	-0.52	1.19	[-2.85, 1.82]	-0.43	0.664	9.40	12.28	[-14.72, 33.52]	0.77	0.444
Observations	668					634				
Marginal R^2 / Conditional R^2	0.001 / 0.088					0.006 / 0.305				

When returning to an interrupted task, switchback accuracy (**Figure 5.8**, left) was higher in the Visual (*M* = 97.2%, *SE* = 1.75, *95% CI* [90.8, 99.2]) condition, compared to the Baseline

(M = 88.5%, SE = 5.44, 95% CI [73.0, 95.6]; β = -1.50, *95% CI* [-2.9852, -0.0174], z = -1.98, p = 0.047) and Sort (M = 88.1%, SE = 5.20, *95% CI* [73.7, 95.1]; β = -1.54, *95% CI* [-2.9487, -0.1292], z = -2.14, p = 0.032) conditions. Time to switch to the previous task (**Figure 5.8**, right) was shorter for the Visual group (M = 1.39s, SE = 0.18, *95% CI* [1.02, 1.76]), compared to Baseline (M = 1.94s, SE = 0.19, *95% CI* [1.57, 2.32]; β = 0.55, *95% CI* [0.0434, 1.063], $t(26)$ = 2.13, p = 0.043) and Sort (M = 2.24s, SE = 0.21, *95% CI* [1.81, 2.66]; β = 0.85, *95% CI* [0.3031, 1.3956], $t(27)$ = 3.05, p = 0.005).

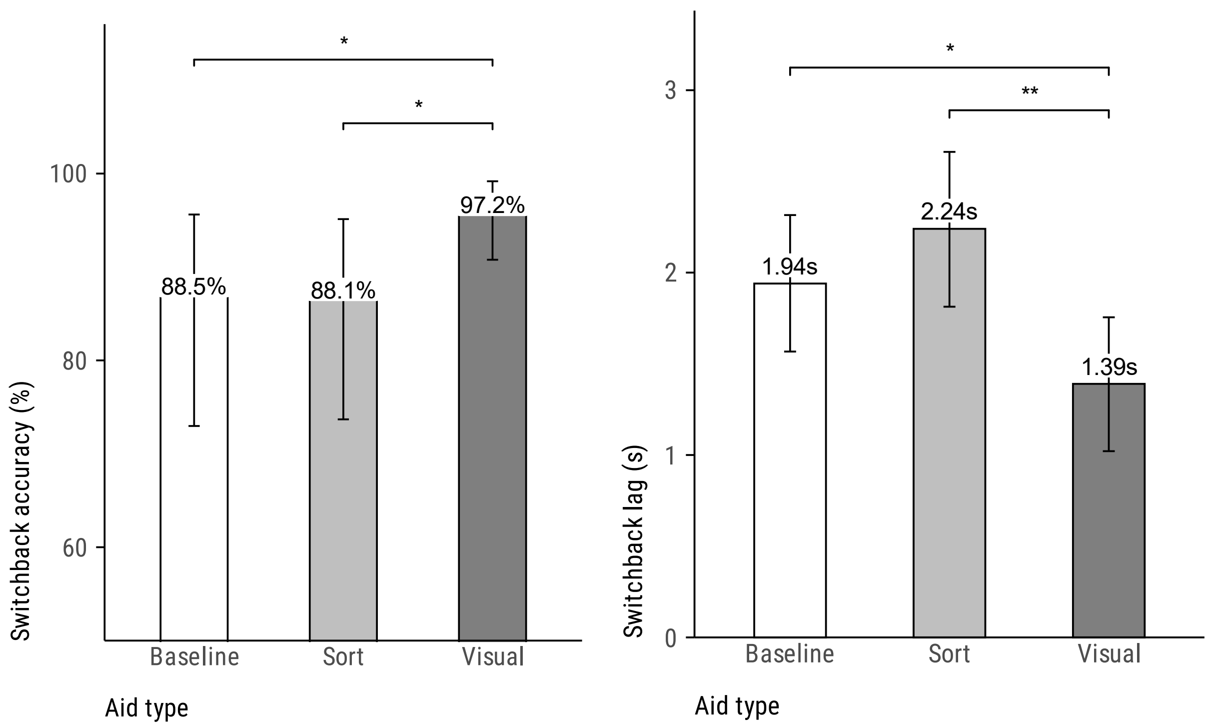

Figure 5.8 Switchback accuracy (left) and time taken to switch back (right) to interrupted task. Error bars show 95% CI.

5.3.6 Survey Results

NASA-TLX ratings are shown in **Figure 5.9**. Only the “Frustrated” dimension showed significant differences between the aid type groups. Participants in the Sort group reported higher frustration (M = 4.67, SE = 0.38, *95% CI* [3.92, 5.42]), compared to the Visual group (M = 3.32, *SE* = 0.37, *95% CI* [2.59, 4.05]; *β* = -1.35, 95% CI [-2.4, -0.3], *t*(108) = -2.56, *p* = 0.012). There was no significant difference in frustration ratings between the Sort and Baseline groups (*M* = 3.84, *SE* = 0.37, *95% CI* [3.10, 4.58]; *β* = -0.83, 95% CI [-1.88, 0.23], *t*(108) = -1.56, *p* = 0.122).

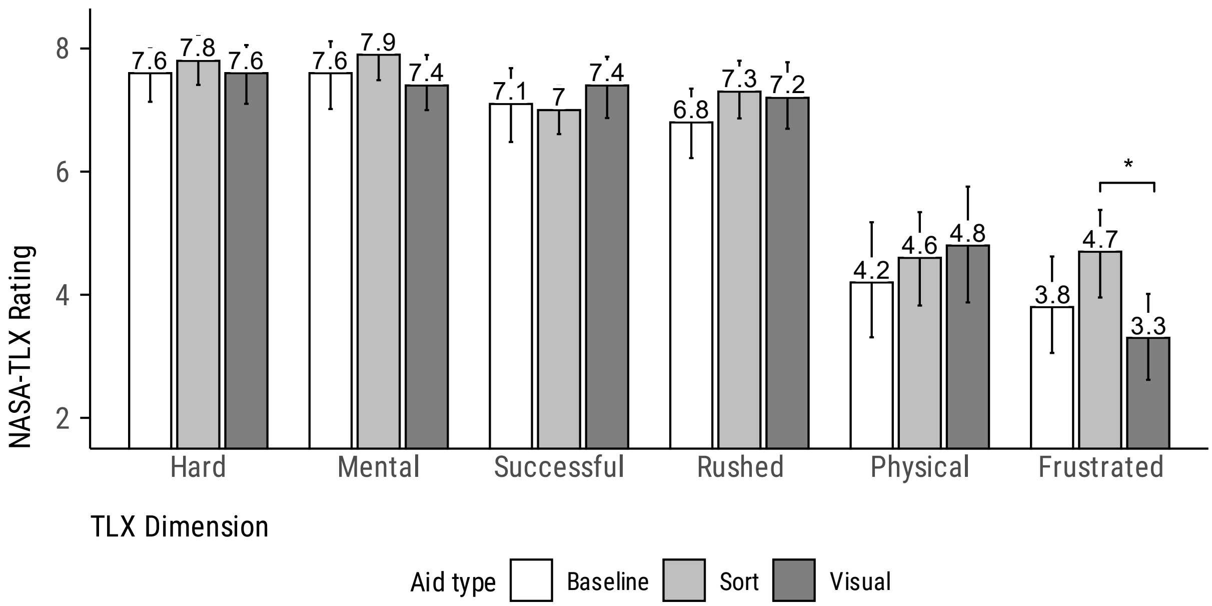

Figure 5.9 NASA-TLX ratings, compared between the Baseline, Sort, and Visual groups. Error bars show 95% CI.

As shown in **Figure 5.10**, participants in the Visual group (*M* = 2.21, *SE* = 0.27, *95% CI* [1.67, 2.75]) found it easier to determine the relative urgency of interrupting tasks, compared to both the Baseline (*M* = 3.28, *SE* = 0.28, *95% CI* [2.72, 3.83]; *β* = 1.07, *95% CI* [0.3, 1.84], *t*(52) = 2.78, *p* = 0.008) and Sort groups (*M* = 3.61, *SE* = 0.28, *95% CI* [3.06, 4.16]; *β* = 1.4, *95% CI* [0.63, 2.17], *t*(52) = 3.64, *p* = < 0.001). Action selection also was rated as being easier by participants in the Visual group (*M* = 3.00, *SE* = 0.26, *95% CI* [2.48, 3.52]), compared to both Baseline (*M* = 3.83, *SE* = 0.27, *95% CI* [3.30, 4.37]; *β* = 0.83, *95% CI* [0.09, 1.58], *t*(52) = 2.24, *p* = 0.029) and Sort (*M* = 4.00, *SE* = 0.27, *95% CI* [3.47, 4.53]; *β* = 1.00, *95% CI* [0.25, 1.75], *t*(52)= 2.69, *p* = 0.010). Similarly, participants in the Visual group rated the simulation interface as more helpful (*M* = 9.05, *SE* = 0.48, *95% CI* [8.08, 10.03]), compared to the Sort group (*M* =

7.56, *SE* = 0.49, *95% CI* [6.55, 8.56]; β = -1.5, *95% CI* [-2.89, -0.1], *t*(35) = -2.17, *p* = 0.037).. A summary of results in provided in **Table 5.3**.

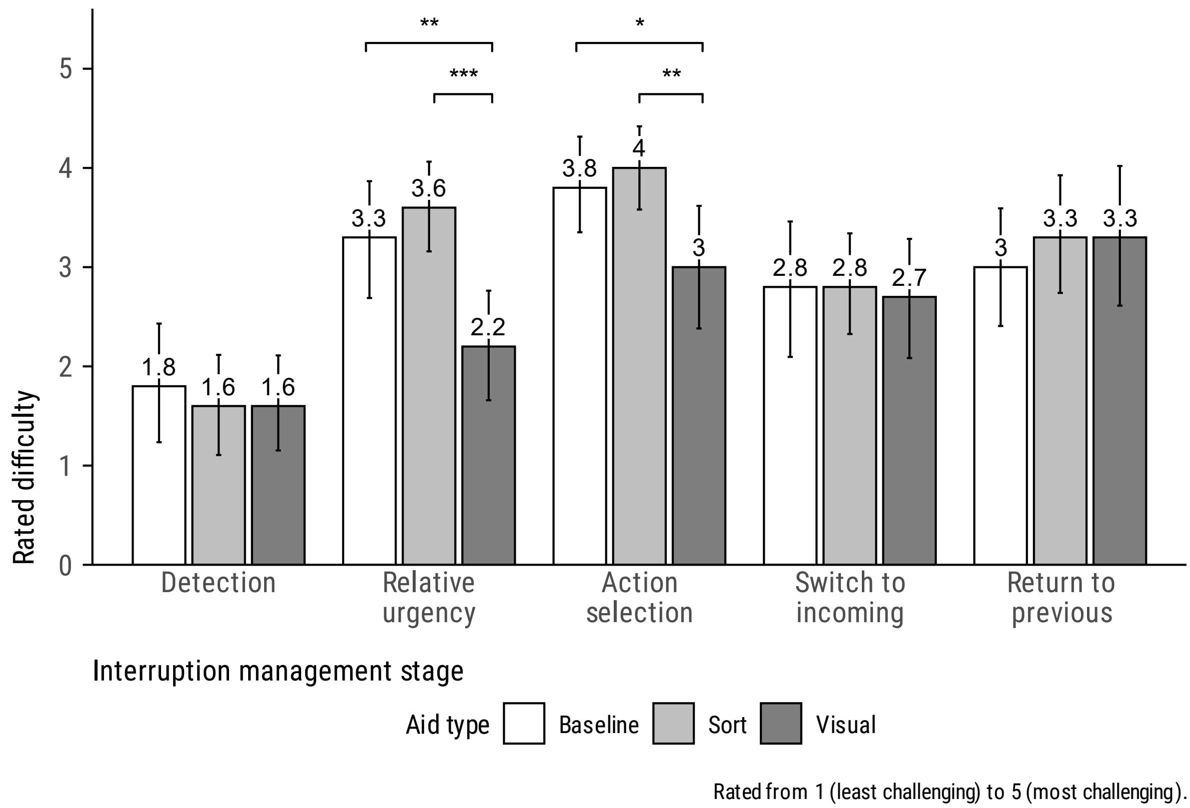

Figure 5.10 Difficulty ratings of interruption management phases, shown as a function of visual aid type. Error bars show 95% CI.

Table 5.3 Summary of study expectations and results.

Expectation	Results
H1: Participants in the Visual group will perform better overall, compared to the Sort and baseline groups. Participants in the Sort group will outperform participants in the Baseline group, but not the Visual group.	Overall score was higher for participants in the Visual group, compared to Baseline, but not Sort. Additionally, younger participants scored significantly higher overall, compared to older participants.
H2: Participants in the Visual group will acknowledge notifications more quickly, compared to the Baseline and Sort groups. Participants in the Sort group will acknowledge notifications more quickly than the Baseline group.	Acknowledgement time was not significantly different between the Visual, Sort, and Baseline groups.
H3: Participants in the Visual group will be more accurate and quicker when deciding whether and when to accept incoming tasks, compared to the Baseline and Sort groups. Decision accuracy and speed will not be different between the Baseline and Sort groups. In the case of automation failures, both the	Decision accuracy was higher in the Visual group, compared only to Sort, but not Baseline. Decision time was longer in the Sort group, compared to Baseline, but not Visual.

Baseline and Sort groups will outperform participants in the Visual group.	In the case of automation failures, participants in both the Baseline and Sort groups outperformed the Visual group, in which not a single participant detected the incorrect classification of notification urgency.
H4: Participants in the Visual group will attend to higher urgency incoming tasks more quickly, compared to the Sort and Baseline groups. This will be true particularly for nested interruptions, compared to single.	Interruption lag was not found to be significantly different between the Visual, Sort, and Baseline groups. There was also no main or interaction effect of nesting level.
H5: Compared to Baseline and Sort, participants in the Visual group will be better at integrating and prioritizing pending tasks. They will be more accurate at prioritizing pending tasks, have a lower delay on flight request task, and be more accurate and faster when switching back to interrupted tasks.	Participants in the Visual group more accurately prioritized tasks, compared to the Baseline group, but not Sort. Performance or time delay were not significantly different between the three groups. Participants made more accurate decisions and switched back to the interrupted task more quickly in the Visual group, compared to both the Baseline and Sort groups.

5.4 Discussion

The goal of this experiment was to assess the effectiveness of automatic interpretation and sorting of interruption notifications (‘Sort’) and of hue- and location-based visualization of relative task urgency (‘Visual’) for improving participants' ability to appropriately schedule and prioritize interrupting tasks. We evaluated overall performance and performance at the individual stages of interruption management across the unaided baseline condition, Sort, and Visual.

Our first expectation was that participants in the Visual group would perform better overall, compared to the Sort and Baseline groups, and that participants in the Sort group would outperform participants in the Baseline group. We found partial support for this expectation—participants in the Visual group performed significantly better than those in the Baseline group (but not the Sort group), by approximately 30 points. Recall that a high urgency task was worth 6 points. Therefore, a difference in overall score of 30 points is roughly equivalent to performing 5 additional high urgency tasks. While not statistically significant, there was a trend towards better performance in the Visual group, compared to Sort. In this case, the difference was 23 points which translates into nearly 4 additional high urgency tasks being performed. In safety-critical environments, being able to perform even one more task can have a major impact. Therefore, the observed differences are operationally highly significant and provide strong support for providing operators with the visual aid.

In addition to aid type, we found that older participants (24-30 years) performed significantly worse than younger participants (18-23 years). One possible reason for this difference could be that older people struggle with processing visual information embedded among temporally distributed distractor items (e.g., Gazzaley et al., 2005). Specifically, they found that older participants were unable to suppress brain-activity associated with task-irrelevant stimuli, which in turn also impaired working memory performance. It is possible that older participants were more susceptible to distraction from the presence of pending tasks and notifications. However, research reporting this decrement involved significantly older participants (e.g., 60+) than what was defined as 'old' in this study (24-30). Still, it is possible that effects of age on interruption management begin much earlier than is currently understood since studies tend to sample only extreme ends of the age spectrum. Interestingly, compared to the younger group, older participants rated the scenarios to be more physically (but not mentally) demanding. Combined with the fact that older participants did not report having to work harder (mentally and physically) to accomplish their respective levels of performance, it seems that the poorer performance was neither a result of age-related difficulties with processing information, nor a lack of motivation and effort. Rather, the higher physical effort rating suggests that older participants struggled more with the *execution* of the necessary steps, such as gathering relevant task information and switching between tasks with sufficient speed—in other words, they struggled not with *understanding*, but with *doing*.

The main purpose of the two interventions in this experiment—automatic sorting and visualization of relative urgency—was to support later stages of interruption management. Benefits were indeed observed at the integration stage where participants in the Visual group selected the correct action based on relative urgency more frequently, compared to the Sort group, but not Baseline. In the post-experiment subjective ratings, participants in the Visual group reported having an easier time, compared to Sort and Baseline, both when determining the relative level of urgency, and when selecting the appropriate action. One participant in the Visual group, for example, noted that "the boxes for urgency and color indicator made relative task urgency very manageable." It was surprising that participants in the Visual group outperformed those in the Sort group, but not the Baseline group. One explanation is that sorting introduced an additional layer of information processing that required participants to a) remember that the notifications are sorted (in the absence of any visual indications of the sorting), and b) keep track of the changing positions of notifications as more tasks are added to the notification panel. In fact, four participants reported at the end of the experiment that they had forgotten about the automatic sorting feature. The above findings may also explain why participants in the Visual group experienced lower frustration (according to their NASA TLX ratings), compared to the Sort group only.

Participants in the Visual group performed better when prioritizing ongoing, pending, and incoming tasks, but only compared to Baseline, not Sort. This difference in performance was reflected in participants' subjective responses. One participant in the Baseline group, for instance, said that "the hardest thing was to make sure [they were] always doing the most urgent assignment and making sure to switch back to tasks after [they were] done with interrupting tasks." Performance likely did not differ significantly between the Visual and Sort groups because the sorting of notifications by urgency level supported which task to switch to, unlike the Baseline condition which did not provide any such support. The improvements in prioritization in the Visual group, compared to Baseline, are likely a result of both a better awareness of the relative urgency of tasks, as well as support for more accurate and faster return to previously interrupted tasks.

Participants in the Visual group also made quicker and more accurate decisions regarding the return to a previously interrupted task, compared to both the Baseline and Sort groups. This was primarily due to the presence of the colored border provided to participants in the Visual group. Recall that a colored border was displayed around a notification when the incoming task had a level of urgency higher than the ongoing task (e.g., a low urgency ongoing flight request task is interrupted by a medium urgency interrupting task). In some cases, this secondary task was again interrupted by a tertiary task of even higher urgency. When the participant switched to this tertiary task, the secondary task was added back to the notifications panel so that it could be switched to and completed later. As a result, the secondary task was displayed with the colored border that it originally had when it interrupted the low urgency flight request task. This colored border likely acted as a prospective memory aid and helped participants return to the interrupted secondary task more reliably and more quickly.

Although visualization of relative urgency improved decision accuracy, task prioritization, and switchback performance, these benefits were limited to cases when the automatic interpretation worked reliably. In the case of automation failures, we found strong evidence for the negative cost of introducing visual aiding on decision accuracy. Not a single participant in the Visual group detected the incorrect categorization of a high urgency task by the automation. In comparison, 80% of participants in the Sort group noticed the error and selected the correct action. Because the urgency visualization supported the determination of *relative* task urgency, participants in the Visual group (unlike those in the Sort group) did not need to review the cargo name to select the appropriate action. As a result, participants in this group relied completely on the automation's assessments. While over-reliance on automation is not a novel phenomenon (e.g., Bainbridge, 1983; Dixon & Wickens, 2006), this finding provides evidence in support of the lumberjack hypothesis (Onnasch et al., 2014), which states that a higher degree of automation is beneficial to the extent that all goes well, but results in just as large of an impairment in performance when the automation inevitably fails. The choice of whether and what degree of automation to introduce into a system therefore involves a tradeoff between the benefits expected when the automation is reliable, and the potential costs when it fails (Sheridan & Parasuraman, 2000). In cases where the nature and circumstances of automation failure are known a priori, training can inform the operator of situations that require more engagement and effortful interpretation. In cases where automation failures are more unpredictable, it may be necessary to select lower levels of automation and use less compelling data aggregation and visualizations to reduce the risk of automation bias (e.g., Kupfer et al., 2023).

Lastly, improvements in decision accuracy, task prioritization and switching performance did not result in improved performance on the ongoing flight request task. Participants in the Visual group were not more accurate and did not experience fewer delays on the flight request task, compared to Baseline and Sort groups. It seems that reduced effort and the time saved in decision making and scheduling of tasks did not directly translate to higher accuracy or shorter delays on the flight request task. At the same time, the introduction of urgency visualization also did not result in *worse* performance. Based on the duration, directness, and dispersion eye tracking metrics, the effects of urgency visualization on scanning behavior were limited to the notifications panel. Only in the notifications panel, mean fixation duration was shorter, scanpath length was longer, and the nearest neighbor index was higher, indicating that participants scanned more quickly and over longer distances. This was likely due to the separation of notifications across bins, in contrast to the Baseline and Sort groups where notifications were populated from top to bottom in sequence. Given that participants in the Visual group gave the task interface a significantly higher rating, compared to the Sort group, the change in scanning behavior did not seem to have a negative effect on the participants' perception of the task interface overall.

One limitation of this study was that it employed a visualization of task urgency based on the redundant display of both color and location. As a result, we were unable to isolate what portion of performance benefits resulted from color coding and what portion was a result of encoding urgency with location. Future studies should compare the color and location dimensions both separately and redundantly to better understand the contribution to performance benefits of each. This may be useful, for example, in cases where limited screen space prevents the encoding of task priority based on location.

Chapter 6 – Conclusion

By some accounts, interruptions are desirable events that carry valuable information to enable people to successfully coordinate activities in highly dynamic, real-world environments. However, the large and increasing number of interruptions in many safety-critical application domains, such as medicine and aviation, has become a challenge for operators whose attentional resources and multitasking abilities are limited. The problem is expected to worsen with the emergence of large-scale multi-agent operations such as Urban Air Mobility or military unmanned aerial system operations which will require a few ground-based operators to maintain awareness of the status and behavior of large numbers of heterogeneous vehicles (FAA, 2020; Mueller et al., 2017).

To date, most research on interruption management has focused on paradigms involving only a small number of interruptions, and immediate and forced interruptions that must be completed at the time of onset. This work fails to address the current and upcoming challenges associated with more frequent interruptions (Andreasson et al., 2017; Baethge et al., 2015; Laarni, 2021). The two main goals of this dissertation were therefore to:

Goal 1: Identify and analyze the challenges that operators encounter when managing negotiated interruptions that are frequent and potentially nested.

Goal 2: Develop and evaluate a set of candidate displays to mitigate the identified challenges.

Goal 1 was addressed by the first two experiments, detailed in Chapters 2 and 3. The first experiment focused on identifying the difficulties faced by operators when coping with frequent and nested interruptions in a supervisory command and control task. Results indicated that frequent and nested interruption notifications were less likely to be acknowledged, compared to less frequent and non-nested ones. Still, the overall rate of acknowledgement was high. Interpretation accuracy suffered more significantly for single and nested interruptions, but it remained unclear whether the performance decrements resulted from problems at the interpretation or the integration stage.

Experiment 2 investigated this issue and examined more closely whether breakdowns in handling frequent and nested interruptions occurred due to the participants' inability to a) correctly classify the urgency of interruption notifications, b) activate the appropriate rule associated with the relative urgencies of ongoing and interrupting tasks, or c) switch between tasks in a timely manner. The results from the second experiment indicated that participants did not struggle to accurately interpret the urgency level of current and incoming interruptions and were able to reliably determine the appropriate action to take based on the relative urgency between ongoing and incoming tasks. However, participants struggled with the appropriate scheduling of incoming tasks, and took longer to switch to nested interrupting tasks of higher urgency, compared to both single and serial interruptions. The longer switch time comprised of delays at earlier stages when acknowledging and interpretating notifications, and due to a resistance to switch away from the ongoing task, even for highly urgent interrupting tasks.

Two candidate displays were developed to address issues with poor scheduling of pending tasks. The first candidate display involved automatic sorting of incoming task notifications by level of urgency, and the second candidate method involved the visualization of the relative level of urgency through the peripheral visual processing of notification color and location. The two mitigation methods were empirically compared against the unaided baseline condition in the third and final experiment in this line of research. Findings indicate that visualization of relative urgency of interrupting tasks improved the overall score, resulted in higher accuracy of selecting the correct action to take, and led to more efficient prioritization of ongoing and interrupting tasks. At the same time, results in the event of automation failures highlighted a potential downside—participants displayed a strong over-reliance on the automatic interpretation of task urgency and failed to select the appropriate action when notifications were categorized into the wrong bin.

6.1 Intellectual Merits and Broader Impact

The theoretical contribution of this research is a better understanding of the challenges and performance breakdowns at each of the three stages of interruption management—particularly when dealing with frequent and nested interruptions and those that are negotiated, rather than forced. This dissertation has examined interruptions across the detection, interpretation, and integrations stages, whereas much of the earlier research has focused primarily on the integration stage. By looking across all stages, we were able to show that interruption management is not a linear process as suggested by early models of IM, but one where behavior and performance at one stage depends on anticipated and experienced difficulties at both earlier and subsequent stages.

Based on the results from the first two experiments, **Figure 6.1** highlights two modifications that were made to the model of interruption management adopted from Latorella (1997) and initially presented in **Figure 1.2** in **Chapter 1**. At the detection stage, we observed that the interruption signal itself is a form of interruption that disrupts ongoing tasks and lines of reasoning. As a result, anticipated difficulties at the subsequent stage of interpretation influenced the time taken to acknowledge interruption signals, particularly during high frequency periods (indicated in **Figure 6.1** as a double-headed blue arrow pointing from the interpretation stage to the acknowledgement stage). First, this suggests that when dealing with frequent interruptions, the need to acknowledge potential interruptions influences operator behavior more strongly than is recognized by current models of interruption management like Latorella's (1997) IMSM. Second, our findings show that in addition to limitations of working memory and attention, as defined by MFG and threaded cognition models, top-down executive control plays a part in how quickly operators acknowledge interruptions. We found this effect to be a result of situational factors like interruption frequency, nesting level, and even interruption position—each were associated with delayed acknowledgement of interruption notifications. In the updated model in **Figure 6.1,** we therefore recognize acknowledgement as a fourth stage in addition to detection, interpretation, and integration**.**

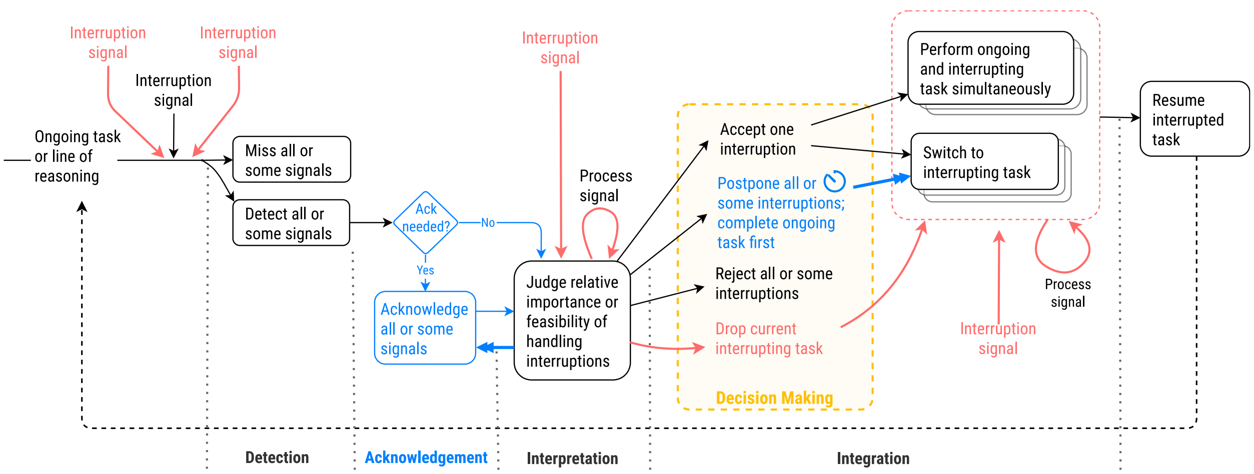

Figure 6.1 The interruption management process, updated based on findings from Experiments 1 and 2. New and modified elements are shown in blue. 'Ack' = Acknowledgement.

The more prominent role of acknowledgement indicates a need for the design of signals that can more effectively guide the operator's attention and are robust to feedback effects from subsequent stages. These feedback effects may be much more pronounced, for example, in the context of alarm floods where a large number of alarms are triggered in a very short amount of time (Perrow, 1999).

Notably, our findings highlight the need to consider not only the reliable detection and identification of signals in alarm floods, but also their appropriate integration, which prior work has largely ignored (Wan, 2019). At the integration stage, nested interruptions (more so than single or serial ones) resulted in delays in switching to pending tasks of higher urgency, partly due to cumulative delays in acknowledgement and interpretation at earlier stages, and partly due to failures of prospective memory to remember to switch to the pending task. This effect is indicated in **Figure 6.1** at the integration stage as a blue timer icon and a double-headed blue arrow next to the 'postpone all or some interruptions; complete ongoing task first' action. These results provide a) confirming evidence that the limitations of prospective memory, such as the intention to switch to a task, are present at short intervals (e.g., Einstein et al., 2003), as well as b) new evidence that these effects are exacerbated by interruptions that are both frequent and nested.

The second contribution of this work is guidance on the design of interruption-resilient interfaces. Results from the third experiment provide empirical evidence that the visualization of relative task priority (not simply automatic sorting according to absolute urgency of the interrupting task) is an effective means of supporting the management of frequent and nested interruptions. Specifically, a location- and color-based preattentive visualization of task priority improves overall performance, increases decision-making accuracy, leads to more efficient prioritization of ongoing and pending tasks, and results in more accurate and faster return to interrupted tasks. Attention-directing benefits of preattentive features are well-documented in current research, but their application is limited either to simple visual tasks outside the context of interruption management (e.g., Barrera-Leon et al., 2023; Wolfe & Utochkin, 2019), or with interruptions presented using transient signals in modalities other than visual (e.g., Hameed et al., 2006; Ho et al., 2004; Jayaraman, 2011; Sarter, 2005), which ignores the need for interruption notifications to be persistent, such as when interruptions are negotiable. Our findings provide evidence that preattentive features can be effectively employed in the visual modality to improve performance managing negotiated interruptions in a busy multi-agent task environment without compromising performance on the ongoing or interrupting tasks.

Finally, the third theoretical contribution of this work is a better understanding of differences between participants who perform well and those who struggle with the management of interruptions. In this line of work, we have found evidence that poor performers are much more likely to reject tasks incorrectly and are less likely to remember task instructions, compared to top-performers. This indicates that operators working with multiple different types of tasks or those with poor working memory may benefit from having persistent and easy access to task instructions and better awareness of task progress, such as through the use of electronic checklists (e.g., Albuquerque et al., 2011; Myers III, 2016; Palmer & Degani, 1991). Additionally, implementation of decision support systems, like the visualization method presented in **Chapter 5**, may be particularly beneficial for some operators for preventing unintentional dismissal of necessary and urgent interruptions.

From an applied perspective, findings from this line of work will help reduce the attentional demands and improve the safety and performance of human-machine teams and the well-being of human operators in a variety of complex event-driven application domains. For example, in the medical domain, emergency rooms and intensive care units closely resemble single-operator-multi-agent operations. In these environments, front-line practitioners must manage multiple patients, each with a unique set of health conditions at varying levels of urgency and criticality. Nurses in particular are susceptible to experiencing nested interruptions (e.g., Sasangohar, 2015), such as when transitioning duties during shift changes, communicating with colleagues, and remembering requests made by doctors. Notably, interruption management performance in these task environments may suffer even more due to higher reliance on working memory, such as the need to remember tasks to be completed without a persistent visual interface, and due to limitations of which tasks can and cannot be interrupted to switch to another task (Sirihorachai et al., 2018). The work presented in this dissertation highlights potential stages during which performance may suffer as a result of handling such frequent and nested interruptions and provides empirically based guidance on ways to support practitioners safely and efficiently mitigate these challenges.

Beyond the medical domain and the supervision of autonomous aerial vehicles, visualization techniques developed in this line of work can complement human performance in other forms of computer-assisted and collaborative work. For example, as the use of virtual and augmented reality becomes more ubiquitous and accessible, priority visualizations can be an effective way to support the management of interrupting tasks that require mobility, such as emergency response and space operations.

6.2 Future Work

The work presented in this dissertation is a first step in understanding the challenges associated with handling negotiated interruptions that are increasingly frequent and potentially nested. Looking forward, as human-machine teams transition from one-to-one control to the one-to-many supervision of semi-autonomous agents, nested interruptions will likely become more prevalent both due to an increase in the frequency of interruptions, and due to unforeseen interactions between various parts of the system. There are several avenues that should be further explored in order to advance our understanding of, and ability to support operators in handling frequent and nested interruptions.

While there was some evidence indicating that more frequent and nested interruptions result in missed or delayed acknowledgement, we did not see major performance breakdowns at the detection stage in this line of work. Part of this may be because the only stimuli presented over the auditory channel were the auditory chimes for interrupting tasks. It is possible that if the auditory channel were subject to stimuli and noise from additional sources, performance decrements would be more prominent. Future studies should explore performance at the detection stage by employing experiment designs where the auditory channel is used for more than notification alerts, such as communication with other team members. In this dissertation, for example, it was assumed that a single operator is solely responsible for managing a set of aircraft from takeoff through landing, rather than multiple, specialized operators handling different aspects of flight operations for a group of aircraft. In the latter case, it would be necessary to share information and coordinate task responsibilities across multiple operators, thereby not only creating more interruptions but also increasing the number of sources of these interruptions. Such a paradigm would have implications at the detection stage due to the need to acknowledge requests made by other operators, and also at the interpretation and integration stages due to the need to balance requests from other operators in addition to requests from vehicles.

The current research necessarily employed a simple mapping of task characteristics to level of priority—three levels of urgency mapped to high, medium, and low priority. Practitioners in operational settings must often consider different or additional dimensions when making decisions about interruption handling, such as importance or difficulty and time required to perform an interrupting task. Consideration and weighing of these additional factors would greatly increase the effort needed to interpret and appropriately integrate tasks into the ongoing workflow. A task of high importance, for example, may not be highly urgent, and should be scheduled differently than a task that is both important and urgent (e.g., Cox et al., 2021; Lanctot & Duxbury, 2022). In turn, operators may be more likely to forget to return to deferred tasks that are important but not urgent. Or they may fail to resume interrupted tasks if the interrupting task requires considerable time to complete.

The amount of time available to complete a task is an important consideration in the assignment of tasks to human and machine agents. For example, emergency situations such as dealing with a battery fire may require a response within a matter of seconds, rather than the one-minute or longer time ranges available to participants in our experiments. In an emergency, it may be more appropriate for the task to be allocated to automation right away, which can then inform the operator once the task is complete. This raises two important questions: 1) on what basis should tasks be assigned to the human operator and the automation and who controls this allocation?, and 2) how can we keep the operator informed about automation actions without creating additional and potentially untimely interruptions for the human operator? These questions are partially addressed by literature on adaptive/adaptable function allocation (Boy, 2009; Chapanis, 1965; Feigh & Pritchett, 2014; Li et al., 2013) and levels of automation (Miller & Parasuraman, 2003; Prinet et al., 2012; Wickens et al., 2010) but they are far from being fully resolved.

Another important challenge for interruption management is that the perceived task urgency and importance often differs between the interrupter and interruptee—what may be highly important and urgent to the caller may be of low urgency to the callee. In these cases, interruption management solutions based on a fixed or static definition of task priority will be ineffective or, worse, counterproductive. Particularly in the context of human-machine teams, it is therefore critical for both agents to understand the goals, needs, and responsibilities of their counterpart to ensure the timely and appropriate initiation of an interruption. This level of understanding has traditionally been limited to human-human teams which share social norms, common sense, and similar reasoning. But the rapid pace of research and development in artificial intelligence may make it possible to establish a bi-directional understanding between human operators and machine agents—in essence, to have the machine know what the operator knows (Whang, 2023). This topic has received increased attention in recent years and is being discussed under the label 'mutual theory of mind' and mutually adaptive interaction (Brinck & Balkenius, 2020). One important issue is, for example, achieving mutual recognition of the expectations, intentions, and social affordances of another agent such that the human and robot become and act as a single system in which the behavior and actions of one depend on the other. Another requirement for technology to evolve into a collaborative partner is access to information about a human operator's baseline levels of performance on specific tasks (including accuracy and time needed to complete a task) to recognize instances when performance begins to suffer or falls below acceptable thresholds of safety (e.g., due to working long hours). Intelligent agents may combine this information with operator's preferred strategies for dealing with frequent and nested interruptions (e.g., only switch at specific breakpoints), as well as the criticality and urgency of pending tasks in order to make better decisions regarding whether and when the operator should be interrupted, and whether the operator will be able to switch to a pending task on time.

Finally, the role of task characteristics should be explored further. We examined but did not observe a significant effect of processing code similarity on task performance; however, other aspects of ongoing and interrupting tasks remain to be evaluated. These may include similarity in content between the interrupting and the ongoing task, the performance level of each task (i.e., skill-based, rule-based and knowledge-based performance) and the level of task engagement. For example, prospective memory may suffer more in case of ongoing tasks that are highly engaging or when these tasks require solving novel or unfamiliar problems. Also, the degree of predictability of interruptions deserves further exploration. Future studies may investigate, for example, how the interruption management process is influenced by frequent and nested interruptions when they occur after long versus short uninterrupted spans of time, and when transitions between low and high frequency interruptions are sudden versus gradual (e.g., Bowers et al., 2014; Moacdieh et al., 2020).

Appendices

Appendix A – Pre-Experiment Questionnaire for Experiment 1

Pre-Experiment Questionnaire

* Indicates required question

Section 1 of 3

Please enter your participant ID *

Short answer text

How often do you play video games? *

- ◯ Rarely or never
- ◯ Several times a year
- ◯ Several times a month
- ◯ Several times a week
- ◯ Daily

Section 2 of 3
Gaming Experience

	No	Sometimes	Yes
Role-playing Games (RPG)	◯	◯	◯
Real time strategy (RTS) or multiplayer online battle arena (MOBA). Examples include StarCraft, DOTA, etc.	◯	◯	◯
Simulation (flight simulation, city building, etc.)	◯	◯	◯

Puzzle (Chess, Sudoku, etc.)	○	○	○
Other	○	○	○

Please list the games you have experience with in the RPG genre:

Short answer text

Please list the games you have experience with in the RTS/MOBA genre:

Short answer text

Please list the games you have experience with in the simulation genre:

Short answer text

Please list any other games that you have experience with:

Short answer text

Section 3 of 3

Please indicate how much you agree with each of the following statements *

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I believe I have the ability to work effectively on more than one task at once	○	○	○	○	○
I believe I have the ability to do several things at once	○	○	○	○	○
I believe I have the ability to work on a number of different tasks at the same	○	○	○	○	○

time

I believe I have the ability to work on different types of tasks during the same time period	○	○	○	○	○
I believe I have the ability to go back and forth between multiple tasks	○	○	○	○	○

Please indicate how much you agree with each of the following statements *

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I believe I have the ability to work effectively even when I'm interrupted	○	○	○	○	○
I believe I have the ability to maintain my concentration even when interrupted by another task	○	○	○	○	○
I believe I have the ability to work effectively even when I'm frequently interrupted to do something else	○	○	○	○	○
I believe I have the ability to get back on track quickly after I've been interrupted while working	○	○	○	○	○

Submit

Clear form

Appendix B – Post-Scenario Questionnaire for Experiments 1-3

Post-Scenario Questionnaire

* Indicates required question

Please enter your participant ID *

Short answer text

How mentally demanding was the scenario? *

	1	2	3	4	5	6	7	8	9	10	
Low	○	○	○	○	○	○	○	○	○	○	High

How physically demanding was the scenario? *

	1	2	3	4	5	6	7	8	9	10	
Low	○	○	○	○	○	○	○	○	○	○	High

How hurried or rushed was the pace of the scenario? *

	1	2	3	4	5	6	7	8	9	10	
Low	○	○	○	○	○	○	○	○	○	○	High

How successful were you in accomplishing what you were asked to do? *

	1	2	3	4	5	6	7	8	9	10	
Poor	○	○	○	○	○	○	○	○	○	○	Good

How hard did you have to work (mentally and physically) to accomplish your level of performance? *

	1	2	3	4	5	6	7	8	9	10	
Low	○	○	○	○	○	○	○	○	○	○	High

How annoyed or frustrated were you? *

	1	2	3	4	5	6	7	8	9	10	
Low	○	○	○	○	○	○	○	○	○	○	High

Submit

Clear form

Appendix C – Post-Experiment Semi-Structured Interview Guide for Experiment 1

At the beginning of the post-experiment semi-structured interview, the researcher will start the audio recording. To help recall specific events or challenges during the interview, participants will be allowed to refer to a screen recording of the two scenarios performed by the participant.

The following questions will be used to guide the discussion:

- How do you think you did in managing the different tasks in the scenarios?
- On what basis do you decide whether or not to switch to the interrupting task?
- What stage/part of handling multiple interruptions presents a major challenge? Reliable detection of notifications, accurate interpretation of the task(s) to be performed, appropriate integration into ongoing workflow, or returning to the ongoing task?
    - Did any of these stages become more difficult when interruptions became more frequent, serial, and/or nested?
    - Can you recall specific instances during the scenarios that were challenging and why?
- What strategies did you use that helped make the interruption management process easier?
    - What helped you more easily return to the interrupted task?
    - Did these strategies change when interruptions became more frequent, serial, and/or nested?
- If you were given a chance to do the experiment again, what, if anything, would you do differently? Would you try a different strategy?
- What tools or interface changes would help you better manage the interruptions?

Appendix D – Pre-Experiment Questionnaire for Experiments 2 and 3

Pre-Experiment Questionnaire

* Indicates required question

Please enter your participant ID *

Short answer text

How often do you play video games? *

- ◯ Rarely or never
- ◯ Several times a year
- ◯ Several times a month
- ◯ Several times a week
- ◯ Daily

Please indicate how much you agree with each of the following statements *

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I believe I have the ability to work effectively on more than one task at once	◯	◯	◯	◯	◯
I believe I have the ability to do several things at once	◯	◯	◯	◯	◯
I believe I have the ability to shift my attention across multiple tasks when I	◯	◯	◯	◯	◯

work

I believe I have the ability to work on a number of different tasks at the same time	◯	◯	◯	◯	◯
I believe I have the ability to work on different types of tasks during the same time period	◯	◯	◯	◯	◯
I believe I have the ability to go back and forth between multiple tasks	◯	◯	◯	◯	◯

Please indicate how much you agree with each of the following statements *

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I believe I have the ability to work effectively even when I'm interrupted	◯	◯	◯	◯	◯
I believe I have the ability to maintain my concentration even when interrupted by another task	◯	◯	◯	◯	◯
I believe I have the ability to work effectively even when I'm frequently interrupted to do something else	◯	◯	◯	◯	◯
I believe I have the ability to get back on track quickly after I've been interrupted while working	◯	◯	◯	◯	◯

Submit

Clear form

Appendix E – Post-Training Questionnaire for Experiments 2 and 3

Post-Training Questionnaire

* Indicates required question

Please enter your participant ID *

Short answer text

What is the maximum number of points you can earn for high urgency tasks? *

- ○ 2
- ○ 3
- ○ 6
- ○ 8

Medium urgency tasks completed more than 2 minutes after notification are worth 0 points.*

- ○ True
- ○ False

If you are currently working on a new flight task with a cargo of vaccines and a notification appears that a UAV holding takeout orders needs help selecting an alternate landing site, the appropriate action to take is to *

- ○ Accept the incoming task as soon as possible
- ○ Delay the incoming task
- ○ Handoff the incoming task

How many bonus points can you get by completing a low urgency task within 2 minutes after notification, as long as the associated new flight task is not delayed? *

Short answer text

What is the maximum number of points you can earn for high urgency tasks? *

- ○ 2
- ○ 3
- ○ 6
- ○ 8

A new flight that is scheduled to carry milk & eggs (L) is delayed. What is the maximum number of points you can get for this new flight task? *

- ○ 6
- ○ 4
- ○ 3
- ○ 2
- ○ 0

Handing off a medium or low urgency task results in a negative score penalty.*

- ○ True
- ○ False

Failure to complete a high urgency task within how many minutes after notification results in a penalty of -2 points? *

- ○ 4
- ○ 3
- ○ 2
- ○ 1

Submit

Clear form

Appendix F – Post-Experiment Questionnaire for Experiment 2

Post-Experiment Questionnaire

* Indicates required question

Please enter your participant ID *

Short answer text

What part of handling multiple interruptions did you find to be a major challenge? Please rank the following from least (1) to most challenging (5).*

	1 (Least challenging)	2	3	4	5 (Most challenging)
Reliable detection of notifications	○	○	○	○	○
Determining the relative urgency of tasks	○	○	○	○	○
Remembering the rules for accepting, delaying, and handing off tasks based on relative urgency	○	○	○	○	○
Remembering to switch to an interrupting task	○	○	○	○	○
Returning to the previous task after completing an interrupting task	○	○	○	○	○

Please explain your answer to the previous question. *

Long answer text

Were there other aspects of interruption management you found challenging? Please explain. *

Long answer text

Submit

Clear form

This form was created inside of University of Michigan. Report Abuse

Google Forms

Appendix G – Post-Experiment Questionnaire for Experiment 3

Post-Experiment Questionnaire

* Indicates required question

Section 1 of 3

Please enter your participant ID *

Short answer text

Assigned Group *

- ○ Group 1
- ○ Group 2
- ○ Group 3

What part of handling multiple interruptions did you find to be a major challenge? Please rank the following from least (1) to most challenging (5). You may choose the same rating for more than one item. *

	1 (Least challenging)	2	3	4	5 (Most challenging)
Reliable detection of notifications	○	○	○	○	○
Determining the relative urgency of tasks	○	○	○	○	○
Remembering the rules for accepting, delaying, and handing off tasks based on relative urgency	○	○	○	○	○
Remembering to switch to an interrupting task	○	○	○	○	○

Returning to the previous task after completing an interrupting task	◯	◯	◯	◯	◯

Please explain your answer to the previous question. *

Long answer text

Section 2 of 3

Group 2

Did you find automatic sorting of notifications based on urgency to be helpful? *

	1	2	3	4	5	6	7	8	9	10	
Not at all helpful	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Extremely helpful

Please explain your answer to the previous question. *

Long answer text

Section 3 of 3

Group 3

Did you find the visualization of notification urgency helpful? *

	1	2	3	4	5	6	7	8	9	10	
Not at all helpful	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Extremely helpful

Did you find reminders to be helpful? Please leave blank if you did not see any reminders. *

1	2	3	4	5	6	7	8	9	10

Not at all helpful	○	○	○	○	○	○	○	○	○	○	Extremely helpful

Please explain your answer to the previous two questions. *

Long answer text

Submit

Clear form

This form was created inside of University of Michigan. Report Abuse

Google Forms

Bibliography

Adamczyk, P. D., & Bailey, B. P. (2004). If not now, when?: The effects of interruption at different moments within task execution. *Proceedings of the 2004 Conference on Human Factors in Computing Systems - CHI '04*, 271–278. https://doi.org/10/bq27cn

Adamczyk, P. D., Iqbal, S. T., & Bailey, B. P. (2005). A method, system, and tools for intelligent interruption management. *Proceedings of the 4th International Workshop on Task Models and Diagrams*, 123–126. https://doi.org/10/bkh3f3

Adler, R. F., & Benbunan-Fich, R. (2013). Self-interruptions in discretionary multitasking. *Computers in Human Behavior*, *29*(4), 1441–1449. https://doi.org/10/f4x6mm

Albuquerque, K. V., Miller, A. A., & Roeske, J. C. (2011). Implementation of Electronic Checklists in an Oncology Medical Record: Initial Clinical Experience. *Journal of Oncology Practice*, *7*(4), 222–226. https://doi.org/10.1200/JOP.2011.000237

Alqahtani, M., & Histon, J. M. (2012). Improving the Management of Interruption through the Working Awareness Interruption Tool: WAIT. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*, 393–397. https://doi.org/10.1177/1071181312561089

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive Science*, *26*(1), 39–83. https://doi.org/10/bcz8nm

Altmann, E. M., & Trafton, J. G. (2004). *Task interruption: Resumption lag and the role of cues*. *26*(26).

Altmann, E. M., Trafton, J. G., & Hambrick, D. Z. (2017). Effects of interruption length on procedural errors. *Journal of Experimental Psychology: Applied*, *23*(2), 216–229. https://doi.org/10/gbkm96

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, *6*(4), 451–474. https://doi.org/10.1016/0010-0285(74)90021-8

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An Integrated Theory of the Mind. *Psychological Review*, *111*(4), 1036–1060. https://doi.org/10/bvmxts

Andreasson, R., Lindblom, J., & Thorvald, P. (2017). Interruptions in the wild: Portraying the handling of interruptions in manufacturing from a distributed cognition lens. *Cognition, Technology & Work*, *19*(1), 85–108. https://doi.org/10/f3tz8v

Andrews, A., Ratwani, R., & Trafton, G. (2009). Recovering from Interruptions: Does Alert Type Matter? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *53*(4), 409–413. https://doi.org/10/fxx472

Baethge, A., Rigotti, T., & Roe, R. A. (2015). Just more of the same, or different? An integrative theoretical framework for the study of cumulative interruptions at work. *European Journal of Work and Organizational Psychology*, *24*(2), 308–323. https://doi.org/10/gjf9xd

Bailey, B. P., & Iqbal, S. T. (2008). Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction*, *14*(4), 21:1-21:28. https://doi.org/10/dtpgjc

Bailey, B. P., & Konstan, J. A. (2006). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior*, *22*(4), 685–708. https://doi.org/10/fbckbg

Bailey, B. P., Konstan, J. A., & Carlis, J. V. (2000). Measuring the effects of interruptions on task performance in the user interface. *Smc 2000 Conference Proceedings. 2000 Ieee International Conference on Systems, Man and Cybernetics. "cybernetics Evolving to Systems, Humans, Organizations, and Their Complex Interactions" (Cat. No.0*, *2*, 757–762 vol.2. https://doi.org/10/c69tmb

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779. https://doi.org/10/fhdj8w

Barrera-Leon, L., Corno, F., & De Russis, L. (2023). How the Preattentive Process is Exploited in Practical Information Visualization Design: A Review. *International Journal of Human–Computer Interaction*, *39*(4), 707–720. https://doi.org/10.1080/10447318.2022.2049137

Basoglu, K. A., Fuller, M. A., & Sweeney, J. T. (2009). Investigating the effects of computer mediated interruptions: An analysis of task characteristics and interruption frequency on financial performance. *International Journal of Accounting Information Systems*, *10*(4), 177–189. https://doi.org/10/cqvbjj

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Berg, P. T. van den, Roe, R. A., Zijlstra, F. R. H., & Krediet, I. (1996). Temperamental Factors in the Execution of Interrupted Editing Tasks. *European Journal of Personality*, *10*(4), 233–248. https://doi.org/10/fxgfhh

Bogunovich, P., & Salvucci, D. (2011). The effects of time constraints on user behavior for deferrable interruptions. *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11*, 3123. https://doi.org/10/bm9wng

Boot, W. R., Kramer, A. F., Becic, E., Wiegmann, D. A., & Kubose, T. (2006). Detecting Transient Changes in Dynamic Displays: The More You Look, the Less You See. *Human Factors*, *48*(4), 759–773. https://doi.org/10.1518/001872006779166424

Borowsky, A., Horrey, W. J., Liang, Y., Garabet, A., Simmons, L., & Fisher, D. L. (2016). The effects of brief visual interruption tasks on drivers' ability to resume their visual search for a pre-cued hazard. *Accident Analysis and Prevention*, *93*, 207–216. https://doi.org/10/f8vp66

Borst, J. P., Taatgen, N. A., & van Rijn, H. (2010). The problem state: A cognitive bottleneck in multitasking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 363–382. https://doi.org/10/bwwtz4

Borst, J. P., Taatgen, N. A., & van Rijn, H. (2015). What Makes Interruptions Disruptive? A Process-Model Account of the Effects of the Problem State Bottleneck on Task Interruption and Resumption. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2971–2980. https://doi.org/10/gjf9xx

Botvinick, M. M., & Bylsma, L. M. (2005). Distraction and action slips in an everyday task: Evidence for a dynamic representation of task context. *Psychonomic Bulletin & Review*, *12*(6), 1011–1017. https://doi.org/10.3758/BF03206436

Bowers, M. A., Christensen, J. C., & Eggemeier, F. T. (2014). The Effects of Workload Transitions in a Multitasking Environment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 220–224. https://doi.org/10.1177/1541931214581046

Boy, G. A. (2009). The Orchestra: A Conceptual Model for Function Allocation and Scenario-based Engineering in Multi-agent Safety-critical Systems. In L. Norros, H. Koskinen, L. Salo, & P. Savioja (Eds.), *Ecce 2009—European Conference on Cognitive Ergonomics: Designing Beyond the Product—Understanding Activity and User Experience in Ubiquitous Environments* (Vol. 258, pp. 187–193). Technical Research Centre Finland.

Braarud, P. Ø., & Kirwan, B. (2011). Task Complexity: What Challenges the Crew and How Do They Cope. In A. B. Skjerve & A. Bye (Eds.), *Simulator-based Human Factors Studies Across 25 Years* (pp. 233–251). Springer. https://doi.org/10.1007/978-0-85729-003-8_15

Brazzolotto, P., Duran, G., & Michael, G. A. (2022). How do we handle interruptions? Investigating the processes underlying the resumption of interrupted tasks. *Psychologie Française*, *67*(1), 77–95. https://doi.org/10.1016/j.psfr.2021.05.001

Brazzolotto, P., & Michael, G. A. (2020). Complexity of interruptions: Evidence supporting a non-interruption-based theory. *Scandinavian Journal of Psychology*, *61*(n/a), 723–730. https://doi.org/10/gg3dxt

Brewer, R. W., Walker, A. J., Pursel, E. R., Cerame, E. J., Baker, A. L., & Schaefer, K. E. (2020). Assessment of Manned-Unmanned Team Performance: Comprehensive After-Action Review Technology Development. In J. Chen (Ed.), *Advances in Human Factors in Robots and Unmanned Systems* (pp. 119–130). Springer International Publishing.

Brinck, I., & Balkenius, C. (2020). Mutual Recognition in Human-Robot Interaction: A Deflationary Account. *Philosophy & Technology*, *33*(1), 53–70. https://doi.org/10.1007/s13347-018-0339-x

Brixey, J. J., Robinson, D. J., Johnson, C. W., Johnson, T. R., Turley, J. P., & Zhang, J. (2007). A Concept Analysis of the Phenomenon Interruption. *Advances in Nursing Science*, *30*(1), E26.

Broadbent, D. E., & Broadbent, M. H. P. (1987). From detection to identification: Response to multiple targets in rapid serial visual presentation. *Perception & Psychophysics*, *42*(2), 105–113. https://doi.org/10.3758/BF03210498

Brudzinski, M. E., Ratwani, R. M., & Trafton, J. G. (2007). *Goal and Spatial Memory Following Interruption*. 6.

Brüning, J., Mückstein, M., & Manzey, D. (2020). Multitasking strategies make the difference: Separating processing-code resources boosts multitasking efficiency when individuals prefer to interleave tasks in free concurrent dual tasking. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(12), 1411–1433. https://doi.org/10.1037/xhp0000865

Cades, D. M., Davis, D. A. B., Trafton, J. G., & Monk, C. A. (2007). Does the Difficulty of an Interruption Affect our Ability to Resume? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *51*(4), 234–238. https://doi.org/10/fxvd9f

Cain, B. (2007). *A Review of the Mental Workload Literature*. DEFENCE RESEARCH AND DEVELOPMENT TORONTO (CANADA). https://apps.dtic.mil/sti/citations/ADA474193

Chapanis, A. (1965). On the Allocation of Functions between Men and Machines. *Occupational Psychology*, *39*(1), 1–11. bth.

Chaparro, A., & Chaparro, M. (2017). Applications of Color in Design for Color-Deficient Users. *Ergonomics in Design*, *25*(1), 23–30. https://doi.org/10.1177/1064804616635382

Chih-Yuan Ho, Nikolic, M. I., & Sarter, N. B. (2001). Multimodal information presentation in support of timesharing and effective interruption management. *20th DASC. 20th Digital Avionics Systems Conference (Cat. No.01CH37219)*, *1*, 5D2/1-5D2/8 vol.1. https://doi.org/10/cvpbb5

Cook, G. I., Marsh, R. L., & Hicks, J. L. (2005). Associating a Time-Based Prospective Memory Task with an Expected Context Can Improve or Impair Intention Completion. *Applied Cognitive Psychology*, *19*, 345–360. https://doi.org/10.1002/acp.1082

Couffe, C., & Michael, G. A. (2017). Failures Due to Interruptions or Distractions: A Review and a New Framework. *The American Journal of Psychology*, *130*(2), 163–181. JSTOR. https://doi.org/10/gg2d2f

Cox, A. L., Bird, J., Brumby, D. P., Cecchinato, M. E., & Gould, S. J. J. (2021). Prioritizing unread e-mails: People send urgent responses before important or short ones. *Human–Computer Interaction*, *36*(5–6), 511–534. https://doi.org/10.1080/07370024.2020.1835481

Cutrell, E., Czerwinski, M., & Horvitz, E. (2001). *Notification, Disruption, and Memory: Effects of Messaging Interruptions on Memory and Performance*. IFIP TC13 International Conference on Human-Computer Interaction. https://www.semanticscholar.org/paper/Notification%2C-Disruption%2C-and-Memory%3A-Effects-of-on-Cutrell-Czerwinski/aa17d9337793fb6676b4f9aad55374ca64a88159

Czerwinski, M., Chrisman, S., & Schumacher, B. (1991). Interactive Posters: The Effects of Warnings and Display Similarity on Interruption in Multitasking Environments. *ACM SIGCHI Bulletin*, *23*(4), 38–39. https://doi.org/10.1145/126729.1056014

Czerwinski, M., Cutrell, E., & Horvitz, E. (2000). Instant Messaging: Effects of Relevance and Timing. *People and Computers XIV: Proceedings of HCI 2000*, *2*, 71–76. https://www.microsoft.com/en-us/research/publication/instant-messaging-effects-of-relevance-and-timing/

Dell'Acqua, R., Jolicœur, P., Sessa, P., & Turatto, M. (2006). Attentional blink and selection in the tactile domain. *European Journal of Cognitive Psychology*, *18*(4), 537–559. https://doi.org/10.1080/09541440500423186

Dewitt, M. R., Hicks, J. L., Ball, B. H., & Knight, J. B. (2012). Encountering items previously paired with prospective memory target events can serve to reactivate intentions. *Journal of Cognitive Psychology*, *24*(8), 981–990. https://doi.org/10.1080/20445911.2012.727389

Di Lollo, V., Kawahara, J., Shahab Ghorashi, S. M., & Enns, J. T. (2005). The attentional blink: Resource depletion or temporary loss of control? *Psychological Research*, *69*(3), 191–200. https://doi.org/10.1007/s00426-004-0173-x

Dismukes, R. K. (2012). Prospective Memory in Workplace and Everyday Situations. *Current Directions in Psychological Science*, *21*(4), 215–220. https://doi.org/10/f4p6pc

Dismukes, R., & Nowinski, J. (2007). Prospective memory, concurrent task management, and pilot error. In *Attention: From theory to practice*. Oxford University Press New York.

Dixon, S. R., & Wickens, C. D. (2006). Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(3), 474–486. https://doi.org/10/b4n4xt

Dodhia, R. M., & Dismukes, R. K. (2009). Interruptions create prospective memory tasks. *Applied Cognitive Psychology*, *23*(1), 73–89. https://doi.org/10.1002/acp.1441

Dörner, D. (1980). On the Difficulties People Have in Dealing With Complexity. *Simulation & Games*, *11*(1), 87–106. https://doi.org/10/c2khs9

Drews, F. A., Markewitz, B. A., Stoddard, G. J., & Samore, M. H. (2019). Interruptions and Delivery of Care in the Intensive Care Unit. *Human Factors*, *61*(4), 564–576. https://doi.org/10/gjf9xb

Duchowski, A. T. (2017). *Eye tracking methodology: Theory and practice* (Third edition). Springer.

Edwards, M. B., & Gronlund, S. D. (1998). Task Interruption and its Effects on Memory. *Memory*, *6*(6), 665–687. https://doi.org/10/bjzfd2

Egeth, H. E., Folk, C. L., Leber, A. B., Nakama, T., & Hendel, S. K. (2001). Attentional Capture in the Spatial and Temporal Domains. In *Attraction, Distraction and Action* (Vol. 133, pp. 93–119). North-Holland. https://doi.org/10.1016/S0166-4115(01)80007-8

Einstein, G. O., McDaniel, M. A., Manzi, M., Cochran, B., & Baker, M. (2000). Prospective memory and aging: Forgetting intentions over short delays. *Psychology and Aging*, *15*(4), 671–683. https://doi.org/10.1037/0882-7974.15.4.671

Einstein, G. O., Williford, C. L., Pagan, J. L., McDaniel, M. A., & Dismukes, R. K. (2003). Forgetting of Intentions in Demanding Situations Is Rapid. *Journal of Experimental Psychology. Applied*, *9*(3), 147–162. https://doi.org/10/c2h26c

Eyrolle, H., & Cellier, J.-M. (2000). The effects of interruptions in work activity: Field and laboratory results. *Applied Ergonomics*, *31*(5), 537–543. https://doi.org/10.1016/S0003-6870(00)00019-3

FAA. (2020). *Urban Air Mobility Concept of Operations* (1.0). Federal Aviation Administration. https://nari.arc.nasa.gov/sites/default/files/attachments/UAM_ConOps_v1.0.pdf

Feigh, K. M., & Pritchett, A. R. (2014). Requirements for Effective Function Allocation: A Critical Review. *Journal of Cognitive Engineering and Decision Making*, *8*(1), 23–32. https://doi.org/10/ggxpxq

Feitshans, G., Rowe, A., Davis, J., Holland, M., & Berger, L. (2008, August 18). Vigilant Spirit Control Station (VSCS): The Face of COUNTER. *AIAA Guidance, Navigation and Control Conference and Exhibit*. https://doi.org/10.2514/6.2008-6309

Ferris, T., Penfold, R., Hameed, S., & Sarter, N. (2006). The Implications of Crossmodal Links in Attention for the Design of Multimodal Interfaces: A Driving Simulation Study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(3), 406–409. https://doi.org/10.1177/154193120605000341

Folk, C. L., & Remington, R. (2006). Top-down modulation of preattentive processing: Testing the recovery account of contingent capture. *Visual Cognition*, *14*(4–8), 445–465. https://doi.org/10.1080/13506280500193545

Fong, A., Hettinger, A. Z., & Ratwani, R. M. (2017). A Predictive Model of Emergency Physician Task Resumption Following Interruptions. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2405–2410. https://doi.org/10.1145/3025453.3025700

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A Description of Think Aloud Method and Protocol Analysis. *Qualitative Health Research*, *3*(4), 430–441. https://doi.org/10/fj64tr

Foroughi, C. K., Malihi, P., & Boehm-Davis, D. A. (2016). Working Memory Capacity and Errors Following Interruptions. *Journal of Applied Research in Memory and Cognition*, *5*(4), 410–414. https://doi.org/10/gjf9xz

Foroughi, C. K., Werner, N. E., McKendrick, R., Cades, D. M., & Boehm-Davis, D. A. (2016). Individual differences in working-memory capacity and task resumption following interruptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9), 1480–1488. https://doi.org/10.1037/xlm0000251

Friedrich, M., & Vollrath, M. (2022). Urgency-Based color coding to support visual search in displays for supervisory control of multiple unmanned aircraft systems. *Displays*, *74*, 102185. https://doi.org/10.1016/j.displa.2022.102185

Gazzaley, A., Cooney, J. W., Rissman, J., & D'Esposito, M. (2005). Top-down suppression deficit underlies working memory impairment in normal aging. *Nature Neuroscience*, *8*(10), 1298–1300. https://doi.org/10.1038/nn1543

Gibson, B. S., & Peterson, M. A. (2001). Inattentional Blindness and Attentional Capture: Evidence for Attention-Based Theories of Visual Salience. In *Attraction, Distraction and Action* (Vol. 133).

Gillie, T., & Broadbent, D. (1989). What makes interruptions disruptive? A study of length, similarity, and complexity. *Psychological Research*, *50*(4), 243–250. https://doi.org/10/btzgdn

Gold, C., Damböck, D., Lorenz, L., & Bengler, K. (2013). “Take over!” How long does it take to get the driver back into the loop? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 1938–1942. https://doi.org/10/gfkv89

Gray, W. D., & Fu, W.-T. (2004). Soft constraints in interactive behavior: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science*, *28*(3), 359–382. https://doi.org/10.1207/s15516709cog2803_3

Greitzer, F. L. (2005). Toward the development of cognitive task difficulty metrics to support intelligence analysis research. *Fourth IEEE Conference on Cognitive Informatics, 2005. (ICCI 2005).*, 315–320. https://doi.org/10.1109/COGINF.2005.1532647

Guo, J., Chen, T., Xie, Z., & Or, C. K. (2021). Effects of interventions to reduce the negative consequences of interruptions on task performance: A systematic review, meta-analysis, and narrative synthesis of laboratory studies. *Applied Ergonomics*, *97*, 103506. https://doi.org/10.1016/j.apergo.2021.103506

Hameed, S., Ferris, T. K., Jayaraman, S., & Sarter, N. B. (2006). Supporting Interruption Management Through Informative Tactile and Peripheral Visual Cues. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(3), 376–380. https://doi.org/10/fzkzx8

Hicks, J. L., Cook, G. I., & Marsh, R. L. (2005). Detecting event-based prospective memory cues occurring within and outside the focus of attention. *The American Journal of Psychology*, *118*(1), 1–11.

Hill, B. P., DeCarme, D., Metcalfe, M., Griffin, C., Wiggins, S., Metts, C., Bastedo, B., Patterson, M. D., & Mendonca, N. L. (2020). *UAM Vision Concept of Operations (ConOps) UAM Maturity Level (UML) 4*. https://ntrs.nasa.gov/citations/20205011091

Hirsch, P., Koch, I., & Grundgeiger, T. (2022). Task Interruptions. In A. Kiesel, L. Johannsen, I. Koch, & H. Müller (Eds.), *Handbook of Human Multitasking* (pp. 145–188). Springer International Publishing. https://doi.org/10.1007/978-3-031-04760-2_4

Ho, C.-Y., Nikolic, M. I., Waters, M. J., & Sarter, N. B. (2004). Not Now! Supporting Interruption Management by Indicating the Modality and Urgency of Pending Tasks. *Human Factors*, *46*(3), 399–409. https://doi.org/10/cqfp5m

Hodgetts, H. M., & Jones, D. M. (2006a). Contextual cues aid recovery from interruption: The role of associative activation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1120–1132. https://doi.org/10/ftn3qk

Hodgetts, H. M., & Jones, D. M. (2006b). Interruption of the Tower of London Task: Support for a Goal-Activation Approach. *Journal of Experimental Psychology: General*, *135*(1), 103–115. https://doi.org/10/c5xcn4

Holden, J., & Goel, N. (2016). *Fast-Forwarding to a Future of On-Demand Urban Air Transportation* (pp. 1–98). Uber Elevate.

Horvitz, E., Apacible, J., & Subramani, M. (2005). Balancing Awareness and Interruption: Investigation of Notification Deferral Policies. In L. Ardissono, P. Brna, & A. Mitrovic (Eds.), *User Modeling 2005* (pp. 433–437). Springer. https://doi.org/10.1007/11527886_59

Iqbal, S. T., & Bailey, B. P. (2005). Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, 1489–1492. https://doi.org/10/btjr6f

Iqbal, S. T., & Horvitz, E. (2007). Disruption and recovery of computing tasks: Field study, analysis, and directions. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 677–686. https://doi.org/10/bf89jz

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. https://doi.org/10.1109/34.730558

Jayaraman, S. (2011). *Supporting Monitoring and Interruption Management in Complex Domains through Graded Multimodal Notifications.*

Jett, Q. R., & George, J. M. (2003). Work Interrupted: A Closer Look at the Role of Interruptions in Organizational Life. *Academy of Management Review*, *28*(3), 494–507. https://doi.org/10/b3dqcp

John, M., Smallman, H. S., & Manes, D. I. (2005). Recovery from Interruptions to a Dynamic Monitoring Task: The Beguiling Utility of Instant Replay. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*(3), 473–477. https://doi.org/10/fzch78

Jones, M. R. (2001). Temporal expectancies, capture, and timing in auditory sequences. In *Attraction, Distraction and Action* (Vol. 133). Elsevier Science.

Klimoski, R., & Mohammed, S. (1994). Team Mental Model: Construct or Metaphor? *JOURNAL OF MANAGEMENT*, *20*(2).

Kupfer, C., Prassl, R., Fleiß, J., Malin, C., Thalmann, S., & Kubicek, B. (2023). Check the box! How to deal with automation bias in AI-based personnel selection. *Frontiers in Psychology*, *14*. https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1118723

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Laarni, J. (2021). Multitasking and interruption handling in control room operator work. In A.-M. Teperi & N. Gotcheva (Eds.), *Human Factors in the Nuclear Industry* (pp. 127–149). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-102845-2.00007-7

Labonté, K., Tremblay, S., & Vachon, F. (2019). Forewarning interruptions in dynamic settings: Can prevention bolster recovery? *Journal of Experimental Psychology: Applied*, *25*(4), 674–694. https://doi.org/10/ggwkfx

Lanctot, A., & Duxbury, L. (2022). It may be urgent, but is it important? A look at how employees evaluate their emails. *Canadian Journal of Administrative Sciences / Revue Canadienne Des Sciences de l'Administration*, *40*(1), 18–32. https://doi.org/10.1002/cjas.1672

Lascara, B., Lacher, A., Matthew, D., Maroney, D., Niles, R., & Vempati, L. (2019). *Urban Air Mobility Airspace Integration Concepts*.

Latorella, K. A. (1997). *Investigating interruptions: Implications for flightdeck performance* [Ph.D., State University of New York at Buffalo]. http://search.proquest.com/docview/304388806/abstract/3CEE3154FAD94AA2PQ/1

Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(5), 802–817. https://doi.org/10.1109/TPAMI.2006.86

Lee, B. C., & Duffy, V. G. (2015). The Effects of Task Interruption on Human Performance: A Study of the Systematic Classification of Human Behavior and Interruption Frequency. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *25*(2), 137–152. https://doi.org/10/gg2dx8

Li, H., Sarter, N. B., Wickens, C., & Sebok, A. (2013). Supporting human-automation collaboration through dynamic function allocation: The case of space teleoperation. *Proceedings of the Human Factors and Ergonomics Society*, 359–363. https://doi.org/10/ggxpws

Li, S. Y. W., Magrabi, F., & Coiera, E. (2012). A systematic review of the psychological literature on interruption and its patient safety implications. *Journal of the American Medical Informatics Association*, *19*(1), 6–12. https://doi.org/10/dtjjtw

Liu, P., & Li, Z. (2012). Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics*, *42*(6), 553–568. https://doi.org/10.1016/j.ergon.2012.09.001

Longo, L., Wickens, C., Hancock, P., & Hancock, G. (2022). Human Mental Workload: A Survey and a Novel Inclusive Definition. *Frontiers in Psychology*, *13*, 883321. https://doi.org/10.3389/fpsyg.2022.883321

Lu, S. A., Wickens, C. D., Prinet, J. C., Hutchins, S. D., Sarter, N., & Sebok, A. (2013). Supporting Interruption Management and Multimodal Interface Design: Three Meta-Analyses of Task Performance as a Function of Interrupting Task Modality. *Human Factors*, *55*(4), 697–724. https://doi.org/10/ggxpvz

Lu, S. A., Wickens, C. D., Sarter, N. B., & Sebok, A. (2011). Informing the design of multimodal displays: A meta-analysis of empirical studies comparing auditory and tactile interruptions. *Proceedings of the Human Factors and Ergonomics Society*, *2*(MacLean 2008), 1170–1174. https://doi.org/10/fzhxj4

Magrabi, F., Li, S. Y. W., Day, R. O., & Coiera, E. (2010). Errors and electronic prescribing: A controlled laboratory study to examine task complexity and interruption effects. *Journal of the American Medical Informatics Association*, *17*(5), 575–583. https://doi.org/10.1136/jamia.2009.001719

Maltz, M., & Meyer, J. (2001). Use of Warnings in an Attentionally Demanding Detection Task. *Human Factors*, *43*(2), 217–226. https://doi.org/10.1518/001872001775900931

Mansi, G., & Levy, Y. (2013). Do instant messaging interruptions help or hinder knowledge workers' task performance? *International Journal of Information Management*, *33*(3), 591–596. https://doi.org/10.1016/j.ijinfomgt.2013.01.011

Mark, G., Gudith, D., & Klocke, U. (2008). The cost of interrupted work: More speed and stress. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 107–110. https://doi.org/10/cjftns

Marsh, R. L., Cook, G. I., & Hicks, J. L. (2006). An Analysis of Prospective Memory. In *Psychology of Learning and Motivation* (Vol. 46, pp. 115–153). Academic Press. https://doi.org/10.1016/S0079-7421(06)46004-7

Martens, S., & Wyble, B. (2010). The attentional blink: Past, present, and future of a blind spot in perceptual awareness. *Neuroscience & Biobehavioral Reviews*, *34*(6), 947–957. https://doi.org/10.1016/j.neubiorev.2009.12.005

Mathur, A., Panesar, K., Kim, J., Atkins, E. M., & Sarter, N. (2019, June 14). Paths to Autonomous Vehicle Operations for Urban Air Mobility. *AIAA Aviation 2019 Forum*. https://doi.org/10/ggxpx8

McComb, S., Kennedy, D., Perryman, R., Warner, N., & Letsky, M. (2010). Temporal Patterns of Mental Model Convergence: Implications for Distributed Teams Interacting in Electronic Collaboration Spaces. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(2), 264–281. https://doi.org/10.1177/0018720810370458

McDaniel, M. A., Einstein, G. O., Graham, T., & Rall, E. (2004). Delaying execution of intentions: Overcoming the costs of interruptions. *Applied Cognitive Psychology*, *18*(5), 533–547. https://doi.org/10/dxdp9f

McDaniel, M. A., Robinson-Riegler, B., & Einstein, G. O. (1998). Prospective remembering: Perceptually driven or conceptually driven processes? *Memory & Cognition*, *26*(1), 121–134. https://doi.org/10.3758/BF03211375

McFarlane, D. C. (1997). *Interruption of People in Human-Computer Interaction: A General Unifying Definition of Human Interruption and Taxonomy* (NRL-FR-5510-97-9870). OFFICE OF NAVAL RESEARCH ARLINGTON VA. https://apps.dtic.mil/docs/citations/ADA333587

McFarlane, D. C. (2002). Comparison of Four Primary Methods for Coordinating the Interruption of People in Human-Computer Interaction. *Human–Computer Interaction*, *17*(1), 63–139. https://doi.org/10/bjkwg9

McFarlane, D. C., & Latorella, K. A. (2002). The Scope and Importance of Human Interruption in Human-Computer Interaction Design. *Human–Computer Interaction*, *17*(1), 1–61. https://doi.org/10/fv8w62

Miller, C. A., & Parasuraman, R. (2003). Beyond Levels of Automation: An Architecture for More Flexible Human-Automation Collaboration. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *47*, 182–186. https://doi.org/10/fx8z7b

Miyata, Y., & Norman, D. A. (1986). Psychological issues in support of multiple activities. *User Centered System Design: New Perspectives on Human-Computer Interaction*, 265–284. https://doi.org/10/gjf9x2

Moacdieh, N. M., Devlin, S. P., Jundi, H., & Riggs, S. L. (2020). Effects of Workload and Workload Transitions on Attention Allocation in a Dual-Task Environment: Evidence From Eye Tracking Metrics. *Journal of Cognitive Engineering and Decision Making*, 1555343419892184. https://doi.org/10/ggxpvt

Mok, B., Johns, M., Lee, K. J., Miller, D., Sirkin, D., Ive, P., & Ju, W. (2015). Emergency, Automation Off: Unstructured Transition Timing for Distracted Drivers of Automated Vehicles. *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, 2458–2464. https://doi.org/10.1109/ITSC.2015.396

Monk, C. A. (2004). The Effect of Frequent versus Infrequent Interruptions on Primary Task Resumption. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(3), 295–299. https://doi.org/10/fx33jp

Monk, C. A., Boehm-Davis, D. A., Mason, G., & Trafton, J. G. (2004). Recovering From Interruptions: Implications for Driver Distraction Research. *Human Factors*, *46*(4), 650–663. https://doi.org/10/chcssj

Monk, C. A., & Kidd, D. G. (2008). *The effects of brief interruptions on task resumption*. *52*(4), 403–407.

Monk, C. A., Trafton, J. G., & Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied*, *14*(4), 299–313. https://doi.org/10/dwjvdr

Monsell, S., Sumner, P., & Waters, H. (2003). Task-set reconfiguration with predictable and unpredictable task switches. *Memory & Cognition*, *31*(3), 327–342. https://doi.org/10.3758/BF03194391

Morgan, P. L., Patrick, J., & Tiley, L. (2013). Improving the effectiveness of an interruption lag by inducing a memory-based strategy. *Acta Psychologica*, *142*(1), 87–95. https://doi.org/10.1016/j.actpsy.2012.09.003

Mueller, E. R., Kopardekar, P. H., & Goodrich, K. H. (2017). Enabling Airspace Integration for High-Density On-Demand Mobility Operations. *17th AIAA Aviation Technology, Integration, and Operations Conference*, *June*, 1–24. https://doi.org/10/ggxpz7

Myers III, P. (2016). Commercial Aircraft Electronic Checklists: Benefits and Challenges (Literature Review). *International Journal of Aviation, Aeronautics, and Aerospace*. https://doi.org/10.15394/ijaaa.2016.1112

National Academies of Sciences, Engineering, and Medicine. (2020). *Advanced Aerial Mobility: A National Blueprint* (p. 25646). National Academies Press. https://doi.org/10.17226/25646

Nikolic, M. I., Orr, J. M., & Sarter, N. B. (2004). Why Pilots Miss the Green Box: How Display Context Undermines Attention Capture. *The International Journal of Aviation Psychology*, *14*(1), 39–52. https://doi.org/10/ffmxmf

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

Palmer, E., & Degani, A. (1991). Electronic Checklists: Evaluation of Two Levels of Automation. *Proceedings of the Sixth International Aviation Psychology Symposium*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.461.1825&rep=rep1&type=pdf

Panesar, K., Mathur, A., Atkins, E., & Sarter, N. (2021). Moving From Piloted to Autonomous Operations: Investigating Human Factors Challenges in Urban Air Mobility. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*, 241–245. https://doi.org/10.1177/1071181321651143

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10/bbntx8

Patterson, M. D., Antcliff, K. R., & Kohlman, L. W. (2018, May 14). *A Proposed Approach to Studying Urban Air Mobility Missions Including an Initial Exploration of Mission Requirements*. Annual Forum and Technology Display, Phoenix, AZ. https://ntrs.nasa.gov/citations/20190000991

Penningroth, S. L., & Scott, W. D. (2013). Task Importance Effects on Prospective Memory Strategy Use. *Applied Cognitive Psychology*, *27*(5), 655–662. https://doi.org/10.1002/acp.2945

Perrow, C. (1999). *Normal accidents: Living with high-risk technologies*. Princeton University Press.

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*(2), 160–174. https://doi.org/10.1037/0096-3445.109.2.160

Prinet, J. C., Terhune, A., & Sarter, N. B. (2012). Supporting dynamic re-planning in multiple UAV control: A comparison of 3 levels of automation. *Proceedings of the Human Factors and Ergonomics Society*, 423–427. https://doi.org/10/ggxpwt

Proctor, R. W., & Proctor, J. D. (2021). Sensation and Perception. In *HANDBOOK OF HUMAN FACTORS AND ERGONOMICS* (pp. 55–90). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119636113.ch3

Puranik, H., Koopman, J., & Vough, H. C. (2020). Pardon the Interruption: An Integrative Review and Future Research Agenda for Research on Work Interruptions. *Journal of Management*, *46*(6), 806–842. https://doi.org/10.1177/0149206319887428

Radović, T., & Manzey, D. (2022). Effects of complexity and similarity of an interruption task on resilience toward interruptions in a procedural task with sequential constraints. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(2), 159–173. https://doi.org/10.1037/xhp0000981

Ratwani, R. M., Andrews, A. E., McCurry, M., Trafton, J. G., & Peterson, M. S. (2007). Using Peripheral Processing and Spatial Memory to Facilitate Task Resumption. *Proceedings of* *the Human Factors and Ergonomics Society Annual Meeting*, *51*(4), 244–248. https://doi.org/10/fzgkmz

Ratwani, R. M., Andrews, A. E., Sousk, J. D., & Trafton, J. G. (2008). The Effect of Interruption Modality on Primary Task Resumption. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(4), 393–397. https://doi.org/10/fzjstz

Ratwani, R. M., & Trafton, J. G. (2004). *A Spatial Memory Mechanism for Guiding Primary Task Resumption*. George Mason University.

Ratwani, R. M., & Trafton, J. G. (2008). Spatial memory guides task resumption. *Visual Cognition*, *16*(8), 1001–1010. https://doi.org/10/bs5k5z

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 849–860. https://doi.org/10.1037/0096-1523.18.3.849

Recarte, M. Á., Pérez, E., Conchillo, Á., & Nunes, L. M. (2008). Mental Workload and Visual Impairment: Differences between Pupil, Blink, and Subjective Rating. *The Spanish Journal of Psychology*, *11*(2), 374–385. https://doi.org/10.1017/S1138741600004406

Riggs, S. L., Wickens, C. D., Sarter, N., Thomas, L. C., Nikolic, M. I., & Sebok, A. (2017). Multimodal Information Presentation in Support of NextGen Operations. *The International Journal of Aerospace Psychology*, *27*(1–2), 29–43. https://doi.org/10/ggxpzc

Rios, D., Katzman, N., Burdick, K. J., Gellert, M., Klein, J., Bitan, Y., & Schlesinger, J. J. (2023). Multisensory alarm to benefit alarm identification and decrease workload: A feasibility study. *Journal of Clinical Monitoring and Computing*, 1–9. https://doi.org/10.1007/s10877-023-01014-4

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207–231. APA PsycArticles. https://doi.org/10/b2qmtj

Rummel, J., Wesslein, A.-K., & Meiser, T. (2017). The role of action coordination for prospective memory: Task-interruption demands affect intention realization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(5), 717–735. https://doi.org/10.1037/xlm0000334

Saleh, H. H. (2019). *Optimal Policy for Nested Task Interruptions in a Workflow Process* [Thesis, Purdue University Graduate School]. https://doi.org/10.25394/PGS.7430654.v1

Salvucci, D. D., & Beltowska, J. (2008). Effects of Memory Rehearsal on Driver Performance: Experiment and Theoretical Account. *Human Factors*, *50*(5), 834–844. https://doi.org/10.1518/001872008X354200

Salvucci, D. D., & Bogunovich, P. (2010). Multitasking and monotasking: The effects of mental workload on deferred task interruptions. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 85–88. https://doi.org/10/c6smn7

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review*, *115*(1), 101–130. https://doi.org/10/c8fhv8

Salvucci, D. D., & Taatgen, N. A. (2011a). *The multitasking mind*. Oxford University Press.

Salvucci, D. D., & Taatgen, N. A. (2011b). Toward a Unified View of Cognitive Control. *Topics in Cognitive Science*, *3*(2), 227–230. https://doi.org/10/cggznt

Salvucci, D. D., Taatgen, N. A., & Borst, J. P. (2009). Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1819–1828. https://doi.org/10.1145/1518701.1518981

Santomauro, C., Powell, M., Davis, C., Liu, D., Aitken, L. M., & Sanderson, P. (2021). Interruptions to Intensive Care Nurses and Clinical Errors and Procedural Failures: A Controlled Study of Causal Connection. *Journal of Patient Safety*, *17*(8), e1433. https://doi.org/10.1097/PTS.0000000000000528

Sarter, N. B. (2005). Graded and Multimodal Interruption Cueing in Support of Preattentive Reference and Attention Management. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*(3), 478–481. https://doi.org/10/fzkh9j

Sarter, N. B. (2006). Multimodal information presentation: Design guidance and research challenges. *International Journal of Industrial Ergonomics*, *36*(5), 439–445. https://doi.org/10/dxc5jg

Sarter, N. B. (2007). Multiple-resource theory as a basis for multimodal interface design: Success stories, qualifications, and research needs. *Attention: From Theory to Practice*, 187–195.

Sarter, N. B. (2013). Multimodal Support for Interruption Management: Models, Empirical Findings, and Design Recommendations. *Proceedings of the IEEE*, *101*(9), 2105–2112. https://doi.org/10/f5bvff

Sasangohar, F. (2015). *Understanding and Mitigating the Interruptions Experienced by Intensive Care Unit Nurses* [Thesis]. https://tspace.library.utoronto.ca/handle/1807/71317

Sasangohar, F., Donmez, B., Easty, A. C., & Trbovich, P. L. (2017). Effects of Nested Interruptions on Task Resumption: A Laboratory Study With Intensive Care Nurses. *Human Factors*, *59*(4), 628–639. https://doi.org/10/gbhdvs

Sasangohar, F., Scott, S. D., & Cummings, M. L. (2014). Supervisory-level interruption recovery in time-critical control tasks. *Applied Ergonomics*. https://doi.org/10/gg2d8x

Sasangohar, F., Scott, S. D., & Donmez, B. (2013). Interruption Management and Recovery in Time-critical Supervisory-level Tasks: A Literature Review. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 1745–1749. https://doi.org/10/gg5hcs

Schaper, P., & Grundgeiger, T. (2018). The effect of different distractions on remembering delayed intentions. *Memory*, *26*(2), 154–170. https://doi.org/10.1080/09658211.2017.1339090

Schrank, D., Eisele, B., & Lomax, T. (2019). *Urban Mobility Report*. Texas A&M Transportation Institute. http://tti.tamu.edu/documents/mobility-report-2019.pdf

Schuler, P. T., & Yang, X. J. (2022). Human operators' blind compliance, reliance, and dependence behaviors when working with imperfect automation: A meta-analysis. *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*, 1–6. https://doi.org/10.1109/ICHMS56717.2022.9980609

Scott, N., Wan, J., Farzan Sasangohar, & Mary Cummings. (2008). Mitigating Supervisory-level Interruptions in Mission Control Operations. *Proceedings of 2nd Int'l Conference on Applied Human Factors and Ergonomics*. https://interruptions.net/literature/Scott-AHFEI08.pdf

Scott, S. D., Mercier, S., Cummings, M. L., & Wang, E. (2006). Assisting Interruption Recovery in Supervisory Control of Multiple Uavs. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(5), 699–703. https://doi.org/10/fxt429

Shen, Y. J., & Jiang, Y. V. (2006). Interrupted visual searches reveal volatile search memory. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1208–1220. https://doi.org/10/c6wvn9

Sheridan, T. B., & Parasuraman, R. (2000). Human vs. Automation in Responding to Failures: An Expected-Value Analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*(13), 1–4. https://doi.org/10.1177/154193120004401301

Sirihorachai, R., Aebersold, M. L., Sarter, N. B., Harris, M., Marsh, V., & Redman, R. W. (2018). Examining Interruptions in the Operating Room Using Simulation. *Clinical Simulation in Nursing*, *24*, 14–24. https://doi.org/10/gfmfs3

Sklar, A. E., & Sarter, N. B. (1999). Good Vibrations: Tactile Feedback in Support of Attention Allocation and Human-Automation Coordination in Event-Driven Domains. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *41*(4), 543–552. https://doi.org/10.1518/001872099779656716

Smith, C. A. P., Clegg, B. A., Heggestad, E. D., & Hopp-Levine, P. J. (2009). Interruption management: A comparison of auditory and tactile cues for both alerting and orienting. *International Journal of Human-Computer Studies*, *67*(9), 777–786. https://doi.org/10.1016/j.ijhcs.2009.05.002

Speier, C., Valacich, J. S., & Vessey, I. (1999). The Influence of Task Interruption on Individual Decision Making: An Information Overload Perspective. *Decision Sciences*, *30*(2), 337–360. https://doi.org/10/fdfp5q

Speier, C., Vessey, I., & Valacich, J. S. (2003). The Effects of Interruptions, Task Complexity, and Information Presentation on Computer-Supported Decision-Making Performance. *Decision Sciences*, *34*(4), 771–797. https://doi.org/10/b6jnpv

Steelman, K. S., McCarley, J. S., & Wickens, C. D. (2011). Modeling the Control of Attention in Visual Workspaces. *Human Factors*, *53*(2), 142–153. https://doi.org/10/fnq6w8

Szumowska, E., & Kossowska, M. (2017). Motivational rigidity enhances multitasking performance: The role of handling interruptions. *Personality and Individual Differences*, *106*, 81–89. https://doi.org/10.1016/j.paid.2016.10.040

Thomas, L., Donohue-Porter, P., & Stein Fishbein, J. (2017). Impact of Interruptions, Distractions, and Cognitive Load on Procedure Failures and Medication Administration Errors. *Journal of Nursing Care Quality*, *32*(4), 309. https://doi.org/10.1097/NCQ.0000000000000256

Trafton, J. G., Altmann, E. M., Brock, D. P., & Mintz, F. E. (2003). Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies*, *58*(5), 583–603. https://doi.org/10/cpm46c

Trbovich, P., Prakash, V., Stewart, J., Trip, K., & Savage, P. (2010). Interruptions During the Delivery of High-Risk Medications. *JONA: The Journal of Nursing Administration*, *40*(5), 211–218. https://doi.org/10/cxdwxx

Treisman, A., & Souther, J. (1985). Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, *114*(3), 285–310. https://doi.org/10.1037/0096-3445.114.3.285

van den Haak, M., De Jong, M., & Jan Schellens, P. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology*, *22*(5), 339–351. https://doi.org/10.1080/0044929031000

van Erp, J. B. (2007). *Tactile displays for navigation and orientation: Perception and behaviour*. Utrecht University.

Wan, Y. (2019). *Identifying and Overcoming Attention Limitations in the Detection and Identification of Alarms in Close Temporal Proximity* [PhD Thesis].

Westbrook, J. I., Woods, A., Rob, M. I., Dunsmuir, W. T. M., & Day, R. O. (2010). Association of Interruptions With an Increased Risk and Severity of Medication Administration Errors. *Archives of Internal Medicine*, *170*(8), 683–690. https://doi.org/10/d56pw8

Whang, O. (2023, March 27). Can a Machine Know That We Know What It Knows? *The New York Times*. https://www.nytimes.com/2023/03/27/science/ai-machine-learning-chatbots.html

Wiberg, M., & Whittaker, S. (2005). Managing availability: Supporting lightweight negotiations to handle interruptions. *ACM Transactions on Computer-Human Interaction*, *12*(4), 356–387. https://doi.org/10/d2q3dj

Wickens, C., & Colcombe, A. (2007). Effects of CDTI Alerting System Properties on Pilot Multi-Task Performance. *2007 International Symposium on Aviation Psychology*, 739–742.

Wickens, C. D. (1980). The Structure of Attentional Resources. In *Attention and performance VIII* (Vol. 8, pp. 239–257). Psychology Press.

Wickens, C. D. (2014). Effort in Human Factors Performance and Decision Making. *Human Factors*, *56*(8), 1329–1336. https://doi.org/10.1177/0018720814558419

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering psychology and human performance* (Fourth edition). Pearson.

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010). Stages and Levels of Automation: An Integrated Meta-analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(4), 389–393. https://doi.org/10/fx3xxb

Wickens, C. D., & McCarley, J. S. (2019). *Applied Attention Theory*. CRC Press. https://doi.org/10.1201/9780429059261

Wickens, C. D., McCarley, J. S., & Gutzwiller, R. S. (2022). Concurrent Task Performance: Time-Sharing and Multiple Resources. In *Applied Attention Theory* (2nd ed.). CRC Press.

Wickens, C., McCarley, J., & Steelman-Allen, K. (2009). NT-SEEV: A model of attention capture and noticing on the Flight Deck. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *53*(12), 769–773. https://doi.org/10/fzh6vz

Williams, C., Morgan, P. L., Christopher, G., Zook, N., & Hoskins, R. (2020). The Effects of Clinical Task Interruptions on Subsequent Performance of a Medication Pre-Administration Task. In T. Ahram & C. Falcão (Eds.), *Advances in Usability and User Experience* (pp. 81–92). Springer International Publishing. https://doi.org/10.1007/978-3-030-19135-1_9

Wilson, M. D., Farrell, S., Visser, T. A. W., & Loft, S. (2018). Remembering to execute deferred tasks in simulated air traffic control: The impact of interruptions. *Journal of Experimental Psychology: Applied*, *24*(3), 360–379. https://doi.org/10/gd7s37

Wilson, M. D., Strickland, L., Farrell, S., Visser, T. A. W., & Loft, S. (2020). Prospective Memory Performance in Simulated Air Traffic Control: Robust to Interruptions but Impaired by Retention Interval. *Human Factors*, *62*(8), 1249–1264. https://doi.org/10.1177/0018720819875888

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, *5*(6), 495–501. https://doi.org/10.1038/nrn1411

Wolfe, J. M., & Utochkin, I. S. (2019). What is a preattentive feature? *Current Opinion in Psychology*, *29*, 19–26. https://doi.org/10.1016/j.copsyc.2018.11.005

Woods, D. D. (1995). The alarm problem and directed attention in dynamic fault management. *Ergonomics*, *38*(11), 2371–2393. https://doi.org/10/dsp2qb

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*(5), 459–482. https://doi.org/10/cfnwhw

Zijlstra, F. R. H., Roe, R. A., Leonora, A. B., & Krediet, I. (1999). Temporal factors in mental work: Effects of interrupted activities. *Journal of Occupational and Organizational Psychology*, *72*(2), 163–185. https://doi.org/10.1348/096317999166581