Department of Information Engineering
and Computer Science

Master's Degree in
Human-Computer Interaction and Design

Final Dissertation

Exploring Pilot Workload Scenarios via Eye-Tracking: An Attempt at Inducing and Identifying Attentional Tunneling in the Cockpit

University supervisor	External tutor	Student, MAT. 239019
Luca Turchet	Maik Friedrich	Elena Rankova

Academic year 2022/2023

Acknowledgements

Throughout this thesis, I have been fortunate enough to always feel supported and receive guidance from exceptional professionals. I would like to take the opportunity here to thank everyone, who helped me along the way. First of all, I would like to thank Luca Turchet for his supervision and for giving me the freedom to pursue this topic. Furthermore, I am deeply grateful to Maik Friedrich for his unwavering support, for the time and effort he invested, and for believing in my capabilities. This experiment brought me a great amount of new skills and knowledge and for all this I have Maik to thank to as he offered me possibilities, guidance, and support that not many tutors demonstrate.

In addition, I extend my gratitude to DLR as an institution for providing me with the necessary materials, equipment, and financial resources, and to all my colleagues for all their support. I am grateful to David for developing the simulator scenarios, to Julia for recruiting the participants, to Matthias for sharing his knowledge and helping us develop and test the experimental design, to Anneke for her valuable input on statistics and study design, and to Boris and Jakob for their insightful recommendations on topics related to machine learning. Moreover, I am thankful to Vicky and Emiel for introducing me to statistical approaches that I was previously not familiar with and for helping me cope with difficulties related to my data.

Finally, I am grateful to my wonderful family and Hido, for being there and for simply being as sweet as they are. Special thanks to my sister for hyping me up and keeping things light, and to Jens for intervening during furious disputes in the WG.

Abstract

Given the profound impact of human errors and the essential role of operators in safety-critical domains, ensuring that operators are in a condition that allows them to adequately perform their tasks is a vital precaution. The timely identification of hazardous cognitive states can reduce accidents and enhance safety across various fields, including aviation. As workload and attentional tunneling are among the cognitive states most frequently associated with human error accidents in aviation, the purpose of this thesis is to explore the possibility of detecting these states using eye-tracking metrics. Attentional tunneling, a term commonly referenced in accident reports, is characterized by the excessive focus on a source of information, hypothesis, or goal to the disregard of other factors. Although previous research has demonstrated the recognition of workload in cockpit settings using eye-tracking metrics, attentional tunneling in simulator environments has rarely been explored. With this study, our aim was to propose and analyze scenarios for inducing and detecting attentional tunneling in simulator environments and to investigate the efficiency of transition frequency, mean saccade length, and entropy as a set of eye-tracking metrics for classifying workload and tunneling states.

As tunneling triggering parameters, the proposed experiment design incorporated a workload-inducing secondary task and an ego-threatening factor in the form of negative auditory feedback on a focus task. Consequently, the occurrence of attentional tunneling was determined based on participants' ability to notice visual cues related to abnormal cockpit behavior. This experimental framework was tested by 15 expert pilots, with data from 12 participants included in the eye-tracking and attentional tunneling analysis. Findings from the workload self-assessment measurements indicated the successful manipulation of workload between conditions. Moreover, the occurrence of attentional tunneling could be observed across one-third of the runs, suggesting that the proposed scenarios have proven efficient.

The statistical analysis of the eye-tracking measurements revealed a significant decrease in the transition frequency and mean saccade length during high workload conditions. The occurrence of attentional tunneling, however, did not seem to significantly impact the recorded gaze measurements. Using the eye-tracking data, three machine-learning pipelines, including Support Vector Machines, Logistic Regression, and Bernoulli Naive Bayes, were trained and tested on their performance across two different classification problems: differentiating between low and high workload states and recognizing instances of attentional tunneling. With mean scores of approximately 50% for both accuracy and precision across all machine-learning approaches, the outcomes of the workload classification did not reach satisfactory performance. Similarly, the effectiveness of the logistic regression and SVM pipelines in classifying tunneling states showcased suboptimal results and a strong bias with relatively high accuracy mean scores and exceptionally low precision scores. Nevertheless, compared to the other two algorithms, the Bernoulli Naive Bayes demonstrated promising results that can be further investigated in future studies focusing on tunneling classification.

Although the employed pipelines were unable to effectively classify the different cognitive states, the lessons learned have been instrumental in developing a strategy for subsequent improvements to our approach, mainly focused on data exploration and restructuring.

Contents

List of Figures

List of Tables

Glossary

ANFIS Adaptive-Network-Based Fuzzy Inference System. 19, 53

AOI Areas of Interest. 16, 18, 19, 24, 30, 40, 50

AR Augmented Reality. 9, 13

ATC Air Traffic Controllers. 22

ATG Attention Getter Panel. 4, 25

CFIT Controlled Flight Into Terrain. 8

ECAM Electronic Centralised Aircraft Monitoring. 4, 24–26, 36, 47, 48, 54

EFIS Electronic Flight Instrument System. 4, 25

EyeTA Eye Tracking Analyser. 27

FCU Flight Control Unit. 4, 25

HUD Head-Up Displays. 9, 13, 22

ILS Instrument Landing System. 4, 24–26, 36, 47–49, 54

ISA Instantaneous Self-Assessment of Workload Technique. 4, 28, 29, 31, 32, 34, 36, 45, 46, 50, 51, 53–55

kNN k-Nearest Neighbor. 19, 53

LDA Linear Discriminant Analysis. 19

LOC-I Loss of Control In-Flight. 8

MATB Multi-Attribute Task Battery. 12, 14

MCDU Multi Purpose & Display Unit. 4, 25

MLP Multilayer Perceptron. 19, 44

NASA-TLX NASA Task Load Index. 12, 13, 19, 28, 31–36, 45–47, 54

ND Navigation Display. 4, 25

NNI Nearest Neighbor Index. 17, 51

OVHD Overhead Panel. 4, 25

PASAT Paced Auditory Serial Addition Test. 13

PFD Primary Flight Display. 4, 12, 14, 23, 25, 30, 47, 50

RA1 Radio Altimeter 1. 4, 24–26, 36, 37, 47, 48, 54

RBF Radial Basis Function. 43, 52, 55

SA Situational awareness. 10

SMI SensoMotoric Instruments. 27

SVM Support Vector Machines. 1, 19, 42–44, 52, 53, 55

SVS Synthetic Vision Systems. 13, 22

SWAT Subjective Workload Assessment Technique. 28

1 Introduction

1.1 Context

An operator's cognitive state is of central importance to the management of safety-critical systems, such as an aircraft. Suboptimal states, such as fatigue and inadequate situational awareness, have been recognized as primary contributors to human-caused errors [1]. With human factors as a leading cause of aviation incidents, advancements in this field are crucial to aviation safety[2].

Workload and attentional tunneling are acknowledged as factors with a strong impact on a pilot's mental state and situational awareness due to their influence on perception and the understanding of a situation [3]. Safety reports frequently mention these states in relation to the most fatal accidents – Loss of Control In-flight (LOC-I) and Controlled Flight Into Terrain (CFIT) [4], [5]. Moreover, studies have identified attentional tunneling as a contributing factor to the majority of military CFIT crashes in the United States [6].

An example frequently employed to describe attentional tunneling in the cockpit is the Everglades 407 flight in 1974, during which a breakdown in a landing gear light caused pilots to focus on identifying the cause of the issue while neglecting to monitor the state of the aircraft. As a result, they failed to notice the disengagement of the autopilot, which led to a constant slow descent ultimately culminating in a fatal crash [7]. Recent accidents, whose investigation has underlined attentional tunneling as a crucial factor include the Singapore Airlines Flight 006 in 2000 [8], West Air Sweden Cargo Flight 294 in 2016 [9], and Tatarstan Airlines Flight 363 in 2013[10]. Each of these cases depicts a situation, in which an unforeseen event imposed an overwhelming workload on the pilots, leading to a state of attentional narrowing that hampered their ability to respond effectively.

Given the central role of pilots' cognitive states within human-caused accidents, the objective identification of a pilot's mental state could strongly contribute to aviation safety. The timely classification of hazardous states could, for example, be operationalized in designing adaptive interfaces that support pilots during safety-critical situations [11]. Moreover, it can be utilized during the testing phase of newly developed interfaces, thereby aiding in the recognition of edge cases and the development of suitable design solutions.

Previous studies have applied various physiological measures, including EEG, heart rate, and eye-tracking, to objectively identify a pilot's cognitive state [12], [13]. Among these measures, eye-tracking stands out as a promising minimally intrusive metric well-suited for the cockpit environment. It has been successfully utilized in classifying mental states, including situational awareness, attention and distraction, workload, attentional tunneling, and fatigue.

This thesis strives to advance research on cognitive state classification via eye-tracking by proposing and evaluating an approach employing machine learning algorithms. The classification and identification of workload and attentional tunneling will be the focus of this study due to the strong correlation between these factors and their significance in decision-making and situational awareness. Moreover, we aim to support future advancements in the research of attentional tunneling among pilots by proposing and analyzing an experimental design applicable in simulator environments. In an effort to achieve these objectives, the study aims to collect data modeled to explore the following research questions:

- Can high, low workload, and attentional tunneling be induced in experimental conditions?
- Can machine learning algorithms classify pilots' workload states based on their transition frequency, mean saccade length, and gaze entropy?
- Can machine learning algorithms classify the occurrence and absence of attentional tunneling among pilots based on their transition frequency, mean saccade length, and gaze entropy?

1.2 Contribution

Due to its relevance in safety-critical environments, workload classification has been the subject of various studies in areas such as aviation, air traffic control, road traffic, surgery, and plant monitoring. The feasibility of identifying high workload using ocular data has been extensively investigated, showing statistically significant outcomes highlighting the relationship between ocular behavior and workload.

Research investigating ocular behavior and its variance under different cognitive states has also focused on attentional tunneling as a topic. The relationship between tunneling and workload is evident in existing research, as the few studies specifically investigating attentional tunneling typically utilize high workload as an independent variable triggering tunneling states [14], [11]. Similarly, multiple studies on workload classification have reported cases of attentional narrowing and indications in their eye-tracking data [15], [16]. However, existing research has focused on either workload classification or attentional tunneling identification. This thesis aims to address this research gap by investigating both conditions within the same context and expanding knowledge of the similarities and differences between the two states by exploring the same eye-tracking metrics for both classification problems. To achieve this, a combination of measurements will be employed that have previously not been applied together but have demonstrated statistical significance in studies on either workload or tunneling classification.

A further objective of this study is the development and analysis of an experimental framework for inducing states of attentional narrowing and high workload in a simulator environment. Due to the constraints of simulators, the specific nature of tasks during a flight, and the visual scanning strategies learned by pilots during their training, a scarcity of experiments proposing experimental designs that potentially trigger states of tunneling in simulator settings exists. The majority of studies have reported attentional narrowing as a result of experiments primarily focused on workload [17] or the evaluation of new technologies, such as Head-Up Displays (HUD) [18] or Augmented Reality (AR) [19]. By developing and analyzing an experimental scenario specifically designed to induce and detect states of attentional narrowing in a cockpit setting, this study hopes to contribute to the knowledge within this domain and to encourage future work in the field.

2 Background

2.1 Situational Awareness, Workload, and Attentional Tunneling

Observing workload and attentional tunneling in a context, that incorporates both aspects, such as situational awareness, effectively illustrates the significance of the two states in safety-critical systems and underlines the inherent connection between the two states.

Within aviation, situational awareness (SA) is commonly described as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future." [3]. Consequently, SA can be considered a state of constant knowledge acquisition and the subsequent adaption of future expectations as a reaction to a dynamically changing environment.

This SA theory separates the process of obtaining awareness into three phases – Level 1: Perception of the Elements in the Environment, Level 2: Comprehension of the Current Situation, and Level 3: Projection of Future Status. During the first level, situational awareness is centered on the perception of critical elements, their attributes, and their current status. At Level 2 the isolated pieces of information about each element are integrated into a comprehensive picture of the situation. Finally, during the last phase, the knowledge obtained from previous levels is applied to build predictions of the future states of the perceived elements [3].

In the process of attaining situational awareness, high workload is considered a stressor that can lead to a narrowing of the perceived information. Such a deficiency at the perceptual level can impair one's ability to maintain SA. While various factors, apart from high workload, can contribute to a diminished situational awareness, such as fatigue or boredom, which are typical for low workload states [3], research frequently highlights findings indicating a decrease in information acquisition during high workload states [20], [15], [21]. The reduced perception of information as a reaction to high workload can result in both increased performance through effective prioritization and attention management, but it can also be the cause of tunneling states and a decline in situational awareness.

While workload is regarded as a stressor in the context of SA, definitions of workload commonly relate the concept to the limited mental capabilities of humans [22]. Correspondingly, Wickens's cognitive resource theory defines workload as the demand produced by one or multiple tasks on an individual's limited capabilities [23].

Complex dynamic systems, such as an aircraft's cockpit, can be especially demanding for mental capabilities, as they allow the simultaneous occurrence of multiple events. This often results in competition for an operator's attention and overlapping demand for processing resources. For example, in situations requiring an operator to read an error message while simultaneously listening to instructions, a decline in performance is expected [23]. During high workload situations, an individual's mental capacities are challenged by the demand for attention incoming from multiple sources. This can potentially lead to a decrease in perceptual capacities, resulting in a diminished Level 1 situational awareness.

The phenomenon of prioritization of certain tasks or information sources as a reaction to a high workload exceeding cognitive capacities can be interpreted as both focused attention and attentional tunneling. However, focused attention describes an intentional concentration on relevant tasks and associated stimuli while neglecting irrelevant information [14]. For an aircraft operator, focused attention is a valuable skill demonstrating effective attention management. It is a central part of pilot training and is reinforced by procedures like the sterile flight deck operation, stating that during critical flight phases, such as landing and take-off, safety-critical activities must be prioritized, while non-essential conversations and non-safety-related announcements are prohibited [24].

Attentional tunneling, on the other hand, represents an involuntary narrowing of attention, which, due to its uncontrollable nature, can be potentially dangerous. States of attentional narrowing are

characterized by the individual's lack of awareness of their narrowed perception as they assess their understanding and mental model of a situation to be complete, allowing uninformed or altered decision-making.

The concept of attentional tunneling as a state, defined by the excessive focus on one information source or goal at the expense of others, exists within current literature under various terms, including attentional tunneling, attentional narrowing, cognitive tunneling, tunnel vision, perseveration syndrome, inattentional blindness, and change blindness. The various terminology describes similar phenomena and is often used interchangeably or to describe closely related but slightly differing cognitive states. The inconsistent usage of the terms makes the research for measurements, an exhaustive definition, experiment design, and treatments especially complex [25]. A definition for this phenomenon commonly accepted within the aviation sector was proposed by Wickens: "the allocation of attention to a particular channel of information, diagnostic hypothesis or task goal, for a duration that is longer than optimal, given the expected costs of neglecting events on other channels, failing to consider other hypotheses, or failing to perform other tasks" [18]. Throughout the remainder of this thesis, the term attentional tunneling will be employed based on Wicken's definition. Following this definition, it can be interpreted that attentional narrowing can be observed whenever an individual persists in directing their focus on a task, information source, or an assumption for an unreasonably long period, causing the neglect of other tasks of potentially critical importance.

2.2 Low and High Workload in Experimental Conditions

Taking into account the previously described definitions of workload, one can conclude that with adjustments in the difficulty or quantity of tasks that require the engagement of similar cognitive resources, the mental capacities of an individual can be challenged to an overwhelming extent ultimately leading to an increased workload. Studies in the aviation field, typically induce different levels of workload by incorporating tasks with increasing difficulty (e.g., landing versus level-flight), by manipulating the complexity of individual tasks (e.g., manual flight, low visibility, or turbulence), or by including supplementary tasks (e.g., arithmetic tasks or audio secondary-task assignments) [26].

Within eye-tracking experiments, it is relevant to avoid factors that could potentially alter participants' eye behavior. Therefore researchers often preserve the same primary task between trials to avoid task-related changes in the ocular data. Instead, workload is often adjusted by incorporating a secondary task [27], [28], [29], or by manipulating the difficulty level of the primary task without changing it fundamentally. This can be achieved by introducing secondary factors like wind, turbulence, or visibility [30], [13]. As workload is one of the approaches utilized for inducing tunneling in experimental conditions, more detailed information regarding strategies for the manipulation of task difficulty and the implementation of secondary tasks are presented in the next section. There, they will be observed as workload-enhancing approaches that can be applied as tunneling triggers.

2.3 Attentional Tunneling in Experimental Conditions

To better understand tunneling and factors that could potentially serve as contributors to such states among pilots, accidents whose investigation has indicated attentional narrowing among the causes will be analyzed as a first step.

Aircraft accident reports suggest that some conditions used to increase workload in experimental settings have also been identified as contributing factors to pilot error and the consequent incidents. The Singapore Airlines Flight 006 in 2000 is an example of pilots overly focusing on a threat, that caused them to neglect the effective monitoring of their surroundings. Due to rapidly degrading weather and the threat of an incoming thunderstorm, the attention of the pilots was preoccupied with preparing for the dangerous wind conditions and the avoidance of the thunderstorm [8]. This stressful situation in combination with poor communication with the air-traffic controllers led the aircraft operators to neglect monitoring the outside conditions to the extent that they failed to notice they entered the wrong runway. As a result, the vehicle crashed at a construction site, causing 83 fatalities. A further tragic example underlining the relevance of attentional tunneling in cockpit settings is the case of the Air France Flight 447 in 2009 [31], where extreme environmental conditions in the form of

turbulence and loud noises caused by both an incoming hail and the aircraft's audio warning system were identified as factors with detrimental effect on the pilots' cognitive state. As a result of their decreased situational awareness and the occurring tunneling states, both pilots failed to correctly interpret an aircraft issue and led the aircraft into a stall causing the plane to crash into the ocean. Similarly, during the West Air Sweden Cargo Flight 294 in 2016 the startle effect originating from an unexpected behavior of one of the pilots' Primary Flight Display's (PFD) caused the operators to overfocus on their PFD displays while neglecting other instruments [9]. Due to the low light conditions during this night flight, the only source of orientation for the pilots was in their instruments. However, as a result of the occurring tunneling states, causing the pilots to disregard relevant information, their loss of orientation ultimately ended in a fatal crash.

Inducing attentional tunneling by mirroring situations from existing accidents in experimental conditions has been considered unfeasible, due to the complex nature of accident causation, commonly referred to as the Swiss cheese model, combined with the technical limitations of a simulator. As a result, existing literature has been examined for experimental conditions with a potential impact on workload and attentional narrowing. It is relevant to note, that due to the limited research on tunneling states, particularly within the field of aviation, the literature research incorporated all previously mentioned terms, used to depict attentional tunneling or states closely resembling it.

The structure of the following text has been based on Prinet's proposal of a comprehensive framework of tunneling triggers [14]. Prinet's model relates states of attentional narrowing to stress and anxiety. By utilizing the existing structure as a basis and adjusting it through simplifications or additions, the proposed categorization of this thesis includes environmental factors, workload, motivational intensity, and novel situations.

Environmental Factors

Stress is commonly understood as a response to the perception of a situation or stimulus as threatening, and has frequently been utilized as a trigger evoking states of cognitive narrowing. Many environmental conditions, when experienced at a high intensity, are perceived as threatening and have been utilized as stressors in previous research. For example, excessively hot or cold temperatures, vibrations, loud noises, and bright lights constitute some of the environmental factors commonly used in experimental conditions [32].

During civil aviation flights, bright light, loud noises, and vibrations in the form of turbulence are commonly experienced situations. As a result, some of these factors have been recreated in experimental conditions exploring their effects on participants' cognitive states. For example, previous studies employing loud noises as a threatening factor discovered that it has a small to medium negative effect on performance along the cognitive, motor, and communicational levels [33]. However, its impact on visual perception appeared to be negligible. In line with these findings, Pirnet's work indicated no significant effect of loud noise on the performance of the monitoring and detection tasks, incorporated in the commonly used multi-tasking environment Multi-Attribute Task Battery (MATB) [14]. Furthermore, the study did not report a significant effect of loud noises on the number of fixations, mean fixation duration, and mean saccade length as eye-tracking metrics. In a study comparing novice and expert pilots, loud noise in the form of an audio warning has been employed as a workload-inducing factor [34]. Its impact on the reported NASA-TLX scores has shown to be significant among pilots with limited training. However, its effect on experienced pilots was observed to be relatively lower. In their literature research paper, Szalma and Hancock examine the different effects that noise characteristics, such as type, schedule, intensity, and duration, have on different levels of mental performance [33]. A relevant observation in this study states that speech-related noise demonstrates a stronger effect compared to non-speech sounds. Furthermore, performance was closely related to the temporal patterns in which sounds were presented. A strong decline in performance was observed in relation to intermittent noises, the effect of which has been associated with the contrast in noise intensity resulting from the sudden sound outbursts. Interestingly short durations of noise exposure showcased stronger detrimental effects on performance in comparison to longer intervals of sound.

In comparison to the detailed examinations of the effects of noise on cognitive capacities, turbulence has been less frequently employed as a stressor. It has been applied as a workload-inducing factor,

assuming a moderate impact on perceived cognitive load [30]. As a moderately demanding workload condition, turbulence has shown a tendency with no significant effect on NASA-TLX scores.

Workload and Information Density

The majority of studies focusing on attentional tunneling in aviation have reported the phenomenon as a result of experiments with innovative interfaces containing high informational density, such as Augmented Reality (AR), Head-Up Displays (HUD), Synthetic Vision Systems (SVS), and Moving Weather Maps. The technology behind both HUD displays and AR presents digital information on top of real-world content. The implementation and research of this technology has indicated an excessive focus on the digital layers, combined with a reduced perception of the underlying real-world information [35], [19]. Similarly, studies incorporating highly visual displays, with a large amount of information, high level of details, and moving objects, such as the three-dimensional SVS interfaces and the dynamic Moving Map displays have reported pilots directing an inappropriate amount of attention on these instruments, while neglecting out-the-window information [18]. The attractiveness of the representation of digital elements has been assumed to be the cause behind this behavior of inappropriate attention allocation.

Workload and Task Difficulty

As previously mentioned, a common approach to manipulating workload states among pilots is the adjustment of task difficulty between scenarios. Different methods for controlling task difficulty within a simulator include weather manipulation, automation levels, unexpected situations, and traffic volume.

In an experiment on inattentional deafness, Dehais et al. evaluated pilots' ability to detect an alarm under high versus low workload conditions [13]. By introducing an unexpected strong wind change, commonly known as windshear, the experimenters successfully increased the experienced workload. This was reflected in the significant changes in both the reported subjective workload self-assessments and the objective heart rate measurements.

Similarly, weather conditions in combination with aircraft failures, traffic deviations, and communication issues have been applied as an approach to manipulating task complexity in an experiment on crew decision-making by Young et al. [17]. This study, although not specifically focused on attentional tunneling, reported multiple cases of aircraft crews failing to notice intruder aircrafts, runway incursions, and aircraft mode changes, ultimately indicating states of cognitive narrowing.

Another experiment, with a focus on attentional tunneling among pilots, evaluated the detection rates of runway incursions during automatic, partially automatic, and manual flight modes [36]. Although the subjective NASA-TLX measurements indicated a significantly higher workload during manual versus automatic flight modes, the detection of incursions between the two scenarios showcased no significant difference. Interestingly, during the partially automated scenarios pilots' detection rates increased significantly.

Workload and Multitasking

Multitasking is an inherent part of daily activities involving safety-critical tasks, such as driving while engaging in a conversation or processing information from air traffic controllers during a flight. Due to its relevance for safety, multitasking has been widely employed in experimental studies in the field of transportation. In the automotive industry, for example, auditory tasks like n-back and Paced Auditory Serial Addition Test (PASAT), have been applied as a workload-inducing factor in both studies on workload [37] and attentional tunneling [21]. Similarly, since multitasking and processing audio information is essential for pilots' capabilities, secondary audio tasks are considered an appropriate approach for manipulating workload in simulator environments [12], [38].

An example of a secondary task, commonly utilized in neuroscientific studies, is the n-back. It has been designed to overload participants' working memory capacities, by incorporating tasks related to monitoring and updating of information, and rule-based decision-making [39]. During an n-back task, participants are presented with a row of numbers and are asked to respond, whenever the current number matches the number n-steps back. Arithmetic audio tasks are another example of

workload-inducing approaches, frequently used in experimental studies [37], [14]. During such tasks, participants hear a series of numbers and are required to perform simple mathematical operations, basing their calculation on their current estimation. However, compared to n-back, arithmetic audio tasks are limited in the amount of performance monitoring they provide, since the final answer resulting from a sequence of calculations is usually taken into consideration and the single responses are rarely analyzed. Consequently, the N-back task provides the benefit of constant information on the participant's accuracy.

Since secondary tasks are employed with the goal of overloading participants' cognitive capacities, an expected outcome, during phases of high workload, is a drop in performance on the secondary tasks, due to participants focusing their available mental resources on solving the primary task [40], [39]. Among pilots this effect can be particularly strong, as they are trained to prioritize flying following a strict hierarchy of actions: aviate, navigate, communicate. Consequently, it's important to choose a secondary task that is not overly complex, causing pilots to completely neglect it. However, it is relevant to bear in mind, that pilots are often well-trained in assignments involving mental calculations and memory tasks, as they are part of the pilot training and certification process. Following these considerations, we assume that the utilization of a higher-level n-back task could be suitable for pilots as a participant group.

Motivational Intensity

Similar to extreme environmental conditions, which can be perceived as a threatening stimulus that causes a stress reaction, situations triggering affective states high in negative motivational intensity have demonstrated narrowing effects on attentional processes [41]. Although both positive and negative affective states have been shown to induce high motivational reactions, resulting in a goal-oriented behavior and consequently cognitive narrowing, the induction of intense positive states in experimental settings has proven challenging [42]. With the ultimate goal of inducing cognitive tunneling through positive affective states high in motivational intensity, Prinet offered a monetary award to the participants achieving the highest scores within the MATB multitasking environment [14]. However, the study's outcomes did not indicate the onset of tunneling states. A possible interpretation, suggested by the author, is that the uncertainty behind achieving the highest score and the participant's lack of confidence has caused them to ignore the motivational effect of a potential monetary award.

In a preceding experiment with the similar goal of investigating performance changes and attentional tunneling, Prinet incorporated a secondary task in combination with an ego-threatening factor in the form of negative performance feedback [14]. The feedback consisted of an alarm, informing participants of their deteriorating performance and was designed to provoke highly motivational negative states. However, the feedback was in reality artificially generated and consisted of a sound playing at random intervals. Regardless of the fictive nature of the alarm, the reported anxiety levels during conditions containing performance feedback and a secondary task increased significantly and a substantial decline in participants' performance was observed. In line with these findings, previous research shows that the utilization of negative affect as an approach to inducing highly motivational states has a longer tradition of being successfully applied in experimental conditions, especially in the context of cognitive narrowing, compared to the usage of positive emotional states as a motivational factor [41].

Novel Situations

The potentially threatening nature of the challenge behind novel and unsolvable situations has been observed to cause goal-oriented behavior, that is characterized by an excessive focus on the unsolvable issue, frequently causing the neglect of other factors, which may be of similar or even higher importance. Closely related to Wicken's definition of attentional tunneling, this phenomenon is evident in some of the previously mentioned accidents, such as the Everglades 407 flight, where the aircraft's crew were overly fixated on a malfunctioning landing gear light, and the West Air Sweden Cargo Flight 294, during which a failure in one of the PFDs caused pilots to neglect communication and other sources of information.

This effect and its applicability as a tunneling trigger in experimental conditions has been investigated in the previously mentioned study, during which Prinet experimented with monetary awards as a factor prompting highly motivational positive affective states [14]. Comparing different tunneling triggers, such as monetary incentives, novel and unsolvable situations, and secondary tasks, Prinet observed that novel and unsolvable situations were particularly effective in triggering attentional tunneling, by causing a significant decline in performance and heightened visual fixation on the issue, as evidenced by the eye-tracking data.

Tunneling Determinants

Within an experiment focused on attentional tunneling, it is crucial to define the parameters for classifying participants' behavior as indicative of occurring tunneling states. Following the definition, attentional tunneling can be observed whenever an individual is focused on a certain source of information, hypothesis, or goal for a period that is longer than optimal, resulting in a neglect of other potentially relevant sources or factors. Consequently, two approaches to determining the occurrence of tunneling have been employed in previous research.

The first approach utilizes visual cues and is based on the assumption that a focused participant, who is not experiencing a state of tunneling, would manage to perceive different sources of information. In studies involving simulated flights, for example, the visual cues are commonly presented as objects, or vehicles on the runway during landing, also known as runway incursions [43], [36]. Similarly, in studies from other domains visual cues include interface changes [44], [45], or "target" elements that are presented along with similar "distractor" objects [46], [47].

The second approach is based on the ability of participants to adjust their goals and assumptions and is therefore focused on examining their decision-making processes [48]. This method has been applied in aviation studies, for example, by exposing pilots to degrading weather conditions and testing their ability to adjust their current approach and make the correct decision of changing their flight path and landing at a different airport [49], [11].

Due to the limited prior research on utilizing visual cues through unexpected cockpit behavior and in-flight events, this study will classify the occurrence of attentional tunneling by incorporating abnormal situations including visual changes.

2.4 Cognitive States and Ocular Behavior

Eye Movements and Terminology

Vision, as one of the fundamental sensory mechanisms for humans, holds a central role in our ability to perceive and interpret the surrounding world. However, due to the anatomical specifics of the human eye, the simultaneous perception of only a small amount of visual information is possible. The cone photoreceptors, responsible for perceiving sharp, highly detailed, and colored visual content, are located in a small central part of the retina, called the fovea, whereas the peripheral regions of the retina are occupied predominantly by the rod photoreceptors, which facilitate vision during low-light conditions and are more sensitive to motion [50].

These characteristics of our visual system highlight the underlying necessity for eye movements, in order to extract information and build a thorough understanding of our surroundings. Furthermore, it suggests that eye movements could reflect important information about the observed object, underlying cognitive processes, expectations, and the interplay between sensory input, selectivity, and cognitive states [51]. By recording a person's point of regard, eye-tracking provides insights into a person's intentional attention allocation (top-down processing), as well as what elements attracted their attention (bottom-up) [14].

Among the gaze behavior, detected by eye trackers, fixations represent a set of miniature eye movements focused on a specific object or region, that result in a stabilized image on the retina, which allows for the actual extraction of visual information [26]. Saccades, on the other hand, constitute the rapid eye movements from one fixation to another and serve as a means to direct the point of gaze [51]. It is generally presumed that during saccadic movements visual perception is suppressed. The resulting combination of fixation points and the saccades connecting them is commonly referred to as

Literature	Transition frequency	Dispersion	Saccade size	Number of AOIs	Number of fixations	Fixation duration	Other
Clémence Prinet, Experiment 1, 2016			average saccade length (↓)		number of fixations (↓)	mean fixation duration (↑)	
Di Nocera et al., 2006		NNI (↑)					
Di Nocera et al., 2007		NNI (↑)					
Di Stasi et al., 2015		entropy (↑)					gaze velocity (↑)
Faulhaber et al., 2020	frequency per minute (↑)			(↑)			dwell time percentage (ns)
Kistakis et al., 2022			saccade amplitude variation during multitasking (↑)		fixation frequency during multitasking (↓)	fixation duration variation during multitasking (↑)	saccade frequency, saccade velocity, saccade duration, blink frequency, blink duration, pupil diameter, (...)
Krejtz et al., 2018							microsaccade magnitude (↑), microsaccade rate (ns), microsaccade peak velocity and magnitude (ns), intra-trial change in pupil diameter (↑), inter-trial change in pupil diameter (↑)
Lu et al., 2020		entropy (↑)			fixation frequency (↑)		average fixation times (↓), saccade frequencies (↓)
Moacdieh et al., 2020	transition rate (↑)	convex hull area (ns), spatial density (↓), stationary entropy (↓)	mean saccade amplitude (↓)			mean fixation duration (ns)	Scanpath length per second (↓), backtrack rate (ns), transition entropy (↓)

Table 2.1: Literature overview of studies examining gaze changes between conditions with low and high workload: increases during high workload(↑), decreases during high workload(↓), not statistically significant (ns).

a "scanpath". In Eye-Tracking analysis, the position of the eye is often calculated based on regions defined by the researcher, called "areas of interest" (AOI).

Eye-Tracking Metrics Related to Workload and Tunneling

The different types of ocular movements that eye trackers provide information about allow the calculation of various metrics from the gathered data. Considering the well-established usage of eye-tracking in scientific experiments, a variety of combinations between the different metrics has been applied and evaluated.

In aviation, the impact of workload as a factor affecting gaze behavior has frequently been investigated and prior studies have demonstrated that pilots' visual scan patterns could be an indicator of variations in workload [52]. An overview of some of the revised literature discussed here concerning the effects of workload and attentional tunneling on gaze changes can be observed in Tables 2.1 and 2.2.

Traditionally the relationship between visual search behavior and workload has been explored by focusing on measures such as changes in the pupil diameter, the duration or frequency of fixations, and the "dwell time", which is a calculation of the time spent within an AOI [26], [27]. Pupillary size fluctuations in particular are a traditional measurement that has proven to be a reliable indicator of shifts in the experienced workload [53], [26], [27]. However as pupillary size is dependent on the available light, this technique is unsuitable for actual flights, causing researchers to seek appropriate alternatives.

A less frequently utilized measurement, promising to provide insights on attentional shifts, and attention allocation is the calculation of transitions between AOIs. Specifically in cockpit environments, where information is categorized across multiple displays in different positions, demanding active visual scanning, transition counts can provide information about a pilot's situational awareness and the rate at which they update their information. Related to this measurement are multiple similar metrics, primarily differing in their temporal definitions: transition frequency, transition rate, and switching

Literature	Transition frequency	Dispersion	Saccade size	Number of AOIs	Number of fixations	Fixation duration	Other
Clémence Prinet, Experiment 3, 2016		convex hull area (ns), NNI (↑)	average saccade length (↓)			average fixation duration (ns)	saccade duration to fixation duration ratio (ns), average dwell duration (AOI specific), percentage od fixation on each task area (AOI specific)
Desmet et al., 2019		horizontal and vertical fixation densities (↑)			number of fixations (↓)	fixation duration(ns)	saccade duration (↑)
Regis et al., 2014	switching rate (↓)			(↓)			heart rate (↑)
Reimer, 2009		horizontal (↓), and vertical (ns) gaze dispersion					horizontal/vertical central gaze location (ns)
Victor et al., 2005		standard deviation of gaze (↓)					glance duration (↑), percent glance duration exceeding 2 seconds (↑), standard deviation of glance duration (↑), glance frequency (↑), total glance duration (↑), percent of fixations on the road centre auditory task (↑), percent of fixations on the road centre task demand (↓),

Table 2.2: Literature overview of studies examining gaze changes between conditions with tunneling triggers and without: increases with triggers (↑), decreases with triggers (↓), not statistically significant (ns).

rate. Transition frequency is defined as the number of transitions within a researcher-defined time span, while transition rate measures the number of transitions per second [54]. Switching rate, on the other hand, is a measurement used in multiple studies involving attentional tunneling and it refers to the transitions within a custom-calculated interval that has been specifically estimated using the data's information density [55].

Research focusing on the effect of workload and attentional tunneling on the transition count has showcased opposing results for the two cognitive states. For example, in a study on the experienced workload during Single-Pilot Operations, a significant increase in the transition frequency between cockpit instruments and the external view was observed during phases of heightened workload [30]. Similarly, an experiment by Moacdieh et al. comparing the gaze variations during sudden and gradual workload changes showcased a significantly higher transition rate under high workload conditions [56]. These results contradicted the anticipated behavior of increased efficiency and, consequently, reduced transition frequency during cognitively challenging situations. Furthermore, transition frequency was the only metric exhibiting more efficient behavior under low workload conditions. Interestingly, in a study focused on attentional tunneling, but unrelated to workload, Nicolas Regis et al. detected a significant reduction in switching rate among participants experiencing tunneling [44]. These contradicting outcomes suggest that states of high workload and tunneling might lead to different results and that transition frequency could be a suitable metric for differentiating between the two states.

Recently, dispersion measures, which quantify the amount of spread between fixation points, have witnessed an increase in application among eye-tracking studies. Based on the theory of predictive coding, it is assumed that the brain constantly processes perceived sensory input, compares it to prior knowledge, and generates expectations based on the available information [57]. This process plays a crucial role in determining an individual's visual scanning strategy, which constitutes the selection of the next area that will be gazed at. It is assumed that the selection of regions for visual sampling involves multiple mechanisms, networks, and subprocesses and that insights into the dispersion of fixations could indicate fluctuations in the underlying processes [57].

In contrast to transition frequency, which provides more generalized insights into behaviors related to context-switching and information-updating, dispersion measures offer more detailed observations of ocular movements and selectivity on a smaller scale. To calculate the dispersion of fixation points, different approaches have been applied including calculations of the observed distance between nearest fixations compared to the expected distance between nearest points in a random distribution (Nearest Neighbour Index (NNI)), representations of the smallest area, containing all fixations (convex hull area), and entropy measures. Findings from studies applying NNI as a measure suggest a more

dispersed pattern during tasks of high temporal demand [54]. This trend is supported by two studies by Di Nocera et al. [58], [59], and an experiment by Prinet [14], all of which show an increase in dispersion through higher Nearest Neighbor Index values under higher workload.

In contrast to this line of research, studies applying stationary entropy have shown differing results, with statistically significant findings indicating both higher and lower entropy under high workload. Stationary entropy is a metric originating from information theory, which depicts the level of randomness or complexity in an observed configuration compared to an estimated maximum level of complexity or randomness [60]. It considers all possible states that can be observed and how these states are distributed. Shannon's entropy, commonly referred to as stationary entropy within eye-tracking research, is calculated using the following equation:

$$H(x) = -\sum_{i=1}^{n}(p_i)\log_2(p_i)$$

Where x represents the observed entropy value, n is the amount of observed states, i denotes individual states, and p(i) indicates the probability of the current state. Measuring complexity and uncertainty using this approach generates a direct connection between the calculated outcome and predictability [57].

In the previously mentioned study on workload changes by Moacdieh et al. [56], entropy was observed to be higher under lower workload, supporting the assumption that gaze dispersion becomes less random with increasing workload. However, multiple other experiments have observed an increase in the randomness of the fixation pattern under higher workload conditions, showing less efficient and less systematic scan patterns under high workload [61], [62].

Numerous studies focusing on attentional tunneling have investigated changes in dispersion during states of cognitive narrowing. Similar to research on workload employing entropy as a measure, the different dispersion metrics utilized in studies on attentional tunneling, showcase outcomes that frequently differ. For instance, within the automotive industry, multiple experiments focusing on driving during conversations demonstrated a statistically significant reduction in the standard deviation of vertical and horizontal gaze positions as a dispersion measure [21], [63]. However, a similar study examining drivers' eye movements during hands-free conversations resulted in a higher horizontal and vertical density of the visual scanning area [64]. Conversely, Prinets' experiment on attentional tunneling during a desktop reaction task observed no significant difference in the convex hull area, which is a similar dispersion metric [14].

Overall, it can be concluded that within research on workload and tunneling a variety of dispersion measures has been implemented. However, the outcomes have demonstrated contrasting results, with the majority indicating increased dispersion under high workload conditions. A potential explanation for these inconclusive results, apart from the differences in the dispersion calculation methods, could lie in the varying environments and tasks between the different experiments. Due to its wider application in studies that involve cockpit environments, entropy has been selected as the appropriate dispersion measure for the current study.

A more traditional approach that has resulted in more consistent findings and has been utilized both in research focused on workload and attentional tunneling, is the measurement of differences in the mean saccade amplitude [27], [56], [26], [54]. Saccade amplitudes describe the distance traveled by the eye during a saccade, which is the movement from one fixation to the next. This measurement is typically expressed in degrees and is calculated based on the AOI size, the Euclidean distance between two successive fixations, and the distance between the eye-tracking glasses and the AOI [65]. Saccade length, or interfixation distances, is an alternative metric representing a simplified approximation of saccade amplitudes. It measures the Euclidean distance between two fixations within an AOI and is often used with low-speed or low-precision eye-tracking devices [54] or whenever the distance between the eye-tracker and an AOI is unknown or changing. Given the consistency of the results from previous research indicating a persistent decrease in saccade length and saccade amplitude in relation to both higher workload and the occurrence of attentional tunneling [14], these measurements are expected to exhibit strong classification capacities. Within the current experiment, saccade length is selected

as a measurement, due to the frequent head movements and the position of the cockpit instruments, which is not perpendicular to the eyeglasses.

Eye-Tracking Metrics Considerations

Many factors can influence eye-tracking measurements, which requires experimenters to be cautious during both the planning and analysis of experiments. Particularly among pilots as a participant group, previous studies have shown the potential influence of factors such as experience [34], flight phase [26], and fatigue [66] on their visual scan behavior [52]. Furthermore, some measurements, such as saccade length, are considered idiosyncratic, meaning that data may vary based on individual differences among participants' eye behavior [54].

To reduce these effects, it is recommended that experimenters minimize heterogeneity among participants as much as possible. Moreover, a within-subject design, combined with the appropriate statistical tests that can account for potential participant effects, is recommended as a strategy reducing individual differences [54].

2.5 Cognitive State Classification via Eye-Tracking

Machine Learning Models for Workload and Tunneling classification

Due to the potential benefits, such as pilot-aware cockpits, in recent years, eye-tracking research has actively attempted to develop approaches for predicting cognitive states. Within studies investing workload state classification based on eye-tracking data, subjective measures such as the NASA Task Load Index (NASA-TLX) have commonly been utilized as a reference for training and testing machine learning models. As part of their literature research, Kaczorowska et al. observed that Support Vector Machines (SVM) are the most frequently employed method for classifying workload levels via eye-tracking data [67]. Moreover, SVM was reported to exhibit strong performance, achieving accuracy rates consistently exceeding 80%. Other popular machine learning approaches for workload classification include Linear Discriminant Analysis (LDA), k-Nearest Neighbors (kNN), Multilayer Perceptron (MLP), linear regression, and neural networks [67].

Within the same study, Kaczorowska et al. compared the performance of various machine learning models in classifying three levels of workload based on 7 eye-tracking metrics from 29 participants. The analyzed models included SVMs with different kernels, Logistic Regression, kNN, Decision Trees, Random Forest, and MLP. The results suggest that SVM with a linear kernel, logistic regression, and MLP achieved the best outcomes within this study.

Similar findings were observed in an experiment that focused on detecting attentional tunneling states based on the self-affinity of gaze direction among 13 participants. The results showcased that SVM and Decision Trees performed better than kNN and Quadratic Discriminant Analysis [45].

In a different study investigating machine learning approaches for detecting attentional tunneling, the performance of SVM and Adaptive-Network-Based Fuzzy Inference System (ANFIS) was compared [44]. The models were fit utilizing physiological data from 18 participants with metrics including the eye-tracking switching rate, the number of AOIs observed, and heart rate. The ANFIS neural network resulted in an error rate of 1.1, demonstrating better performance in comparison to SVM with a rate of 1.9.

With a focus on regression analysis, Bitkina et al. compared multiple regression models, including simple linear, polynomial, S-shaped value, conjunctive, and disjunctive models [68]. Within this study, the experimental outcomes indicate that polynomial and conjunctive models exhibited a better performance.

Suggested Methodology

Two separate discussions with machine learning professionals were conducted in order to establish the best approach for our research taking into account the literature research, type of data, and experiment constraints. In line with findings from the literature research, indicating SVM and logistic regression among the supervised machine learning methods to be approaches, exhibiting strong performance both in classifying workload and tunneling states, the discussions suggested these methods as potentially the

most suitable approach for this study. Both methods perform well when utilized with limited amounts of data and are less prone to overfitting than neural networks, which is why they were recommended. Furthermore, the outcomes of these methods are less reliant on the experimenter's experience, proving them suitable for this study.

A different approach recommended during some of the discussions was to consider employing XGBoost[1] - an open-source implementation of the gradient boosted trees algorithm. XGBoost is increasingly applied in recent studies and has shown strong performance, for example within competitions on platforms like Kaggle [69].

Further recommendations included simplifying the process by comparing multiple approaches at once using the TPOT[2] tool, which automatically evaluates multiple machine learning pipelines and recommends the best-performing strategy most suitable for the data at hand. Additionally, it was recommended to utilize cross-validation as a method that could enhance the performance of the selected approaches.

Within this study, the performance of the different approaches will be judged based on their precision and accuracy scores. A minimum score of 70% or higher for both accuracy and precision will be taken as a reference indicating a fair performance. This criterion is based on common recommendations within the field [70], [71]. Considering that the ultimate goal of this study is to potentially contribute to the development of a pilot-aware cockpit, precision stands out as an especially relevant factor. Inadequate precision could result in an excessive amount of false alarms, erroneously categorizing the pilot's cognitive state as high workload or attentional tunneling, which can potentially lead to pilot distraction, frustration, and an additional decline in situational awareness.

[1] https://xgboost.readthedocs.io/en/stable/index.html, last accessed: Dec. 2023

[2] https://epistasislab.github.io/tpot/, last accessed: Dec. 2023

3 Experiment

3.1 Study Goals and Hypotheses

The study described in the following text has been approved by the Ethics committee of the DLR Institute. A central goal of the study was to investigate the proposed experimental design and the effects of the selected conditions on workload and the occurrence of tunneling states. A further objective of this thesis was to explore different abnormal events suitable for cockpit environments that can potentially serve as factors determining the occurrence of attentional tunneling and to assess their perceptibility for pilots. We expect to detect variations in participants' eye behavior between different workload levels and between participants experiencing tunneling states and those who did not. The final objective of this work is to evaluate the performance of minimum two machine learning algorithms in classifying states of workload and tunneling. The following hypotheses have been formulated as a result of these objectives:

- Can high, low workload, and attentional tunneling be induced in experimental conditions?

  H1.1: Both low and high workload can be induced within the experimental conditions of a cockpit simulator.

  H1.2: States of attentional tunneling can be induced within the experimental conditions of a cockpit simulator.

- Can machine learning algorithms classify pilots' workload states based on their transition frequency, mean saccade length, and gaze entropy?

  H2.1: Transition frequencies increase during states of high workload.

  H2.2: The mean saccade length decreases during states of high workload.

  H2.3: The observed entropy increases during states of high workload.

  H2.4: Machine learning algorithms can successfully classify low and high workload states based on transition frequency, mean saccade length, and entropy.

- Can machine learning algorithms classify the occurrence and absence of attentional tunneling among pilots based on their transition frequency, mean saccade length, and gaze entropy?

  H3.1: Transition frequencies decrease during states of attentional tunneling.

  H3.2: The mean saccade length decreases during states of attentional tunneling.

  H3.3: The observed entropy decreases during states of attentional tunneling.

  H3.4: Machine learning algorithms can successfully classify the occurrence and absence of attentional tunneling based on transition frequency, mean saccade length, and entropy.

3.2 Participants

A total of 15 male expert pilots with a mean age of 41.01 (SD = 8.79) participated in this study. Among them, 11 of the 15 pilots held an A320 certification, while 2 pilots possessed an A330 license in combination with either an A350 or A340 license. Due to the similarities between the A330 and A320 cockpits, all 13 Airbus pilots were treated as a single group. Notably, two expert pilots certified for other types of aircraft, but without an Airbus license were included due to their willingness and the limited number of participants available. Including them in this study aimed at collecting as much data as possible and potentially investigating, whether the eye-tracking data reflected the varying

experience levels within the A320 cockpit among the two pilot groups. A detailed overview of the participants and their averaged data can be observed in Table 3.1

Aircraft	Count	A320	Total flight hours (M / SD)	Flight hours this year (M / SD)
Airbus	13	11	9015.38 / 5299.61	513.85 / 177.79
Other	2	0	7000.0 / 6000.0	630.0 / 170.0

Table 3.1: Participants' experience level.

3.3 Experimental Conditions and Design

Building on insights from the literature research, the experiment was designed with the main objective of investigating experimental scenarios suitable for a simulator cockpit, that have the potential to induce high workload levels and trigger tunneling states.

An overview of the literature research findings illustrating potential triggers of attentional tunneling is presented in Figure 3.1. Some of the described triggers were unfeasible within this experiment due to technical limitations, such as the lack of HUD and SVS displays in the simulator and the challenges associated with replicating air traffic control (ATC) communication and air traffic fluctuations, which would require the recruitment of ATC officers in addition to pilots. Other approaches, such as workload manipulation through auto and manual flight modes, were deemed inappropriate due to their potential impact on participants' eye behavior. Furthermore, scenarios incorporating extreme environmental conditions such as loud aperiodic noises or intense lighting were excluded, due to the potential danger for both participants and experimenters, when exposed to these factors for longer periods, for example during multiple experiments in a single day.

To optimize our suggestions and ensure a balance between pilots' expectations, skills, motivations, and professional traits, our proposed scenarios were discussed with an experienced pilot, who supported us in effectively blending the literature findings with practical considerations and technical limitations. Given that both workload and tunneling are main topics within this study, the experimental design aimed at incorporating one condition with workload as an attentional tunneling trigger and another containing a different tunneling trigger.

Workload as a Tunneling Trigger

Among the workload-inducing factors, presented in the literature review, multitasking via an auditory secondary task was selected as an appropriate approach for this experiment, due to its common usage in prior studies, proven effectiveness, and minimal influence on eye-tracking metrics.

Our literature research revealed a wide variety of auditory secondary tasks used in experiments to challenge participants' working memory capacities. The identified approaches include n-back tasks [21], arithmetic calculations within numerical sequences [29], reverse word spelling, backward repetition of number sequences, and tasks related to recognizing and keeping count of specific sounds [72], [73]. From these options, the n-back task was preferred due to its flexibility in adjusting task difficulty and its ability to provide continuous information about participants' performance. Unlike tasks focused on numerical counting, which typically require participants to provide their responses after each run, and tasks involving arithmetic operations, where a single mistake in a series of operations results in an error for the entire sequence, n-back tasks demonstrate higher flexibility and time-sensitivity.

N-back tasks involve sequences of either numbers or words, to which participants must respond when the currently presented element matches the element situated n-steps back. For instance, if the sequence 1, 2, 3, 4, 3, 5, 4 is presented as part of a 2-back task, participants should respond upon hearing the second occurrence of the number 3. In the case of a 3-back task applied to the same sequence, participants are expected to react when hearing the second occurrence of the number 4. The following text will refer to the numbers, that a participant is expected to react to as "targets"

or n-backs, whereas the numbers that do not correspond to the previous n-step will be considered "empty" elements or "non n-backs." To ensure that participants experienced an increased workload without being overwhelmed by the complexity of their responsibilities, which can potentially cause them to completely neglect the secondary task, this study has utilized a 2-back task.

Ego Threat as a Tunneling Trigger

A relevant insight from the discussions with the previously mentioned expert pilot, who supported us in developing the experimental design, was the identification of ego threats as tunneling triggers potentially suitable for pilots specifically. Although negative performance feedback has previously been explored and has demonstrated an insignificant impact on cognitive narrowing [14], pilots, as a participant group, could presumably be more susceptible to its influence. Due to their professional training in simulators, pilots frequently express high expectations for their own performance and a strong motivation to demonstrate their skills. This presumably amplifies the motivational intensity of receiving negative feedback and having their attention drawn to their failure. To establish measurable criteria for a pilot's performance, participants were assigned a focus task of closely following a predefined flight path, based on which the feedback on their performance was assessed. This task additionally simplified the subsequent eye-tracking calculations, as it focused the participants' gaze predominantly on the main flight instrument - the Primary Flight Display (PFD). The selected negative feedback consisted of an aperiodic beep sound, that resembles aircraft alarms. Unfortunately, due to technical limitations, it was not possible to provide live feedback evaluating participants' per-

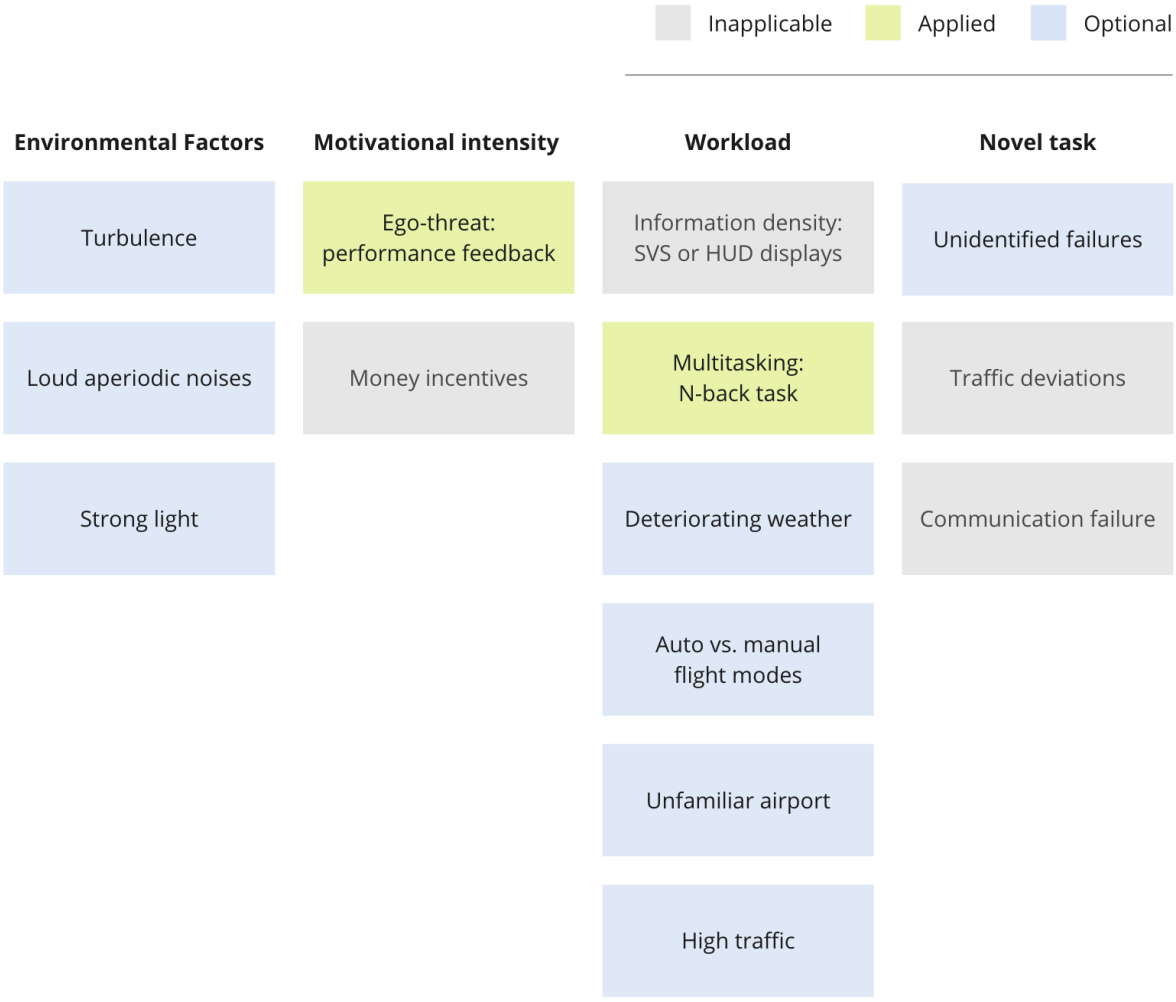


Figure 3.1: Triggers of attentional tunneling.

formance. Therefore, pseudo-feedback was generated by presenting participants with an audio file containing randomized beeps. To increase the effect of the static feedback and maintain participant engagement, even if the beeps didn't align with the actual performance, participants were informed that their performance is evaluated by a complicated algorithm developed by other researchers, whose work needs to stay undisclosed.

Tunneling Determinants

As part of the experimental framework a criteria determining the onset of attentional tunneling needed to be defined. Following findings from the literature research, participants' attention could be assessed by either incorporating events testing participants' ability to notice them or by testing participants' decision-making. Consequently, unexpected events encompassing visual cues were selected as tunneling determinants, in order to reduce the repetition of similar situations across multiple scenarios and to ensure that determinants are presented in the same flight phase, in order to avoid unwanted effects on the eye-tracking data. Even though novel and unsolvable situations have shown to be a strong trigger of cognitive narrowing [14], the development of a large number of untypical aircraft behaviors was not possible due to the technical characteristics of our simulator. Therefore within this experiment, unexpected situations have been utilized as events, testing whether a participant is currently experiencing cognitive narrowing, rather than as tunneling triggers. In the following text, these events, evaluating the participant's attention and cognitive state will be referred to as either tunneling determinants, abnormalities, or simply events. The selected tunneling determinants included:

1. An unexpected flickering of the Instrument Landing System (ILS) button. The flickering frequency was set to 1 Hz, corresponding to a rate of 1 second. The choice behind this frequency was based on recommendations from a previous study on automobile brake lights and drivers' reaction times. The optimal flickering rate for break lights, to achieve standard reaction times among car drivers under normal conditions, was established to be 4Hz [74]. Following the assumption that pilots are well trained and expected to react to less salient stimuli, combined with our goal to challenge participants' attention, frequencies encompassing rates lower than the suggested optimal were chosen for this experiment.

2. A Radio Altimeter 1 (RA1) fault warning on the Electronic Centralized Aircraft Monitor (ECAM) was displayed without the usual accompanying sound alarm.

3. Switching between multiple pages on the lower ECAM without an apparent reason. The frequency of the page switches was set to 4 seconds (0.25 Hz). Due to the larger size of the lower ECAM display and its significant role during flight operations, compared to the previously mentioned ILS button, a lower frequency was chosen for this determinant. This decision was influenced by findings from previous research suggesting that higher flicker frequencies result in faster reaction times [75].

4. A balloon passing nearby.

A visualization of the described events and their location within the cockpit can be observed in Figure 3.3. Furthermore, Figure 3.2 illustrates the position and names of the different displays, which additionally served as AOIs within the eye-tracking setup. The placement of the determinants in different locations within the visual field was designed to prevent participants from concentrating their gaze on a specific region. Additionally, we aimed to explore whether a determinant's position had an impact on its visibility and participants' reaction times.

Trigger + Determinant Combinations

We assume that both tunneling triggers, represented by different experimental conditions, and tunneling determinants, in the form of various types of abnormal events, affect how often an event is noticed. However, due to the limited participant pool, a statistical evaluation of the combined effects of the triggers and determinants was unfeasible within this study. As a result, the participants were

Figure 3.2: An overview of the A320 cockpit displays and the corresponding AOIs: Overhead Panel (OVHD), Flight Control Unit (FCU), Electronic Flight Instrument System (EFIS), Attention Getter Panel (ATG), Electronic Centralised Aircraft Monitoring (ECAM), Navigation Display (ND), Primary Flight Display (PFD), lower ECAM Display, Multi Purpose & Display Unit (MCDU). Credit: DLR (CC BY-NC-ND 3.0) [1]

Figure 3.3: An overview of the utilized tunneling determinants: 1. ILS Button Flicker, 2. RA1 Warning, 3. Lower ECAM Page Switch, 4. Balloon. Credit: DLR (CC BY-NC-ND 3.0) [1]

separated into two groups, experiencing different types of trigger + determinant combinations with the ultimate goal of providing insights and recommendations based on the descriptive observations resulting from the comparison. To ensure a sufficient number of attentional tunneling instances, the first group, consisting of 8 participants, experienced a predefined set of combinations, that was optimized for maximizing the occurrence of attentional tunneling. Conversely, the second participant group, comprising 4 participants, was presented with randomized combinations of triggers and determinants. The decision to divide the participant pool was driven by the desire to explore the impact of different triggers and determinants on the occurrence of attentional tunneling and to evaluate the effectiveness of the proposed combinations. Additionally, it aimed to identify events that are challenging enough for the attentional capacities of pilots to the extent that not all of them would take notice, thereby distinguishing these events as suitable indicators of pilot attentiveness.

To increase the occurrence of attentional tunneling, conditions and events were combined, guided by assumptions linked to stimulus-driven visual search. Consequently, the following combinations were selected:

[1] https://www.dlr.de/de/service/impressum, last accessed: November 2023

- *Baseline* + RA1: *Baseline* , as the presumably easiest condition was paired with the event, assumed to be the most challenging to notice. The RA1 warning is a small red text that appears on the ECAM and stays static on the display. This event was expected to be the most difficult to detect due to its small size, its remote positioning from the focus area (the PFD), and its static nature, making it less noticeable for the peripheral vision, which is sensitive to movement.

- *Performance* + ILS Flicker: Both the flickering ILS and the lower ECAM page switch were assumed to be similarly easy to detect. The ILS button flickered at a frequency higher than the rate of the ECAM page switch, and it was located in closer proximity to the PFD, both of which increased its probability of being noticed [75]. Conversely, the lower ECAM display flickered at a lower frequency but included a much larger element containing highly relevant information about the aircraft's state.

- *Workload* + Lower ECAM Page Switch: Similar to the ILS button flicker, the page switch on this display was believed to be noticeable even during challenging scenarios.

- *Combined* + Balloon: For the *combined* condition, which incorporated two tunneling triggers simultaneously, an event was chosen that is observable for a larger amount of time, includes a bigger element, and is unlikely to occur in a real-life flight scenario, thereby making it easier to notice.

The randomized combinations, on the other hand, included the following pairings: *baseline* + ECAM (x2), *baseline* + balloon, *baseline* + ILS, *performance* + RA1 (x2), *performance* + ECAM, *performance* + balloon, *workload* + balloon (x2), *workload* + ILS, *workload* + RA1, *combined* + ILS, *combined* + ECAM, *combined* + RA1, *combined* + ILS

Experiment Design

With the aim of gathering as much data as possible with the available resources, a within-subject design was chosen for this experiment. To minimize order effects, the conditions were presented in a pseudo-randomized order employing a Latin-Square design. As previously mentioned, participants were separated into two groups: the first group experienced a combination of conditions and determinants that was specifically designed to maximize the potential onset of attentional tunneling, while the second group encountered a randomized combination. Notably, the differences between the two groups are relevant only concerning the question of whether tunneling can be induced in experimental conditions and in relation to analyzing the frequency of tunneling onset. Within the eye-tracking analysis and machine learning training, both groups were combined and treated as a single cohort.

<table>
<tr><th colspan="2" rowspan="2"></th><th colspan="2">Performance feedback</th><th>Tunneling determinants</th></tr>
<tr><th>No</th><th>Yes</th><td rowspan="3">Flickering LS button;
Lower ECAM page switch;
RA1 warning without sound;
Balloon.</td></tr>
<tr><td rowspan="2">N-back task</td><td>No</td><td>baseline</td><td>performance</td></tr>
<tr><td>Yes</td><td>workload</td><td>combined</td></tr>
</table>

Figure 3.4: Experiment design overview.

3.4 Materials and Methods

Simulator

The experiment was conducted at the Air Vehicle Simulator (AVES) at the Institute for Flight Systems at DLR Braunschweig. It is a large-scale motion simulator with modular components, enabling the integration of different cockpits. For this study, the A320 cockpit was used. The simulator holds the highest level of qualification for a simulator, attributed to its accurate replication of real-world aircraft performance.

The inside of the simulator allows space for up to five persons, as illustrated in Figure 3.6: 1. *Operator (left)*, 2. *Operator (right)*, 3. *Observer*, 4. *Captain*, 5. *First Officer*. As part of this experiment, the seat of the *Operator (left)* was occupied by a specialized simulator operator. Their responsibilities included starting and adjusting the simulation, activating the audio recordings, and providing simulator-specific instructions. Next to the simulator operator, one of the two experimenters was seated in the *Operator (right)* position. During the experiment, their responsibility was to follow the N-back task and document participants' answers. The *Captain* position was taken by the second experimenter, who briefed participants on the experiment's context, tasks, and procedures. Additionally, they recorded participants' workload self-assessment and the timing when the tunneling determinants were noticed. Participants were offered the *First Officer* seat.

Recorded data from the simulator included simulation time, tunneling determinant event start and end times, chronometer readings, and GPS information about the latitude and longitude coordinates of the aircraft.

Figure 3.5: AVES simulator
Credit: DLR (CC BY-NC-ND 3.0) [2]

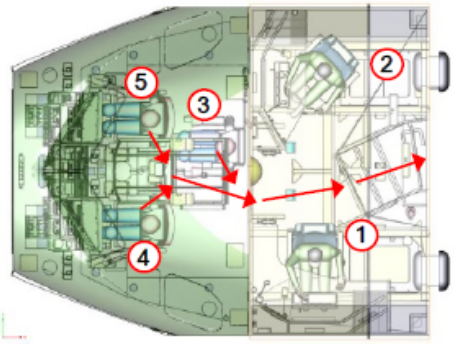


Figure 3.6: AVES A320 cockpit configuration

Eye-Tracking Hardware and Software

The SensoMotoric Instruments[3] (SMI) glasses with a 60 Hz sampling rate were utilized in this experiment. The glasses adopt a video-based dark-pupil technology for estimating the pupil position. This technology consists of infrared-blocking glasses, two infrared cameras, and infrared lights [76]. By illuminating the eye with infrared beams, the relative position of the reflected lights in relation to the center of the pupil is estimated and tracked [51]. Monitoring the reflections enables the calculation of the eye position relative to the glasses. However, since this estimation is linked to the glasses' location only, and considered that pilots are expected to move during the experiment, a 6D head-tracking system was integrated into the simulator. The system consists of five cameras and head-tracking targets attached to the glasses. The cameras monitor the targets and determine the gaze position by combining the location data of both the pupil and the head.

The raw eye-tracking data captured by the SMI glasses was recorded in synchronization with the simulator data and has been post-processed using an in-house software, called Eye Tracking Analyser (EyeTA) [77]. By utilizing a velocity-based algorithm, the EyeTA software differentiates between

[2] https://www.dlr.de/de/service/impressum, last accessed: November 2023
[3] https://gazeintelligence.com/smi-product-manual, last accessed: Dec. 2023

fixations and saccades within the raw data. The algorithm is based on the assumption that the eye-tracking device retains a constant sampling rate. The distances between consecutive points are then calculated and compared to a predefined angular velocity threshold to differentiate between the two eye movements [78].

Self-Assessment Workload Measurements

During this experiment workload was measured utilizing two subjective self-assessment techniques – a NASA Task Load Index (NASA-TLX) questionnaire evaluated the overall workload, encountered during a run, while the Instantaneous Self-Assessment (ISA) of workload technique provided continuous information about a participant's workload during the flight.

NASA-TLX

The NASA-TLX workload evaluation method is a multi-dimensional scale, originally developed to estimate operators' workload in the aviation sector [79]. It is one of the most common measures for subjective workload and has been applied in various fields. Tests on its validity and reliability have demonstrated its consistent efficiency across various participant groups [80], [81]. Within the NASA-TLX questionnaire, workload is evaluated across six dimensions: mental demand, physical demand, temporal demand, effort, performance, and frustration. The different dimensions are represented by scales consisting of values from 0 to 100 in increments of five. There are two methods for interpreting NASA-TLX. One of the approaches consists of directly analyzing each reported value, also referred to as NASA-TLX Raw. This method results in 6 answers corresponding to each of the 6 dimensions with every questionnaire completion. Conversely, in addition to the NASA-TLX scale itself, the second approach requires participants to submit a pairwise comparison between dimensions, determining which factor has a stronger impact on their personally experienced workload. Based on the pairwise comparison, an individual assessment of the weight of each dimension is calculated for every participant. Participants' answers on the dimensions are then multiplied by the dimension's weight. Subsequently, the weighted responses given for each NASA-TLX completion are summed, yielding a single answer between 0 and 100 every time the questionnaire is submitted. Within this experiment, the weighted NASA-TLX has been employed, due to its ability to capture participants' personal interpretation of workload and the convenience associated with working with a single value instead of six. Participants were asked to fill in the pairwise comparison at the beginning of the experiment, while the NASA-TLX questionnaire was presented at the end of each run.

ISA

A less common approach for estimating subjective workload is the Instantaneous Self-Assessment of workload technique (ISA), which has been developed by the UK Civil Aviation Authority, originally aimed at assessing air traffic controllers' workload during multitasking activities. Unlike the NASA-TLX measurement, ISA evaluates workload based on a unidimensional assessment, making its implementation during challenging tasks easier. Its advantage over other subjective workload measurements lies in its ability to provide continuous, time-sensitive information about a participant's workload, allowing researchers to investigate factors potentially causing fluctuations in participants' experienced workload [82]. The ISA rating is gathered at regular intervals by asking participants to evaluate their workload on a scale ranging from one to five, with one translating to low workload and five indicating a very high workload. In previous studies, participants' ISA self-assessment answers have been collected using either verbal responses or integrated interfaces with input fields.

In a study by Tattersall et al., the efficiency of the ISA technique in representing workload has been compared to physiological measurements and the more common Subjective Workload Assessment Technique (SWAT), which, similarly to the NASA-TLX, is a multidimensional questionnaire presented upon task completion [83]. The outcomes revealed that ISA ratings were often more sensitive to workload than SWAT. There was a consistent correlation between ISA ratings and SWAT responses, with a stronger correlation observed across the mental effort and concern dimensions and a weaker correlation in the time-pressure dimension. Notably, it was observed that employing ISA prompts

during a task reduced performance for a certain amount of time, after which participants regained their original performance.

Within our experiment, the ISA measurement has been automated by utilizing an audio file that prompted participants with the question "Your workload from one to five" every two minutes. The participants were expected to answer the question verbally, in order to avoid distraction from the main tasks and to reduce discrepancies in the eye-tracking data. They were instructed on how to interpret this scale as part of the briefing.

The ISA technique was considered a valuable method for this study, as it provides time-sensitive information and can support detecting potential fluctuations in the workload within different flight situations. Furthermore, it provides a detailed overview of participants' cognitive changes, which is crucial for the correct interpretation of the highly time-sensitive eye-tracking data.

N-back Accuracy Measurements

The accuracy of participants in completing the n-back task was collected throughout both the *workload* and *combined* conditions. Participants' responses were monitored and manually documented by one of the experimenters. Following the Signal Detection Theory, false responses were categorized into two types - "misses" and "false alarms" [84]. An answer was considered a "miss", whenever a participant failed to report a number that corresponded to the penultimate one, while "false alarms" were recorded when a participant reported an n-back, that did not match the penultimate number.

Despite the common usage of n-back tasks as a means to increase participant workload, studies frequently neglect reporting the method used for assessing participants' performance. As a result, standardization for evaluating n-back accuracy is lacking [40]. Many studies fail to differentiate between omission errors, or "missed" n-backs, and commission errors, or "false alarms." It has been observed that omission errors occur more frequently and are positively correlated to reaction times, unlike "false alarms" [85]. The two kinds of errors are assumed to correspond to differing cognitive processes, making evaluations without differentiation between the two potentially misleading. Consequently, within this study, the two error types have been analyzed separately.

Eye-Tracking Measurements

As previously mentioned, this study focuses on eye-tracking data depicting the transition frequency, entropy, and saccade length. This section gives an overview of how the separate metrics have been calculated.

Transition Frequency

As per its definition, transition frequency quantifies the number of transitions between AOIs within a time interval defined by the experimenter. In this study, the transition frequency was examined within a 30-second time window. This specific duration has been selected with the goal of enabling comparability between workload and tunneling data. The choice aligns with the temporal windows employed in other studies on attentional tunneling using transition quantification measurements [44], [55]. Additionally, the two-minute interval between two workload self-assessments can easily be divided into multiple intervals of 30-second intervals, thereby simplifying the comparison between the two cognitive states.

Entropy

As mentioned earlier, cockpit displays are designed to separate information, often presenting different content types on different displays. Since the transition frequency already provides information about a participant's choice to switch between displays and thereby change contexts, for the entropy analysis, the decision was made to focus on the entropy changes within the PFD display specifically. This choice is justified by the fact that the overall entropy is highly dependent on the cockpit design, which could potentially lead to generalized information that overlaps with insights from the transition frequency.

Considering the assumption that visual scanning behavior rests upon multiple underlying processes and that entropy could indicate changes and deficiencies within these processes [57], we assume that

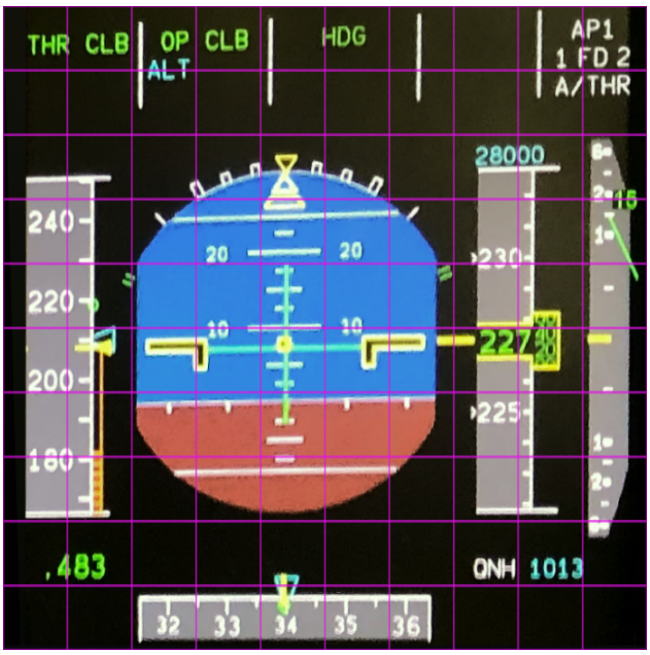


Figure 3.7: Entropy state space (bin) division grid.

focusing on entropy fluctuations within the PFD display could provide detailed insights into shifts within the cognitive functions.

Entropy calculations typically involve dividing the space in which entropy is evaluated into multiple areas, referred to as bins or state spaces. This process of discretization has a significant impact on the entropy calculations and the way data is interpreted. However, there is no widely accepted standard method in current research. Three different approaches to defining the bins are typically utilized [57]: using a grid to separate the area in equal spaces [61], dividing the area into content-driven segments (similar to AOIs) [56], and data-driven approaches based on identifying regions of clusters. Studies involving entropy calculations frequently fail to report how the space has been divided, adding to the challenge of selecting an appropriate approach.

To ensure the best possible results for our data, the PFD display was split into a 10x10 grid that approximately divides the interface into areas with different content, as illustrated in Figure 3.7. This approach combines a grid-driven method while also taking content separation into account.

It is important to note that the entropy calculations were based on the code provided in a study by Shiferaw et al. [57].

Mean Saccade Length

Similar to our approach of assessing entropy with a focus on data from the PFD display, saccade lengths have been calculated exclusively within the PFD. If saccades were computed across the entire visual field, occurring transitions would have a great impact on the mean saccade length. Considering that the length of saccades during transitions is determined by the distance between different displays, including them in the mean calculation would essentially only indicate instances of transitions. Moreover, due to the varying sizes and interfaces of cockpit displays, comparing the mean saccade length across all displays would be highly dependent on the specific interface characteristics. As a result, the mean saccade lengths have been calculated exclusively from fixations within the PFD, ensuring an analysis that is less influenced by interface variations.

3.5 Experiment Procedure

At the beginning of each experimental session, participants received a presentation covering the experiment's background, objectives, the upcoming procedure, the route they would be flying, and their assigned tasks. These tasks consisted of the focus assignment of following a predefined flight path as closely as possible, in addition to run-specific conditions such as an auditory n-back task or au-

dio performance feedback on the focus task. Additionally, they were informed that within each run unexpected behaviors or events will occur and that they should report as soon as they notice something unusual. As a next step, participants were clarified on how the ISA self-assessment would be collected and what scale their answers should correspond to. At the end of the presentation, they were reminded they could always abort the session without negative consequences and were asked to sign a form of consent. Following this, participants were requested to complete a questionnaire with demographic questions, information about their flight hours and certifications, and the pairwise comparison, necessary for the NASA-TLX data analysis.

As a next step within each session, the simulator operator provided a presentation and a tour with safety instructions for behavior in the simulator. After the tour, participants were offered a pause. Once the pause was over, participants took part in a test flight covering the last 10 minutes of the route. After the test flight, they received a training session on the n-back task. To ensure the task was well understood, participants were allowed to proceed with the experiment, only if they achieved an accuracy of 60% or higher in the n-back task. As a last step in the preparation for the initial run, the eye-tracking glasses were calibrated.

At the beginning of each run, participants were informed of the upcoming condition and whether they would be expected to engage in a secondary task or receive feedback on their focus task. Additionally, they were reminded to pay close attention to unexpected events and to report them immediately. At the end of each run, participants were asked to fill in a questionnaire and were offered a pause. The questionnaire included the NASA-TLX rating, a self-evaluation question of their performance during the last run, a question asking them to describe any abnormal situations they encountered, and a question regarding the simulation quality.

Finally, at the end of each session, participants had the opportunity to ask questions and share insights into their overall experience.

4 Results

4.1 Data

Workload

To ensure that the evaluation of the workload data from both Airbus and non-Airbus pilots together is an appropriate approach, the initial step within the data analysis included a comparison between the NASA-TLX and ISA responses of both groups. Boxplot comparisons and histograms revealed no substantial differences in the self-assessments of the two groups. Furthermore, it was observed that data from the non-Airbus group predominantly lies within the mid-range values and contributes to the variety of the dataset. As a result of this observation, data from both groups were included in the workload dataset and were treated collectively within the workload analysis.

Throughout the experiments, during one run the NASA-TLX responses from a participant were lost due to technical issues, resulting in a missing NASA-TLX value. In contrast, the ISA scores of all 15 participants have been fully recorded. However, as some runs were finished earlier than others and participants were asked every two minutes what their workload was, the number of documented ISA results per participant and condition frequently differ.

N-back Accuracy

The n-back responses of all participants have been comprehensively documented and no data has been lost. However, it is noteworthy that the responses were recorded and transcribed manually, introducing the possibility of human errors in the data.

Tunneling

Similar to the n-back accuracy documentation, reaction times and the onset of tunneling were recorded manually and no data was lost. All analyses in the context of attentional tunneling were based on data from Airbus pilots exclusively. Non-Airbus pilots were intentionally disregarded, as the effective detection of abnormal behavior requires extensive knowledge of the standard behavior of an Airbus cockpit, which may be lacking among non-Airbus pilots.

Additionally, one of the Airbus pilots wore high-diopter glasses and participated without them due to the eye-tracking equipment. However, he reported experiencing difficulties with his tasks. Consequently, his data was excluded from the tunneling analysis, resulting in data from 12 participants for the tunneling evaluation.

Eye-Tracking

Comparisons of the ocular data from Airbus and non-Airbus pilots indicated differences between the two groups. As a result, the eye-tracking analysis focused on data from Airbus pilots only.

Furthermore, the validity of the eye data was assessed, and only data achieving a score above the 50 % threshold on the validity test was further utilized in the analysis. As a result, the eye-tracking data from ten runs was excluded due to its insufficient validity. None of the data from the previously mentioned participant with high diopter was incorporated into the analysis, due to its low validity score in combination with the assumption that his eye movements could have been irregular due to his limited sight. Ultimately, the data employed in the eye-tracking analysis depicted the gaze behavior of 12 individuals and a total of 41 runs.

4.2 Validation of the Workload Manipulation

The following section investigates the impact of the experimental independent variables on participants' workload. This has been achieved by analyzing participant's workload self-assessment answers

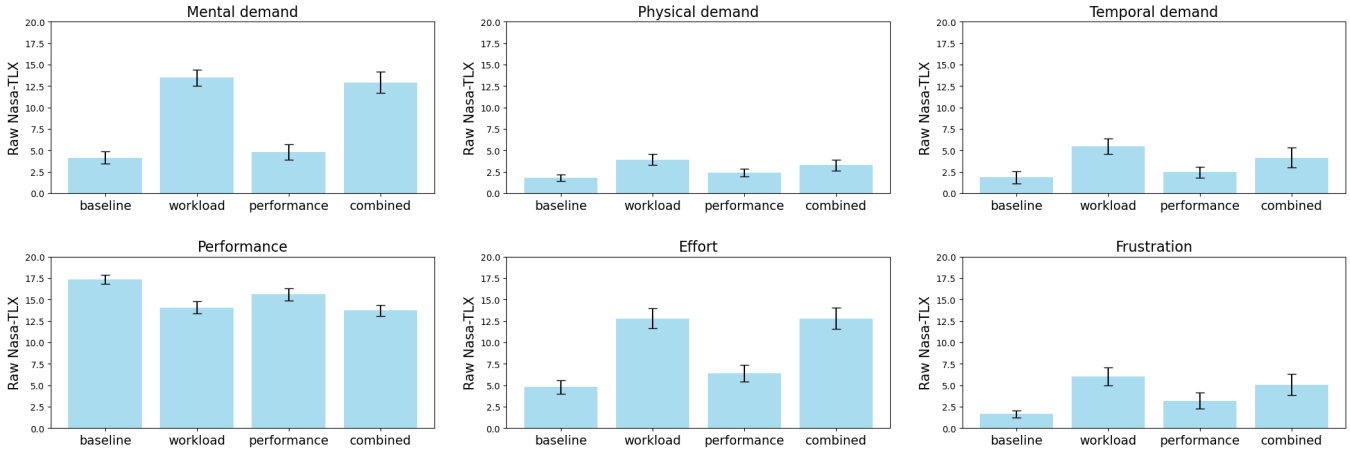


Figure 4.2: Overview of the mean Raw NASA-TLX results for each dimension across all conditions.

between conditions. As mentioned earlier, for the analysis of the workload self-assessment measurements data from both Airbus and non-Airbus pilots has been utilized.

NASA-TLX

The NASA-TLX data consisted of two types of measurements for each run - the raw values for each dimension and one weighted score, which was calculated based on all raw values and the participant's pairwise weighting submitted as part of the briefing questionnaire before the experiment began. An overview of the mean values within each condition can be found in Figures 4.1 and 4.2.

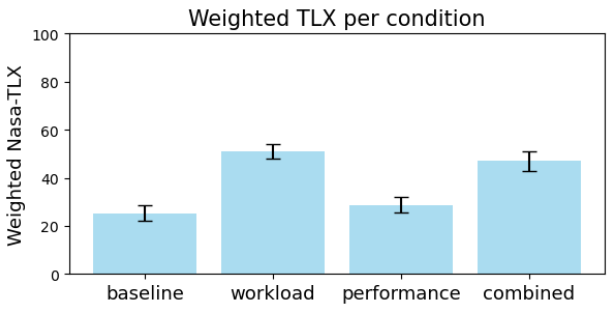


Figure 4.1: Overview of the mean weighted NASA-TLX results per condition.

The evaluation of our experimental design and its impact on the reported NASA-TLX utilized the weighted NASA-TLX value, as it is a single measurement representing all dimensions. The outcomes for the weighted NASA-TLX indicate mean values of 28.07 (SD = 16.26) for the *baseline* condition, 28.64 (SD = 12.5) for the *performance* condition, 50.91 (SD = 11.79) for the *workload* condition, and 47 (SD = 16.09) for the *combined* condition.

Cumulative link mixed models have been chosen as the appropriate approach for the statistical evaluation of the effect of the independent variables "n-back task" and "performance feedback". During the planning of our data analysis, we noticed that a large part of our data contains unequal sample sizes. Since the majority of the traditional tests can not handle missing values and for some of our measurements, for example, the tunneling gaze metrics, artificially balancing the data by removing parts of it, wouldn't be possible as this would leave us with only a few data points, we sought a statistical test, that is suitable for unequal sample sizes in a within-subject design. As a result, we employed different types of mixed-effects models within our statistical analysis. In structuring and presenting our findings, we followed the guidance provided by Meteyard et al. [86].

As mentioned, for the analysis of the NASA-TLX values a cumulative link mixed model was selected. This test was preferred over linear or generalized mixed-effects models, due to the ordinal nature of our dependent variable. To avoid overfitting and insufficiently supported assumptions, for our analysis, a model was employed, that is as simple as possible. Consequently, our tests adapt the intercepts of the individual slopes' by introducing participants as a randomizing factor. However, the slope variations themselves have not been randomized for the different participants. The fixed effects within our model were set to be the presence of an n-back task and accordingly the presence of performance feedback within a condition. An overview of the resulting model and its outcomes can

be found in Table 4.1. All models within our statistical analysis have been tested for homoscedasticity using the performance package and its check_heteroscedasticity function in R. Furthermore, the normal distribution of the residuals has been assessed utilizing Q-Q plots. As a result, the model's outcomes are reported here only if both assumptions have been met.

Findings from the NASA-TLX model indicate that the presence of an n-back task significantly increases the perceived workload. Performance feedback, however, does not display a substantial influence on the reported workload. Interestingly a significant interaction between the performance and n-back variables can be observed. This indicates that the presence of performance feedback while a participant is additionally engaged in an n-back task, significantly decreases the workload-inducing effect of the secondary task.

Fixed Effects					
	Estimate	**SE**	**97.5% CI**	**z-value**	**p**
Performance	0.73	0.65	2.01	1.11	0.26
N-back	6.79	1.13	9.02	5.97	**2.28e-09**
Performance: N-back	-1.89	0.96	-0.0008	-1.96	**0.049**
Random Effects					
			Variance	**SD**	
Participant (Intercept)			9.6	3.1	
Key: p-values and confidence intervals have been calculated in R 4.3.1 using the ordinal package with default settings.					
CLMM equation: *weighed_TLX ~ Performance*N-back + (1 | ParticipantID)*					
Number of observations = 59, groups: ParticipantID = 15;					

Table 4.1: Cumulative link mixed model assessing the effect of the independent variables on the reported TLX scores.

ISA

As an initial step in understanding our data related to the ISA self-assessed workload, normality tests were applied to the reported ISA scores. The results revealed that under some conditions the data did not follow a normal distribution. Even after removing outliers, the data did not achieve normality. As a result, we attempted to better understand our data investigating it in detail.

One of our assumptions was that the reported ISA scores varied in relation to the flight phase. It has been recognized that within a flight, different tasks correspond to different flight phases. For example, during landing, the amount of responsibilities a pilot needs to take care of is substantially larger than during a cruise. Following this assumption, we decided to explore the correlation between the ISA number and the corresponding answer. The ISA number indicates the sequential position of the ISA question within a run. A Spearmann correlation test was run on the ISA Answers and the corresponding ISA numbers resulting in a correlation of $r(423) = 0.27$ with a p-value < 0.001. Following these results, we attempted to maximally reduce the correlation, while preserving as much of the available data as possible. The ISA values were separated into three phases - beginning, middle, and end. Multiple approaches for splitting the data were attempted and tested for correlation. Finally, removing the first ISA from a run and the last two ISA's was identified as the most appropriate approach as it resulted in low correlation while preserving as much of the available data as possible.

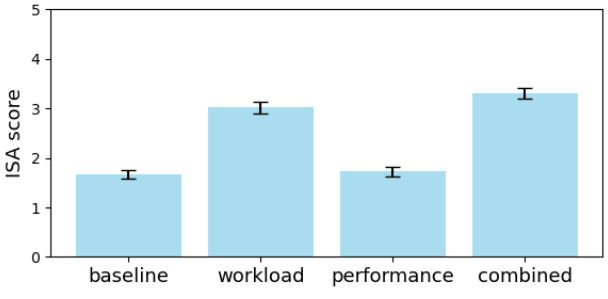


Figure 4.3: Overview of the mean ISA results per condition after reducing the data.

Consequently, the data has been reduced in alignment with our findings. The resulting mean values for each condition were 1.67 (SD = 0.08) for the *baseline* condition, 1.72 (SD = 0.09) for the *performance* condition, 3.01 (SD = 0.11) for the *workload* condition, and 3.29 (SD = 0.1) for the *combined* condition.

Within the assessment of the reported ISA scores, we followed the cumulative link mixed-effects approach utilized in the NASA-TLX analysis, as our data constituted different sample sizes and the target variable consisted of ordinal data. The outcomes of the model can be observed in Table 4.2. Considering the results outlined in Table 4.2, a significant increase in the reported workload during conditions incorporating an n-back task can be observed. Similar to the NASA-TLX outcomes, the presence of performance feedback does not indicate a significant effect on the reported ISA scores. However, in contrast to the results of the NASA-TLX model, no significant effect followed the presence of performance feedback in addition to n-back tasks. Ultimately indicating that performance feedback did not influence the effect of the n-back task on the reported ISA.

Fixed Effects					
	Estimate	**SE**	**97.5% CI**	**z-value**	**p**
Performance	0.00	0.39	0.79	0.02	0.98
N-back	4.60	0.52	5.63	8.79	**<2e-16**
Performance: N-back	0.94	0.54	2.01	1.71	0.09
Random Effects					
			Variance	**SD**	
Participant (Intercept)			3.7	1.92	
Key: p-values and confidence intervals have been calculated in R 4.3.1 using the lmerTest package with default settings.					
CLMM equation: *ISA ~ Performance*N-back + (1 | ParticipantID)*					
Number of observations = 243, groups: ParticipantID = 15;					

Table 4.2: Cumulative link mixed model assessing the effect of the independent variables on the reported ISA scores.

As an additional step in evaluating the validity of the experimental design, the presence of any order effects was tested, despite the pseudo-randomization of the condition order. A cumulative link mixed model was fitted using the NASA-TLX values. Having the run number as a fixed effect and participants as a randomizing factor, the model outcomes indicated no significant effects. Similarly, to test the potential effects of the run number on the ISA responses, a cumulative link mixed model was applied using ISA answers as dependent variables, run numbers as a fixed effect, and Participant ID as an intercept randomizer. No statistically significant findings were identified. Due to the purely prophylactic nature of this test, no detailed table with results is presented.

4.3 N-back Accuracy

Psychological evaluations incorporating secondary tasks similar to n-back and the earlier described arithmetic calculations are a central part of pilot certification exams for various types of aircraft in Germany [87], [88]. Since all pilots, regardless of the type of aircraft are trained in solving secondary auditory tasks, both Airbus and non-Airbus pilots within this study were included in the n-back accuracy calculations, resulting in data from 15 participants.

Two error types were recognized within participants' data - the percentage of missed n-backs and the percentage of false alarms. To ensure that data is meaningfully interpreted between runs with different lengths, both error types were calculated as a percentage of the amount of the corresponding n-back stimulus types. As a result missed items were calculated as a percentage of all n-back targets, whereas false alarms were measured as a percentage of all inputs that were not n-back targets. As only the *workload* and *combined* conditions involved an n-back task, the n-back accuracy data contains

information concerning these conditions only. The accuracy results indicate a mean value of 8.6 % (SD = 3.9) for missed n-backs in the *workload* condition and a value of 11.2 % (SD = 6.34) in the *combined* condition. On the other hand, false alarms resulted in a mean value of 1.67 % (SD = 1.04) for the *workload* condition and 1.87 % (SD = 1.35) for the *combined* condition.

To investigate, whether the audio performance feedback during the *combined* condition, affected participants' n-back accuracy, mixed effects models with performance feedback as a fixed effect and participants as a randomizing factor have been utilized. Due to the continuous nature of the n-back accuracy measurement, mixed effects have been preferred over cumulative link mixed models. Two different models have been fitted, one examining the effects on the percent of missed items, and one examining false alarms. Both statistical tests revealed no significant impact of performance feedback on participants' n-back accuracy. To ensure the adequate interpretation of the statistical analysis, the normal distribution and homoscedasticity of both models have been confirmed. As these tests are rather explorative, do not indicate significant values, and are not the focus of this study, no tables with detailed results are provided.

Since high workload has been identified within the literature research as a potential cause for impaired accuracy within the n-back task, the correlation between reported workload scores and pilots' accuracy has been explored. Considering the possibility that a pilot's perception of their own accuracy might affect their workload self-assessment, a direct causal relationship between the two measurements is potentially lacking. Therefore, a statistical correlation test instead of a relationship test has been selected as the appropriate approach. The results from Spearman's correlation tests showcase a potential association between the ISA self-assessment and the percentage of missed items, resulting in a coefficient of 0.208 and a p-value of 0.003. However, between the ISA score and the percentage of false alarms, no statistically significant correlation could be found. Similarly, the comparison between the NASA-TLX scores and the n-back performance does not showcase a correlation between the reported workload and any of the two accuracy metrics. An overview of the mean accuracy values per ISA score can be observed in Figures 4.4 and 4.5. Figure 4.4 demonstrates that the mean percent of missed items gradually increases with the growing ISA scores.

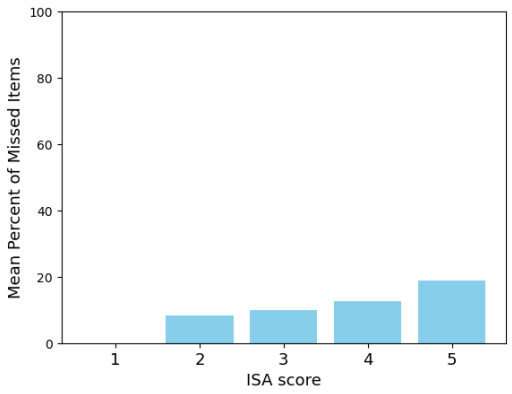


Figure 4.4: Overview of the mean percent of missed items per ISA score.

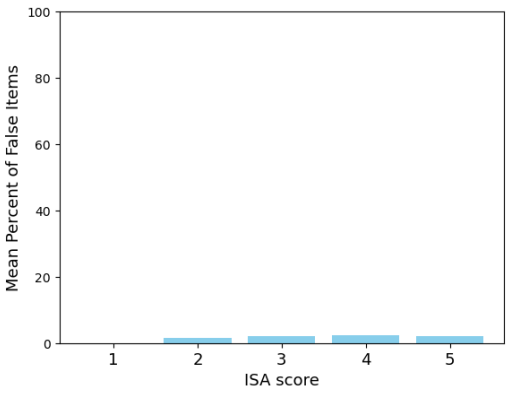


Figure 4.5: Overview of the mean percent of false items per ISA score.

4.4 Tunneling

Reaction times

The evaluation of tunneling data was narrowed down to Airbus pilots without visual deficiencies. As a result, the data included a total of 12 participants for the analysis.

To calculate the reaction times of participants both manually documented data and data generated by the simulator have been used. The onset and end times of the RA1, ILS, and lower ECAM events were contained within the simulator output data, whereas the appearance of the balloon within the pilots' visual field was manually documented by the experimenter sitting next to the participant. Similarly, the moment, in which pilots reacted to an occurring event was manually recorded. Due to

potential delays or human errors in the documentation of the reaction moment, response times have been analyzed with caution.

The reaction times were estimated by subtracting the event's start time from the moment when the pilot noticed the event. This measurement was mainly used for descriptive purposes and to recognize outliers within the data. Unusually long reaction times were considered outliers indicating attentional tunneling and were included in the dataset describing the presence of tunneling states. One outlier related to the RA1 determinant was identified using a boxplot in combination with a z-score method. As a result, data related to this outlier has been considered an instance of tunneling. Table 4.3 showcases in detail the resulting reaction times and frequencies, whereas Figure 4.6 illustrates visually the results.

Determinant	Event count	Notice frequency	Missed frequency	Reaction time (M/SD)	Event Duration
ILS	12	7	5	16.29 / 11.97	29,93
Lower ECAM	12	6	6	9.48 / 3.25	29,92
RA1	12	8	4	4.98 / 3.8	30,93
Balloon	12	11	1	30.85 / 13.22	50

Table 4.3: Overview of the reaction times and notice frequencies.

Tunneling Frequency, Triggers and Determinants

As shown in Table 4.3, 16 cases of attentional tunneling were identified, whereas in 32 cases, participants did not experience the onset of tunneling. As mentioned earlier, participants were divided into two groups consisting of 8 pilots, experiencing a predefined combination of conditions + events and 4 participants exposed to randomized combinations. To meaningfully compare the results from both groups the frequency of tunneling cases has been calculated as a percentage from the total number of occurrences of a tunneling event within the respective participant group, as can be observed in Figure 4.7. The interpretation of the results needs to be regarded with caution due to the small sample size, which could result in small differences having a large impact on the outcomes.

Our findings indicate that for two of the four events, the predefined combinations resulted in more cases of tunneling. However, the RA1 event combined with the *baseline* scenario within the non-randomized group appears to have resulted in an unintended higher detection rate.

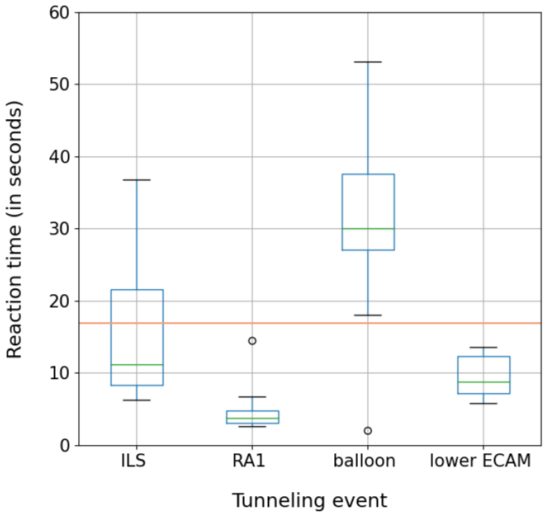


Figure 4.6: Boxplot of participants' reaction times per tunneling event.

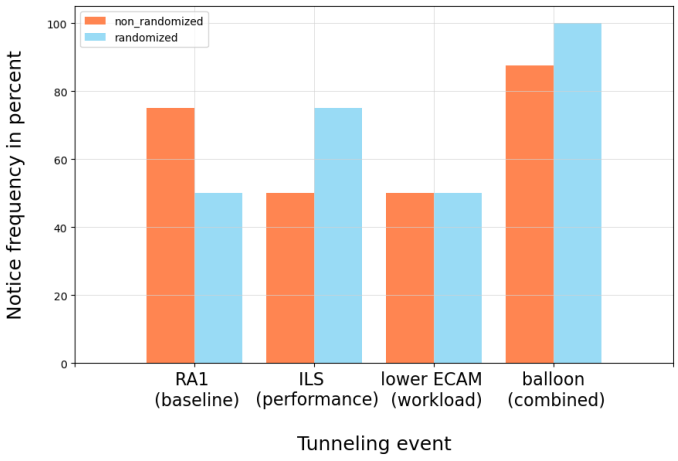


Figure 4.7: An overview of the frequency of tunneling states calculated as a percentage of the number of events within each determinant group. The conditions in the brackets represent the event + condition within the non-randomized group only.

4.5 Workload Gaze Analysis

To enable an analysis comparing workload-related gaze data and data depicting tunneling states, the recorded eye-tracking values were split into 30-second intervals. An overview of how this segmentation would take place within a run is presented in Figure 4.8.

Given the earlier demonstrated significance of the impact of an n-back task on the experienced workload, eye-tracking data originating from conditions containing a secondary task was regarded as high workload content. As a result data from the *baseline* and *performance* conditions have been categorized as low workload, whereas eye-tracking data during the *workload* and *combined* conditions were assigned to the high workload class. As discussed in section 4.2, variations in the reported workload within a run were found to be dependent on the flight phase. To guarantee a correct workload classification based on the condition type, the gaze analysis included only data from the earlier described middle phase.

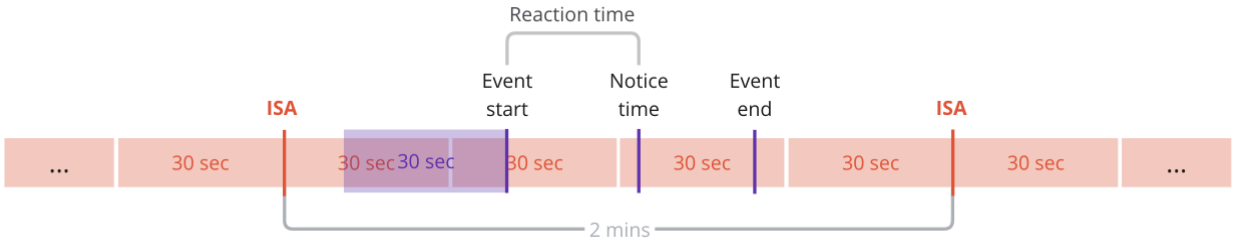


Figure 4.8: Temporal division of the eye-tracking data. The red values describe segments related to workload data, whereas the purple segment demonstrates the interval used within the analysis of tunneling data.

Workload and Transition Frequency

As previously mentioned, the transition frequency was calculated in intervals of 30 seconds. Within these 30 seconds, our results show a mean value of 13.67 (SD = 4.40) transitions during low workload and a mean value of 13.00 (SD = 4.98) during high workload conditions. For the statistical analysis of the transition frequency, a mixed-effects model was utilized, due to the discrete nature of the transition calculation and due to the unequal sample sizes and the commonly occurring personal differences within ocular behavior. Similar to the design of the cumulative link mixed models used in the self-assessment workload analysis, the simplest version of the model has been chosen in order to prevent overfitting. Additionally, the homoscedasticity of each of the models related to eye-tracking data has been tested in R using the performance package and its check_heteroscedasticity function.

The first attempt at fitting the model did not meet the assumption of variance homogeneity within the residuals. Although plots suggested that the heterogeneity could be a result of the considerably fewer data points from some participants compared to others, we decided to transform our data using a logarithm on the dependent variable. This approach was selected as it has frequently been recommended [89], [90], and because the results following the transformation did not significantly change the outcomes. The final findings, after applying a logarithm on the dependent variable, are presented in Table 4.4.

As can be observed, high workload significantly decreases the transition frequency. To allow interpretation of these results, the estimate of the workload effect within the logarithmically adjusted model has been recalculated following instructions on how to interpret results from logarithmic transformations [91]. The exponential of the estimate has been calculated and subtracted from one and finally multiplied by 100. The result of this calculation describes with how many percent the dependent variable changes, whenever the fixed effects class increases by one. In the context of this study, this calculation would indicate by how many percents the transition frequency increases or decreases, whenever a switch from low to high workload occurs. Following these calculations, the resulting estimations indicate that switching from low to high workload decreases the transition frequency by 11%. Finally, to confirm all assumptions have been met all models analyzing eye-tracking data have been

tested for the distribution of their residuals by visually examining the normality of the data using Q-Q plots. The results of the transition frequency Q-Q plots indicate a satisfaction of the normality assumption.

Fixed Effects					
	Estimate	**SE**	**97.5% CI**	**t**	**p**
Intercept	2.56	0.05	2.67	46.233	1.51e-15
Workload	-0.11	0.03	-0.05	-3.902	**0.000105**
Random Effects					
			Variance	**SD**	
Participant (Intercept)			0.03	0.17	
Residual			0.14	0.37	
Key: p-values and confidence intervals have been calculated in R 4.3.1 using the lmerTest package with default settings.					
LMER equation: *log(Transition frequency) ~ Workload + (1 | ParticipantID)*					
Number of observations = 712, groups: ParticipantID = 12;					

Table 4.4: Liner mixed-effects model assessing the effect of workload on transition frequency.

Workload and Mean Saccade Length

The descriptive calculations characterizing the mean saccade length of participants within this study displayed a mean value of 46.22 mm (SD = 16.07) during low workload conditions and a value of 43.84 mm (SD = 14.73) during high workload.

Similar to the approach used in evaluating the transition frequency, the effect of high workload on mean saccade length has been evaluated using a simple mixed-effects model. For the initial model, once again, homoscedasticity tests indicated heterogeneous variance within the residuals. As a result, a logarithm of the dependent variable has been utilized. The outcomes of the final model can be found in Table 4.5. By applying the earlier-mentioned calculations for interpreting logarithmically transformed data, it can be concluded that within our results high workload had a significant effect on saccade mean length by decreasing it by 7%.

Fixed Effects					
	Estimate	**SE**	**97.5% CI**	**t**	**p**
Intercept	3.79	0.06	3.93	58.417	1.5e-15
Workload	-0.08	0.03	-0.03	-2.961	**0.00317**
Random Effects					
			Variance	**SD**	
Participant (Intercept)			0.04	0.2146	
Residual			0.12	0.3495	
Key: p-values and confidence intervals have been calculated in R 4.3.1 using the lmerTest package with default settings.					
LMER equation: *log(Mean Saccade Length) ~ Workload + (1 | ParticipantID)*					
Number of observations = 711, groups: ParticipantID = 12;					

Table 4.5: Liner mixed-effects model assessing the effect of workload on mean saccade length.

Workload and Entropy

Our approach to calculating the effect of workload on entropy has mirrored the previously described analyses. It is relevant to mention, that our analysis incorporated the observed raw entropy values, rather than utilizing a normalized variation of the measurement. As mentioned in the literature research, it has been advised to report entropy values using normalized estimations, in order to increase replicability and understanding, since entropy is tightly related to the size of the AOI it has been observed in [57]. To assist better understanding of our results, we follow this recommendation. However, to avoid unintended averaging effects on the analysis of the data, the mixed-effects model was fitted using the observed raw values. Within our data, a mean value of 0.47 (SD = 2.94) was estimated for the normalized entropy during conditions of low workload, whereas during high workload conditions, the mean value was calculated to be 0.50 (SD = 2.99). Unlike the initial models for transition frequency and mean saccade length, the mixed-effects model for the observed entropy did not fail the homoscedasticity tests. Therefore no logarithmic transformations were applied to the entropy data. Results from the mixed-effects model indicate no significant impact of workload levels on the estimated entropy.

Fixed Effects					
	Estimate	**SE**	**97.5% CI**	**t**	**p**
Intercept	2.95	0.06	3.08	44.89	7.65e-15
Workload	0.05	0.03	0.11	1.37	0.17
Random Effects					
			Variance	**SD**	
Participant (Intercept)			0.04	0.21	
Residual			0.19	0.44	
Key: p-values and confidence intervals have been calculated in R 4.3.1 using the lmerTest package with default settings.					
LMER equation: *Observed Entropy ~ Workload + (1 | ParticipantID)*					
Number of observations = 707, groups: ParticipantID = 12;					

Table 4.6: Liner mixed-effects model assessing the effect of workload on entropy.

4.6 Tunneling Gaze Analysis

Tunneling and Transition Frequency

For the calculation of the eye measurements relating to attentional tunneling, an interval of 30 seconds before the onset of the tunneling determinant event was utilized. The outcomes of our estimates display a mean value of 16.4 transitions (SD = 6.47) within data not related to tunneling and a mean value of 13.5 transitions (SD = 6.07) for the intervals of participants experiencing attentional tunneling.

The statistical approaches for analyzing workload data have been further employed within the evaluation of the tunneling measurements. As a result, an initial model was fitted and tested for homoscedasticity. The model did not pass the test and consequently, the logarithm of the transition frequency was taken as a dependent variable. The outcomes of the resulting model can be viewed in Table 4.7. No statistically significant changes have been detected across the transition frequencies between participants experiencing tunneling states and those who did not. It is notable to mention, that the results from the initial model before the logarithmic transformations didn't showcase significant outcomes either.

Fixed Effects					
	Estimate	**SE**	**97.5% CI**	**t**	**p**
Intercept	2.72	0.08	2.54	30.68	2.26e-13
Tunnel	-0.24	0.15	-0.54	-1.53	0.135
Random Effects					
			Variance	**SD**	
Participant (Intercept)			2.54	2.9	
Residual			0.54	0.06	
Key: p-values and confidence intervals have been calculated in R 4.3.1 using the lmerTest package with default settings.					
LMER equation: *log(Transition frequency) ~ Tunnel + (1 | ParticipantID)*					
Number of observations = 40, groups: ParticipantID = 12;					

Table 4.7: Liner mixed-effects model assessing the effect of tunneling states on transition frequency.

Tunneling and Mean Saccade Length

Within this experiment, participants not experiencing attentional tunneling displayed a mean saccade length of 46.58 mm (SD = 14.47), while results from participants encountering tunneling showcased a mean value of 53.06 mm (SD = 14.44).

The initial linear mixed-effects model for analyzing the mean saccade length during states of tunneling and no tunneling did not cover the assumption for homoscedasticity. As a result, a logarithmic transformation was applied. The produced model and its outcomes are displayed in Table 4.8. For this eye-tracking metric, the occurrence of attentional tunneling did not yield significant effects on the observed measurements.

Fixed Effects					
	Estimate	**SE**	**97.5% CI**	**t**	**p**
Intercept	3.79	0.07	3.94	48.72	3.32e-16
Tunnel	0.16	0.12	0.40	1.32	0.19
Random Effects					
			Variance	**SD**	
Participant (Intercept)			0.03	0.18	
Residual			0.08	0.28	
Key: p-values and confidence intervals have been calculated in R 4.3.1 using the lmerTest package with default settings.					
LMER equation: *log(Mean Saccade Length) ~ Tunnel + (1 | ParticipantID)*					
Number of observations = 40, groups: ParticipantID = 12;					

Table 4.8: Liner mixed-effects model assessing the effect of tunneling states on mean saccade length.

Tunneling and Entropy

As a final step in the statistical analysis of the eye-tracking data, entropy changes between states of tunneling and no tunneling were examined. The normalized results of our entropy measurements indicate a value of 0.45 (SD = 0,09) for data related to no tunneling and a value of 0.48 (SD = 0.04) for data depicting states of tunneling.

Similar to the evaluation of entropy in the context of workload, no logarithmic transformations were needed for the mixed-effects model for entropy, as testing the residuals indicated no heteroscedasticity. The model and its results are displayed in Table 4.9. As can be observed from the outcomes, no statistical significance was found in relation to entropy changes during states of attentional tunneling.

Fixed Effects					
	Estimate	**SE**	**97.5% CI**	**t**	**p**
Intercept	3.03	0.10	2.83	29.24	2e-16
Tunnel	0.18	0.21	-0.23	0.84	0.41
Random Effects					
			Variance	**SD**	
Participant (Intercept)			0.00	0.00	
Residual			0.3237	0.5689	
Key: p-values and confidence intervals have been calculated in R 4.3.1 using the lmerTest package with default settings.					
LMER equation: *Observed Entropy ~ Tunnel + (1 | ParticipantID)*					
F-statistic = 40, groups: ParticipantID = 12;					

Table 4.9: Liner mixed-effects model assessing the effect of tunneling states on entropy.

4.7 Machine-Learning Classification

With the objective of accurately predicting workload and tunneling states, a performance evaluation for three machine-learning approaches has been carried out. The classification of workload and tunneling states has been employed separately.

Workload Classification

Data used in the workload classification has been categorized using the same approach as within the statistical analysis - eye-tracking data from the *workload* and *combined* conditions has been classified as high workload, whereas data originating from the *baseline* and *performance* conditions has been classified as low workload.

As mentioned in the background chapter, based on findings from studies utilizing machine learning pipelines for workload or tunneling classification and discussions with professionals in the field of machine learning, we selected SVM and logistic regression as appropriate methods for our purposes. All machine learning approaches have been executed using the scikit.learn[1] library in Python.

The implemented data preparation pipeline is based on tutorials by Hailat [92] and Brownlee [93]. As one of the first steps, the order of the data was shuffled to avoid ordering effects and to improve generalizability. Following this, the data was transformed using a centralization method, due to the differences and unequal scaling within the numerical values of the three predictor variables: transition frequency, mean saccade length, and entropy. This was achieved by subtracting the mean of each eye-tracking measurement from each data point within this measure. As the next and final step in the centralization, the result of the subtraction was divided by the standard deviation. The described process of data preparation was applied to all algorithms. Before the final performance cross-validation for each algorithm, an exploratory step of individually training and testing the algorithm was employed, in order to investigate the individual predictions in detail. To accomplish this, the data was split into a training and a testing set. 80% of the data was set aside for training, whereas 20% was utilized for testing, resulting in 568 and 143 data points, respectively. The final step, in examining each algorithm involved a 5-fold cross-validation investigating the accuracy and precision of the pipeline.

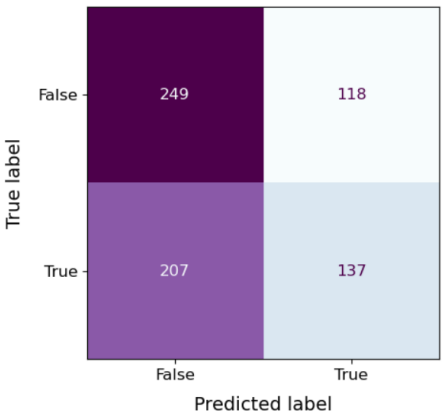


Figure 4.9: Example confusion matrix from the 5-fold SVM cross-validation trained on workload data.

[1] https://scikit-learn.org/stable/, last accessed: Dec. 2023

The SVM classifier was fit utilizing a linear kernel and a regularization parameter with the default value of 1. This value was selected after testing multiple lower parameters, which resulted in worse performance. As mentioned earlier, in addition to performing a 5-fold cross-validation, the model was individually fitted per hand to investigate its predictions. The outcomes from both the separate trainings and the 5-fold cross-validation showcased mean values of around 51% for accuracy and 40% for precision when applied to the testing dataset. An example of the predictions can be observed in the confusion matrix in Figure 4.9.

Logistic regression with a "lbfgs" solver was the next algorithm that was evaluated. Once again the regularization parameter was set to 1 and a "l2" type of penalty was applied. Similar to the outcomes of the SVM classifier, the individual trainings and the cross-validation resulted in comparable values. Both the precision and accuracy of the logistic regression showcased a mean value of around 51%. The outcomes are illustrated in the confusion matrix in Figure 4.10.

As a final step in our attempt to identify a suitable classifier for predicting workload states based on eye-tracking metrics, the TPOT automated machine learning tool was used. This tool consists of a Python library that simultaneously tests multiple machine-learning pipelines on data provided by the user and proposes the best-performing approach. TPOT's configurations have been set to 5 generations and 100 populations. This selection would result in iterating 5 times an optimization process on 100 pipeline suggestions, ultimately generating and assessing 500 recommendations. Based on our dataset the suggested pipeline consisted of a Bernoulli Naive Bayes excluding prior fitting and with an alpha value of 0.01 in combination with a Radial Basis Function (RBF) Sampler with gamma equal to 0.1. The suggested pipeline resulted in mean values of 55% for accuracy and 53% for precision across the different performance assessment methods. The confusion matrix illustrating the prediction outcomes is displayed in Figure 4.11.

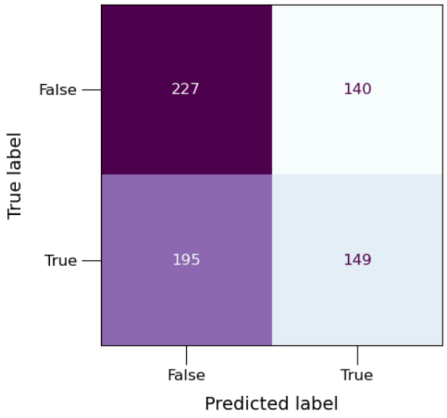


Figure 4.10: Example confusion matrix from the 5-fold logistic regression cross-validation trained on workload data.

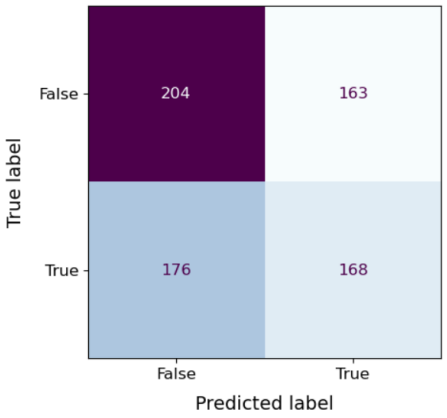


Figure 4.11: Example confusion matrix from the 5-fold Bernoulli Naive Bayes cross-validation trained on workload data.

Tunneling Classification

During the evaluation of the machine-learning approaches for tunneling classification, the same techniques for data preparation have been utilized as described in the previous section. However, due to the smaller amount of data related to attentional tunneling, the exploratory individual trainings included a set of 32 data points for the training dataset and 8 data points within the testing dataset.

Mirroring the approach used within the workload classification, the SVM tunneling classification was employed with a linear kernel and a regularization parameter equal to one. Both the exploratory training and testing of the SVM classifier and the cross-validation assessment indicated a relatively high accuracy score of around 75%. However, the algorithm's precision resulted in remarkably low scores of around 30% during individual training and testing, and a mean precision score of 0% for

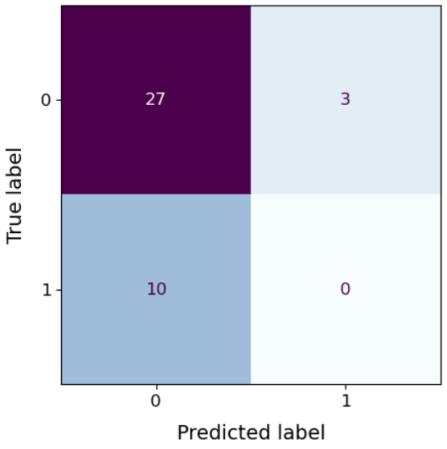


Figure 4.12: Example confusion matrix from the 5-fold SVM cross-validation trained on tunneling data.

the 5-fold cross-validation. By investigating the individual values of the predictions and refitting the model with new combinations of training and testing data, it was evident that due to the limited tunneling data points, the SVM algorithm with linear kernel always predicts a state of no attentional tunneling. This can be observed within the confusion matrix in Figure 4.12. Moreover, decreasing the number of folds within the cross-validation to 3 increased the precision score to 0.3%, suggesting that 5-fold cross-validation might not be suitable for this unbalanced and limited dataset.

The logistic regression model was fit using an "l2" penalty, a regularization set to 1, and an "lbfgs" solver. The results of this pipeline aligned entirely with the observed outcomes from the SVM approach. The cross-validation of the logistic regression resulted in a mean accuracy score of 72% percent and a precision of 0%. An example of the predictions from the 5-fold cross-validation can be observed in Figure 4.13.

The last classification approach, once again, involved utilizing the TPOT pipeline generator with the previously described specifications. The first recommendation suggested applying a Multi-Layer Perceptron (MLP) classifier neural network. However, this approach led to overfitting to the train data, which resulted in a prediction accuracy of 100% on the train set and an accuracy of 30% on the test set. The TPOT specifications were then adjusted and the population size was reduced to 50. Similar to TPOT's suggestion for our workload data, a Bernoulli Naive Bayes with an alpha value of 0.01 and no prior fitting was recommended. With a mean accuracy of 65% and a precision of 45% during the 5-fold cross-validation, this classifier demonstrated a significantly improved performance compared to the SVM and logistic regression. The individual training and tests indicated values of around 70% for accuracy and 59% for precision. An example confusion matrix can be observed in Figure 4.14.

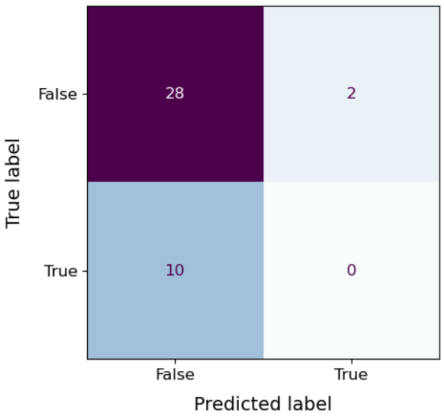


Figure 4.13: Example confusion matrix from the 5-fold logistic regression cross-validation trained on tunneling data.

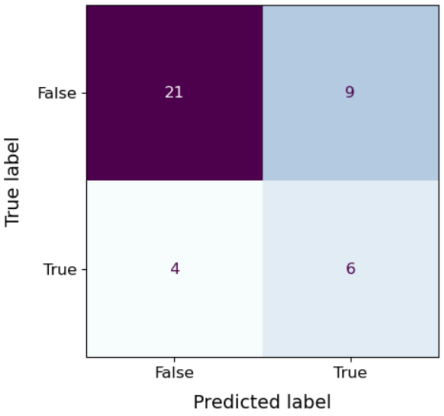


Figure 4.14: Example confusion matrix from the 5-fold Bernoulli Naive Bayes cross-validation trained on tunneling data.

5 Discussion

This study aimed to explore opportunities for inducing high workload and attentional tunneling in a cockpit simulator and to detect the two cognitive states using eye-tracking data. This was achieved by focusing the pilot's attention on a main task and testing their awareness and correspondingly tunneling state through unexpected events. Workload manipulation as one of the tunneling triggers was executed by assigning a secondary n-back task under certain conditions. An additional approach to inducing tunneling states was attempted by imitating negative performance feedback related to the focus task through an audio signal playing at random intervals. The measurements utilized in the gaze analysis included transition frequency, mean saccade length, and entropy.

The following text discusses the results from the previous chapter by comparing them with our hypotheses, the existing literature on the topic, and our research questions. Furthermore, recommendations and suggestions for future studies are considered. Results related to participants' performance within the secondary task are discussed in the first section, as they relate to the evaluation of the chosen workload-inducing factors.

5.1 Workload in Experimental Conditions

Validation of the Workload Manipulation

To obtain a comprehensive overview of the effects of incorporating an n-back task as a workload adjustment factor, participants' self-assessed workload was collected every two minutes during each run, using the temporally sensitive ISA technique, as well as at the end of each flight-session employing the common NASA-TLX questionnaire.

Analysis of the results from both measurements showcases similar outcomes. Using a cumulative mixed model, fitted with participants as an intercept randomizing factor, the effect of the independent variables on the NASA-TLX was explored. Results from the analysis indicate that the n-back task significantly increased the experienced workload, whereas negative performance feedback doesn't showcase an effect. This aligns with our expectations since the performance variable was employed as a tunneling triggering factor, expected to induce ego-threatening reactions that were not assumed to influence workload. Since literature research suggests that workload is an effective approach for inducing tunneling states, the n-back task has been applied to support both topics of this research. Interestingly, the NASA-TLX analysis indicated a significant decrease in the effect of the n-back task during the *combined* conditions, when performance feedback was present in addition to the secondary task. During conditions including audio feedback, participants often mentioned an irritation from the audio signal, as they could not determine how their performance was being evaluated. Sometimes they pointed out that they ignored the incoming sounds. This could potentially have led to some sort of contempt towards the system and consequently an effect on the perceived workload. However, the significance of the interaction between feedback and the n-back task is relatively low, so no conclusive statement can be based on this finding.
Comparing these results with outcomes from the same statistical approach employed on the ISA self-assessment indicates similarities on all levels except the interaction between fixed effects. Analogous to the NASA-TLX, the n-back task influenced the reported workload significantly, while performance feedback did not indicate significant effects. However, unlike the NASA-TLX results, no significant interaction between the n-back task and performance feedback could be detected. This contrast between the two results combined with the temporal differences between the ISA and NASA-TLX techniques may be interpreted as the effect of performance on the general impression of a run. As stated before, however, we do not consider these findings to be conclusive indications of how performance feedback impacts the efficiency of an n-back task.

Results from both workload self-assessment measurements support the statement that n-back tasks affect the perceived workload by intensifying it. These findings align with previous research testing the relationship between scenarios incorporating n-back tasks and the ISA reporting technique [12], as well as studies demonstrating the effect of n-back tasks on the NASA-TLX measurement [81]. Furthermore, our findings related to the effect of negative audio performance feedback on workload also match results from existing research [94]. However, to the extent that our literature research covered, we could not discover previous work indicating interactions between performance feedback and n-back tasks. This finding offers an interesting topic for future work in this field.

Overall, relating our analysis to the first research question, our results suggest that workload manipulation within the experimental conditions of an aircraft simulator is possible. By applying an n-back task in combination with a *baseline* scenario both high and low workload states can be achieved. Furthermore, the flexibility of n-back tasks offers opportunities for future research to experiment with multiple n-back levels and to explore the possibility of detecting multiple levels of workload.

Workload Fluctuations within a Flight

In addition to our main analysis, which focused on the effects of the selected independent variables on the experienced workload, our data indicated some additional observations. As earlier mentioned, we noticed changes in the reported workload throughout the different temporal phases of the flight. Our analysis indicated a correlation between the moment in which the participant has been asked to assess their current workload and their response. Although no detailed statistical analysis of the direction in which the response changes has been initiated, this is a valuable observation relevant for future studies in the field. This finding suggests that the careful selection of a time frame for the eye-tracking analysis is of high importance. Especially in studies utilizing self-assessment measurements that are not time-sensitive, such as the NASA-TLX, the possible differences related to the flight phase, task type, or unexpected situational variations need to be taken into account when defining a time window for the eye-tracking analysis, in order to avoid discrepancies in the data and misleading statistical outcomes.

This finding is not unique in its nature or unexpected, as it has been supported by previous studies [95] and it illustrates the general knowledge that landing and take-off involve considerably more tasks and consequently require more cognitive capacities [96].

N-back Accuracy and Workload

As part of the data related to the workload-inducing conditions, participants' performance on the n-back secondary task has been recorded and analyzed. Given that the utilization of an n-back task as a workload manipulation technique is based on Wicken's Multiple Resources Theory [23] and functions by overloading participants' cognitive capacities, it is assumed that there is an inherent correlation between n-back accuracy and reported workload. The connection between the two factors manifests itself in the decrease in performance accuracy during increases in workload.

This assumption has been supported by our results, which showcase a significant correlation between the reported ISA workload and the percentage of errors. As mentioned in the Experiment chapter, two measurements were recorded in relation to n-back accuracy - the percentage of missed targets and the percentage of false alarms. While both measurements indicate a correlation, only the visualizations of the missed items demonstrate a clear and gradual increase in the percentage of error with the increasing workload self-assessment. This does correspond to assumptions from other experiments stating that false alarms and missed percent originate from different cognitive processes [40].

It is relevant to mention that our documented NASA-TLX scores did not showcase a correlation between the amount of error and the reported workload. Since the n-back task was employed only in two conditions and at the same 2-back level, these results are not surprising as there was not enough workload variance between the two conditions. This observation is additionally supported by our statistical analysis of the relationship between the independent variables and accuracy, which resulted in non-significant values. Although results from the NASA-TLX analysis in the previous subsection suggest a significant interaction between the audio performance feedback and the n-back task on the reported workload, our analysis of the error rates does not reveal a significant difference between

conditions with and without performance feedback. This finding further underlines the statement that the results indicating significant interactions between independent variables should be interpreted with caution.

Finally, we can convey that our results align with existing research and the expected outcomes. Some suggestions for future research would be to utilize multiple levels of n-back tasks, in order to meaningfully investigate the relationship between NASA-TLX and participant accuracy.

5.2 Tunneling in Experimental Conditions

A further relevant topic, central to our research questions, is whether states of attentional tunneling can be induced among pilots as a participant group and within a cockpit simulator environment. The proposed experimental approach, consisting of tunneling triggers and tunneling determinants, has been analyzed by comparing the outcomes of two participant groups - one experiencing a predefined combination of tunneling triggers and tunneling determinants and another one presented with a set of randomized combinations. As previously stated, this strategy was chosen due to the limited participant pool, which would not have allowed us to draw statistically relevant conclusions if all combinations had been randomized. Consequently, we are basing our interpretations on a descriptive evaluation of the outcomes. Here, it is relevant to underline that the following conclusions are based on a limited amount of data and should be regarded with caution.

As mentioned in the previous chapter, our experiment resulted in 16 cases of attentional tunneling and 32 cases indicating no irregular attentional states. Separating these results based on the two participant groups reveals an outcome of 11 cases of tunneling within the non-randomized group, which consisted of 32 observations, and 5 cases of tunneling in the randomized group, comprising 16 observations. Converting these results in percentages indicates a tunneling occurrence of 34% within the non-randomized group, compared to 31% of tunneling instances among the randomized participant group. The observed outcomes suggest no substantial differences between the two groups.

To better understand our results, we have compared the amount of tunneling occurrences between the two groups based on the separate determinant events. As can be seen within the outcomes, for two of the trigger + determinant combinations, almost no difference in the frequency of tunneling occurrence between the randomized and non-randomized groups is evident. These observations encompass the lower ECAM and balloon determinants and accordingly the combinations of lower ECAM + *workload* condition and balloon + *combined* condition. Within our experiment, it was evident that the balloon as an abnormal event, was easily detected by our participants, only one of which didn't notice the passing by object. Even though the balloon was presented during the *combined* condition, which was presumably the most demanding one due to the presence of two tunneling triggers, this determinant did not seem to challenge participants' attentional capacities. During the initial trial of the experimental design, the testing pilot commented, that it is highly implausible to observe unexpected vehicles or objects on the route during real-world flights. Hence, we assume that the improbability of the event, combined with its dynamic nature, and its size upon nearing, make this determinant rather unsuitable for experiments investigating pilots' attentiveness.

The lower ECAM page switch, on the other hand, appeared to be a more challenging approach, resulting in a 50% detection rate in both participant groups. This balanced outcome suggests that this determinant has the potential to be a good indicator of the awareness of pilots. This statement can be further supported by the fact that the information on the lower ECAM is considered of high significance, as its different pages communicate the current health of relevant aircraft components, such as the engine, pressurization, and more. Therefore, changes on this display that remain unseen by a pilot, could have detrimental outcomes, especially in complicated situations.

Unlike the observations from the balloon and lower ECAM determinants, the comparison of tunneling occurrences between the randomized and non-randomized groups indicated that the proposed combinations for the RA1 warning and the ILS button flicker did not result in the expected outcomes. As previously mentioned, it was assumed that the appearance of the RA1 determinant would be difficult to notice due to its small size, static nature, and distance to the main display (the PFD). It was therefore combined with the *baseline* scenario, which was assumed to be the least demanding one. However, our outcomes indicate that this predefined combination allowed the majority of par-

ticipants to notice the warning, suggesting that it wasn't challenging enough as a tunneling identifier. A possible explanation for this tendency could lie behind the relevancy of the display that showcases the RA1 determinant. Since the upper ECAM display is the location, where almost all warnings are showcased, pilots presumably are well-trained in monitoring this part of the aircraft. Furthermore, results from the randomized group imply that assigning more demanding conditions to the RA1 determinant increased the difficulty of detecting the warning. Although this observation is based on a very small amount of data, the outcomes combined with the relevancy of the ECAM display could be an indicator that warnings in this area are relatively easy to spot by pilots and should be incorporated carefully as tunneling determinants.

Conversely, the comparison of outcomes related to the ILS button flicker indicated that combining the determinant with the *performance* condition challenged participants to a considerable degree, whereas incorporating the flicker with other conditions resulted in a higher reaction rate. These results could be an indicator that the performance feedback had a tunneling effect on participants. Since our experiment design could not provide participants with real live feedback on their performance due to technical limitations, multiple participants expressed irritation with the suggested alternative consisting of a randomized audio alert, which did not behave as they expected. As a result, some of the participants indicated that they experimented with multiple approaches to improve their performance. This unexplainable behavior could have had as a consequence a cognitive narrowing effect similar to those related to unsolvable situations. Although within this study a robust and statistically relevant analysis of performance feedback as a tunneling trigger was not feasible due to the limited amount of data available, the attained outcomes suggest that an explicit study focusing on tunneling triggers and involving a larger participant pool should be conducted to better understand the different approaches and their efficiency.

Existing studies on cognitive narrowing in the cockpit frequently identify attentional deficits by testing pilots' ability to notice static or moving objects on the runway during landing. With the aim of expanding existing methodologies, our experimental design proposes an alternative approach based on irregularities in the behavior of cockpit instruments and unexpected visual cues in-flight. Similar approaches have been implemented in studies in other fields such as an experiment by Regis et al., which based the classification of participants' attentional state on whether they noticed an interface warning or not [44]. In line with findings from this experiment, our outcomes indicate multiple instances of attentional deficits, even though our participant pool was well acquainted with the interfaces and expected behavior within a cockpit, which eliminated design characteristics, confusion, and lack of knowledge as possible causes for overlooking the events. Although the earlier mentioned assessment of tunneling occurrence based on participants' reaction to runway incursions has successfully been applied in multiple studies, this approach might be unsuitable for experiments utilizing eye-tracking measurements due to the specific tasks related to the landing phase and their potential impact on the pilots' ocular behavior [18], [36]. Hence, the outcomes of our study complement the existing methodology and expand the opportunities for future experiments using eye-tracking data. Furthermore, due to the variety of proposed tunneling determinants, our approach is suitable for experiments with a smaller participant pool, that employ a within-subjects experimental design.

As mentioned in the literature research, another interesting method applied in studies on cognitive narrowing in aviation encompasses the alternative interpretation of attentional tunneling as the persistence of an erroneous decision by disregarding other solutions or possible explanations. An example of a possible implementation of this was executed in a study by Iani et al. investigating the effects of a 3D display on decision-making [49] and another one by Dehais et al. focusing on perseveration syndrome [11]. Both of these studies based the assessment of a pilot's cognitive state on whether they continued their current flight path, which involved entering a hazardous weather condition, or they chose to adjust their path and avoid the upcoming weather difficulties. Although this approach offers interesting opportunities for insights and avoids the ocular discrepancies potentially arising by incorporating visual events, a wider variety of decision-based scenarios is needed, in order for it to be applicable for experiments with a smaller participant pool and a within-subject design. Consequently, a potential topic for future research could be the development of multiple scenarios for the identification of decision-based attentional tunneling, also known as perseveration syndrome.

Following the completion and assessment of our experiment, further ideas for experimental designs have been gathered. For example, an approach that has been commonly implemented in other fields involves presenting participants with multiple tunneling determinants and identifying attentional tunneling based on the amount of noticed elements. This has frequently been achieved using the MATB tasking framework [14], [45], however, it can also be integrated into a cockpit using the tunneling determinants proposed in this study. For example, the ILS button flicker could be presented multiple times within a run for a shorter period of time. This determinant could, additionally, be combined with the ECAM page switch, thereby increasing the amount of focus required from the participant's side to successfully perform. On the one hand, this approach of incorporating multiple determinants can be useful as it could provide several data points indicating tunneling occurrence within a run. On the other hand, though, it imitates artificial tasks, untypical for a cockpit environment, raising the question of whether an implementation in a simulator is needed for this type of experiments or whether a simple desktop task could be sufficient. A possible approach that mitigates these unwanted effects, while also potentially increasing the volume of generated tunneling data within a run, could be to limit the number of presented events to a moderate amount and to incorporate for example, two to three determinants per run.

In general, comparing our outcomes to the research question of whether attentional tunneling can be implemented in experimental conditions, we can conclude that, since within this study multiple cases of tunneling could be observed, the induction and detection of tunneling states in a simulator is feasible. Even if the comprehension of what defines an occurrence of attentional tunneling is of a complicated nature and could be interpreted in different ways, the examples provided within this study align with previous work from both aviation and other branches. Unfortunately, due to the small sample size, no insights on the efficacy of workload and ego-threat as tunneling triggers could be drawn. However, the initial goal of tunneling induction has been achieved, suggesting that workload and performance feedback could be a good starting point for future work.

5.3 Ocular Behavior and Workload

As part of this experiment, changes in eye behavior under conditions of high workload have been estimated based on transition frequency, mean saccade length, and entropy as eye-tracking measurements. Following our hypotheses, an increase in transition frequency, a decrease in mean saccade length, and an increase in entropy were expected during high workload conditions. Our results, however, indicate statistically significant changes in the transition frequency and mean saccade length, whereas entropy data revealed no significance.

Contrary to our expectation of an increase, a significant decrease in the transition frequency during high workload phases has been detected. Within our results, the decrease has been calculated to be 11% under high workload conditions. Although our hypothesis has been based on the results from previous research employing this measurement [30], [56], our outcomes contradict those studies. However, they do follow the assumption of the earlier mentioned Multiple Resources Theory, which states that during an overload of the available resources, a more task-oriented, efficient, and less random behavior can be expected as a reaction [23]. In their study, Moacdieh et al. underline themselves that the observed significant increase in the transition frequency was unexpected and was the only measurement showing an increase in inefficient behavior under high workload [56]. When comparing our results with these studies, it is relevant to mention that in their experiment Moacdieh et al. calculated the transition frequency per second, whereas Faulhaber et al. estimated the transitions between the cockpit instruments and the environment outside the window in a one-minute interval. Therefore a possible explanation for the difference between our findings and the ones from these experiments could lie in the different time-frame. Furthermore, the specific design of the high workload scenario within the experiment by Faulhaber et al., which investigated participants' behavior during Single-Pilot operations, could be considered an influencing factor on the observed increase in transitions. Since the high workload condition consisted of an unexpected engine failure, which required a particular procedure, the differences between the eye movements in the baseline and abnormal scenario could be explained by the tasks involved. Consequently, we could summarize that our results align with the

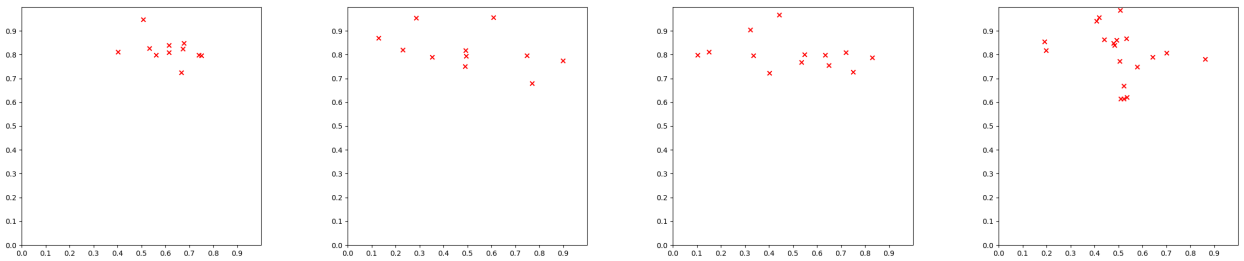


Figure 5.1: An example plot of a participant's fixation coordinates: ISA = 3; ISAnr. = 5; condition = *baseline*.

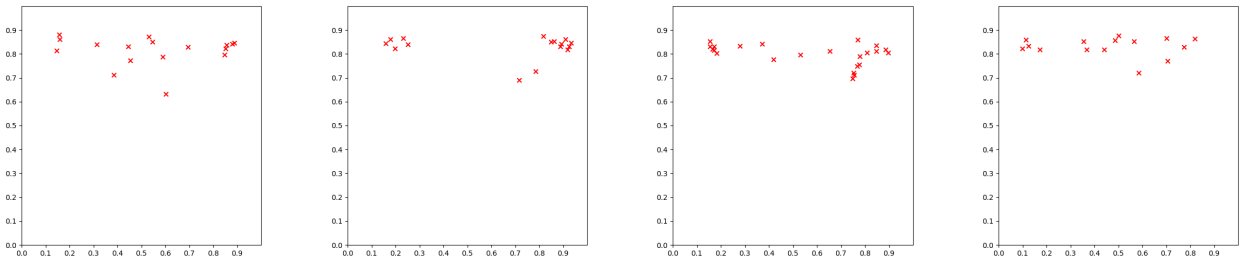


Figure 5.2: An example plot of the same participant's fixation coordinates: ISA = 4; ISAnr. = 5; condition = *workload*.

theoretical concepts relating to cognitive changes under high workload but contradict the outcomes from some of the previous research.

Nonetheless, the study's outcomes in relation to the mean saccade length confirmed our hypothesis, by showcasing a decrease by 7% during phases of high workload. This finding additionally reinforces the outcomes related to the transition frequency by supporting the assumption that high workload stimulates a more efficient and goal-oriented behavior. Our findings indicating a decrease in the mean saccade length are consistent with the results from previous studies measuring the central tendencies of saccades [56], [14]. Interestingly, even though both eye-tracking measurements showcase a strong significance of $p < 0.05$, a comparison of the statistical results from the mean saccade length and transition frequency suggests that workload has a stronger effect on transition frequency than on saccade length. This observation is noteworthy since previous studies consistently indicate that saccade size is a strong predictor for workload levels, whereas transition frequency is less commonly applied and results from prior research demonstrate the opposite of our findings. With this in mind, our outcomes might indicate the need for further investigations of measurements related to the quantification of transitions and the potential of these metrics.

As previously mentioned, entropy - the last gaze measurement used in this study did not showcase significant effects. Even though the statistical tests on entropy have been applied directly to the observed values and no logarithmic transformations were needed, the outcomes suggest a considerably insignificant increase in the documented values. Although this measurement has demonstrated significance in multiple previous studies, indicating an increase with growing workload [61], [62], it is generally considered a relatively unstable metric [26], since contradicting outcomes are frequently reported [56]. Considering that the entropy measurement is dependent on two characteristics - the chosen temporal duration and the selected bin division of the AOI, adjustments in these specifications could improve the outcomes and result in new findings. Within this study the PFD display, used for the entropy calculations, has been divided into a 10x10 grid, resulting in bins of circa 1.7 cm. Other studies either use similar grids separating the space in bigger bins, relative to the contextual information on the display [57], [56], or divide the space into fine-grained patterns with sizes of around a pixel [61]. Exploring the fine-grained visual selection patterns of participants might unveil additional insights from this study, therefore an approach, that we would like to apply to our data in

future iterations would be to follow a smaller grid, utilizing the specific fixation coordinates, rather than larger bins. Another insight that we noticed by comparing participants' fixation locations within the 30-second interval that has been utilized for the eye-tracking measurements in this study, is that this time interval might be too short to draw relevant conclusions on changes in entropy. As can be observed in Figures 5.1 and 5.2, which depict the fixation positions of one participant across the different 30-second intervals between two ISA questions, the amount and position of the fixations fluctuate between the different intervals, regardless of the reported ISA self-assessment and the condition. This could be an indicator that entropy is a measurement more suitable for data depicting longer periods. In their literature review, Shiferaw et al. propose a good overview of studies utilizing entropy and the temporal duration that has been used [57]. Within this overview, it can be observed that there are multiple studies incorporating short durations such as 5 to 30 seconds, however, those studies are mainly related to the free-viewing of pictures or paintings. Experiments investigating more complicated activities, such as surgical or flight tasks, have calculated entropy within longer periods exceeding 2 minutes, emphasizing the assumption that a 30-second interval might be insufficient for entropy insights related to pilot activities.

5.4 Ocular Behavior and Tunneling

Utilizing the same eye-tracking metrics the relationship between attentional tunneling and gaze behavior has been tested. The initial hypotheses expected our data to indicate a decrease across all metrics, including transition frequency, mean saccade length, and entropy. However, our results did not detect any significant effects of tunneling on the eye movements of participants.

Due to the small amount of experiments focused on attentional tunneling, a comprehensive comparison with the literature review for each measurement is challenging. As mentioned within the Background chapter, switching rate has been the only measurement related to transition frequency that our literature research could identify as being successfully employed in previous work on the topic. In a study by Regis et al., switching rate has displayed a significant decrease during tunneling states [44]. Similarly, although insignificant, the results in our data also indicate a change of the frequency in a decreasing direction. Notably, within the study by Regis et al., the switching rate is evaluated by calculating the transition frequency within a 10.5-second interval and then generating an estimation of the transition frequency for a 1-minute interval based on the documented 10.5 seconds. Even if the switching rate and transition frequency are estimated differently, they represent the same measurement and constitute a very similar time interval. Therefore, we assume that the chosen parameters for our calculation are suitable and a possible explanation for the inconclusive results of our analysis might lie in the limited available data.

Although a previous study has showcased a significant decrease in the average saccade length during states of tunneling [14], our outcomes did not detect any significance in the data. Furthermore, our statistical tests indicate an insignificant increase in the saccade length during tunneling. This difference in the direction is hard to interpret since our data is insignificant and the literature research did not lead us to other studies employing saccade sizes in the context of attentional tunneling. On the one hand, a decrease in the mean saccade length is logical, since the lack of appropriate monitoring is characteristic of states of attentional tunneling. However, larger saccade sizes could be an indicator of a more chaotic visual search strategy. Further research utilizing saccade length as a measurement is needed to better understand these findings and gain insights into the processes behind them.

To the best of our knowledge, no previous studies have utilized entropy in the context of attentional tunneling. However, other dispersion metrics, such as the NNI [14], and vertical and horizontal standard deviation of fixations [21], [63] have previously been employed and have indicated significant differences in the fixation patterns. Nevertheless, the outcomes from the different studies are inconsistent and showcase contradicting results. Unfortunately, our experiment could not contribute to clarifying the discrepancies within previous findings. However, for future iterations and work utilizing this metric, the same recommendations and approaches as described in the discussion of entropy in relation to workload can be applied here.

Since none of the eye-tracking metrics in this study displayed significance in relation to attentional tunneling, this might be an indication of an issue in the analytic approach. A possible explanation

could be, firstly, that due to the limited data points related to tunneling states, no meaningful comparison between the occurrence and absence of tunneling could be estimated or that generally the amount of data was insufficient. To improve our evaluation and potentially overcome the restricted data volume, within the next iteration, we intend to generate multiple time windows within a run. Subsequently, we will initiate a new analysis of the generated time frames comparing them between participants experiencing tunneling and not. This method would be similar to the study by Regis et al. [44], who separated their experiment into phases and compared the corresponding changes between the ocular behavior of participants in tunneling and non-tunneling states.

Furthermore, a substantial difference between our approach to selecting the time window for our data and the methods used in other studies is that we have selected a 30-second interval before the onset of a tunneling identifying event, whereas other studies commonly utilize an interval after the onset of a tunneling identifying event [44], [14]. However, we assume that the sole occurrence of a tunneling determinant inherently forces a difference between the eye-tracking data of participants in tunneling and non-tunneling states. For example, participants not experiencing tunneling will certainly glance at the determinant, thereby increasing their transition frequency by two. Additionally, since the ultimate goal of experiments on attentional tunneling during safety-critical activities is preventive and aims at the early recognition of attentional deficits, studies should be able to provide insights into the general changes in ocular behavior during such states. Therefore, we consider the proposed analysis of time slots before the appearance of a tunneling determinant a more suitable approach.

5.5 Machine Learning for Workload and Tunneling Classification

As a final step in the examination of our outcomes, the data was used to train and test three machine-learning pipelines. The selected algorithms were tested first in classifying states of low and high workload, and subsequently in recognizing the occurrence of tunneling and non-tunneling states. As mentioned in the Background chapter, a score of 70% or higher for both precision and accuracy will be taken as a minimum score and a reference point indicating fair performance.

The performance of the SVM classifier showcased a mean accuracy of 51% using 5-fold cross-validation on the workload dataset. This is a similar outcome to some of the binary classification experiments on workload summarized in the literature research by Kaczorowska et al. [67]. However, the majority of the reported results achieve an accuracy above 80%. A binary classification problem with an accuracy of around 50% combined with similar results for precision is generally considered a low value due to the small number of classes the algorithm needs to predict. Compared to our outcomes, the experiment by Kaczorowska et al. showcases a much better accuracy rate of 97%. However, their SVM algorithm, was trained utilizing 7 different eye-tracking metrics and 3 different classes - low, medium, and high workload. This could suggest that a possible approach for improving the performance of our algorithm could be to include further eye-tracking measurements in the dataset.

The logistic regression classifier within our study resulted in approximately the same mean accuracy as the SVM model. Moreover, similar to the comparison of the performance of our SVM model with the one from the experiment by Kaczorowska et al., the logistic regression accuracy within our experiment showcases considerably lower performance. Therefore, the same recommendations as stated in the context of the SVM method can be incorporated here to improve the logistic regression outcomes.

As mentioned in the results section, the last pipeline tested on our dataset included the Bernoulli Naive Bayes combined with an RBF Sampler, which was recommended by the TPOT pipeline generator tool. The resulting mean values of 55% for accuracy and 53% for precision, although slightly better, are similar to the outcomes from the SVM classifier and the logistic regression. However, the recommended RBF Sampler presented us with the idea of retraining the SVM classifier using an RBF kernel, instead of a linear one, which increased the mean precision score of the cross-validation with 13%, whereas the accuracy improved with around 3%. This suggests that the data represents rather complicated relationships that don't follow a linear separation. Since none of the workload classification models achieved satisfying results, although the amount of data could be considered sufficient, further ideas for improving the models have been gathered. For example, since some correlations between the reported workload and the flight phase have been identified, a possible reason for the low classification accuracy could be that the dataset includes multiple levels of workload exceeding the

two low and high classes. Therefore, follow-up work on this study will explore training the classifiers using several of the levels provided by the ISA workload self-assessment.

The training of the SVM algorithm, logistic regression, and Bernoulli Naive Bayes into classifying states of tunneling and non-tunneling resulted in accuracy outcomes slightly better than those of the workload classification. However, the precision scores were substantially poorer. During the individual training and testing sessions for each model, large fluctuations in the accuracy results could be observed. It was strongly evident that due to the small sample size of 40 data points, and the substantially fewer data representing states of tunneling, the models showcased a strong bias. Essentially both SVM and logistic regression often resulted in a relatively high accuracy score of around 70%, but plotting the predicted values revealed that the algorithms assigned only one class - a state of no tunneling. Interestingly, although both SVM and logistic regression regularly predicted only one class, the Bernoulli Naive Bayes suggested by the TPOT pipeline generator, consistently resulted in more varied classifications. Furthermore, due to the unbalanced dataset, a 5-fold cross-validation proved unfeasible for the SVM classifier and the logistic regression. The resulting precision score of 0% indicated that the data subsets were insufficient for training and testing the models. This became apparent as decreasing the amount of folds improved the precision score. In this context, the Bernoulli classifier, once again, outperformed SVM and logistic regression by achieving scores above 0% even with the utilization of a 5-fold classification. With considerably higher mean scores for precision (45%) and similar outcomes for accuracy (65%) compared to the mean values of around 75% for accuracy and 0% for precision among the logistic regression and SVM classifiers, the results of the Bernoulli Naive Bayes suggest that, for future iterations, this pipeline could prove to be appropriate. Although the majority of improvement suggestions mentioned in relation to the workload classification problems can be applied here as well, the main drawback in the tunneling classification pipeline can be attributed to the limited available data.

Potential approaches for improving the performance could include the utilization of more sophisticated models. For example, in their study, Regis et al. utilized an Adaptive Neuro-Fuzzy Inference System (ANFIS) algorithm, which resulted in better performance, although the number of participants was similar [44]. It is also relevant to mention the experiment by Berthelot et al., which similar to our study included data from 10 participants and employed traditional algorithms such as SVM, kNN, and decision trees [45]. The results from this study showcased an accuracy of around 90%, which is substantially higher than the accuracy of our outcomes. A potential reason for the large performance difference might be attributed to the fact that the data within the study by Berthelot et al. included both a reference state and a state of tunneling for each participant. In contrast, our dataset comprises multiple participants whose data relates only to states of tunneling or only to states of no tunneling. Consequently, a revision of the dataset and the identification of non-tunnel states for some participants could be a further point of improvement. As discussed in the statistical analysis of our tunneling gaze data, further explorations of the data by experimenting with different time windows, similar to the approach used in the study by Regis et al., could prove to be beneficial and might contribute to expanding the dataset.

Unfortunately, none of the selected methods surpassed the minimum threshold of 70% set for both precision and accuracy. In relation to our research questions, these outcomes reject our hypotheses and indicate that the available data combined with the selected pipelines are not optimal at their current state. Future iterations could potentially improve the results of the two classification problems by incorporating additional eye-tracking measurements. Moreover, the data of each of the classification problems can be revised and adjustments can be explored in later studies.

6 Conclusion

The primary aim of this thesis has been to induce instances of attentional tunneling together with high and low workload states in a simulator experimental environment. Additionally, this research intended to investigate the potential for classifying these states using machine-learning approaches trained on eye-tracking data.

Inducing Workload and Attentional Tunneling in Experimental Conditions

Following the findings from previous research on attentional tunneling, the induction of tunneling states has been achieved by incorporating workload manipulation techniques in addition to ego-threatening factors. Participant workload has been modified utilizing an auditory n-back task. By using a secondary auditory task, participants' cognitive capacities have been challenged, without hindering their abilities to perform their primary tasks, which involved manually operating the aircraft. Self-assessment techniques have been employed to evaluate participants' workload including the common NASA-TLX questionnaire at the end of each run and the time-sensitive ISA technique comprising a verbal response, which has been collected every two minutes. The results of the participants' reported workload demonstrate that high and low workload has successfully been induced by utilizing an n-back task as an independent variable. Furthermore, the analysis of participants' accuracy within the n-back task demonstrated a correlation between workload and accuracy, confirming the Multiple Ressource Theory [23], which forms the foundation of workload manipulation techniques. In addition to the workload-inducing conditions, an ego-threatening factor in the form of negative auditory feedback on a focus task has been implemented, in order to stimulate states of attentional tunneling. Based on the assumption, that pilots are well trained in manually flying a simulator and frequently showcase a determination to demonstrate their proficiency, negative performance feedback was introduced as an ego-threatening factor with the goal of triggering affective states with strong motivational intensity. Participants were assigned the focus task of flying the aircraft as closely as possible to a predetermined flight path and were informed that auditory feedback would be given, whenever their performance on the main task deteriorated. Although a live evaluation of participants' actual performance was not feasible in this simulator, negative audio feedback was presented at random intervals.

Overall, the outcomes of the study indicate that attentional tunneling has successfully been induced during one-third of the experimental runs. The occurrence of tunneling states has been assessed by examining participants' ability to notice abnormal events, including an ECAM Radio Altimeter warning with no sound, a flickering ILS button, a continuous switch between lower ECAM pages and a balloon passing by. The tunneling-determining events have been specifically chosen to represent multiple types of situations unfolding at different locations within the visual field, in order to explore the efficiency of the different approaches. Due to the limited participant sample size and the potential impact of the determinant type on the occurrence of tunneling, the effects of workload and negative performance feedback as tunneling triggers could not be statistically assessed. However, the descriptive findings from our data revealed that detecting a balloon as an abnormal event might not have posed a sufficient challenge to participants' attentional capacities, suggesting it may not be suitable as a tunneling determinant. Furthermore, our observations indicate that the majority of participants easily detected the RA1 warning. For this reason, if utilized, we recommend presenting ECAM warnings under more demanding conditions. Within this investigation the remaining two events - ILS button flicker and lower ECAM page switch, proved to be suitable tunneling determinants. They provided participants with a sufficient level of difficulty, while also representing relevant events in the cockpit.

In general, we consider the proposed experimental design to have met our expectations and would conclude that the induction of high, low workload, and attentional tunneling in the experimental conditions of an aircraft simulator has been successfully demonstrated. Furthermore, our work showcased instances of attentional tunneling by utilizing workload and negative performance feedback, suggest-

ing that these tunneling triggers could be a good starting point for future research. In general, this study has provided innovative suggestions on experimentally determining the occurrence or absence of attentional tunneling in cockpit environments together with actionable learnings for future work in this field.

Transition Frequency, Mean Saccade Length, and Entropy as Eye-Tracking Metrics for Cognitive State Classification via Machine Learning

The analysis of the impact of workload on the proposed eye-tracking metrics showcased a significant decrease in the transition frequency and mean saccade length under high workload conditions. While these outcomes contradict some previous studies and our hypothesis, that high workload increases the transition frequency, they do support the assumption that an overload of cognitive resources stimulates more efficient ocular behavior. This trend, however, could not be observed among the entropy measurements, which showed no significant differences between workload levels.

Interestingly, although the majority of the selected eye-tracking metrics showcased a significance between workload levels, the utilized machine-learning pipelines did not manage to successfully distinguish high and low workload despite the extensive data available. All tested models, including SVM, logistic regression, and Bernoulli Naive Bayes with an RBF sampler, resulted in 5-fold cross-validation values of around 50% for accuracy and precision, which is below the results commonly observed in other studies on workload classification. Furthermore, an accuracy of 50% for a binary classification problem suggests that the classifier is not reliable. These results decline our research question and indicate that more work needs to be invested in both improving the machine-learning pipelines and potentially the refinement of the workload data.

Conversely, during the statistical assessment of the eye-tracking metrics and their variance between instances of attentional tunneling and its' absence, no significant differences in the ocular behavior could be identified. Overall, the performance of the machine-learning pipelines employed in the context of attentional tunneling suggests that no adequate evaluation of the methods was possible due to the limited available data. However, although the SVM and logistic regression pipelines showcased a strong bias and poor precision in classifying tunneling states, the Bernoulli Naive Bayes exhibited promising results, suggesting it could be a potentially appropriate choice for future studies in the field. Even though none of the machine-learning pipelines achieved satisfactory performance levels, the positive findings suggest a new perspective for future research in the field. Furthermore, the gathered data and learnings will be leveraged to improve our approach during the following iteration.

Future Improvements

Although the experimental design has fulfilled our expectations and has positively addressed one of our research questions, the machine-learning approaches employed within the limits of this thesis have not proven to be fully successful. However, the learnings from this work will be utilized in an upcoming iteration attempting to improve the proposed methodology. In response to the observations from the workload classification problem, the dataset will be revised and potentially separated into multiple levels, based on the reported ISA self-assessment. Additionally, an exploration of the tunneling data will be attempted, aiming to expand the dataset by potentially identifying and comparing different states within the same participant.

Bibliography

[1] H. Kharoufah, J. D. Murray, G. Baxter, and G. Wild, "A review of human factors causations in commercial air transport accidents and incidents: From to 2000–2016," *Progress in Aerospace Sciences*, vol. 99, pp. 1–13, May 2018. doi: 10.1016/j.paerosci.2018.03.002.

[2] International Air Transport Association, "2021 Safety Report, Edition 58," Apr. 2022.

[3] M. R. Endsley, "Toward a Theory of Situation Awareness in Dynamic Systems," *Human Factors*, vol. 37, pp. 32–64, Mar. 1995. doi: 10.1518/001872095779049543.

[4] International Air Transport Association, "Loss of Control In-Flight Accident Analysis Report. Edition 2019. Guidance Material and Best Practices," 2019. Accessed: Dec. 2023. [Online]. Available: https://www.iata.org/contentassets/b6eb2adc248c484192101edd1ed36015/loc-i_2019.pdf.

[5] European Union Aviation Safety Agency, "Loss of Control in General Aviation." Accessed: Dec. 2023. [Online]. Available: https://www.easa.europa.eu/sites/default/files/dfu/Loss%20of%20Control%20in%20General%20Aviation%20-%20update%2017112016-%20sourcedoc-final.pdf.

[6] S. A. Shappell, "A human error analysis of general aviation controlled flight into the terrain accidents occurring between 1990-1998," 2003. Accessed: Dec. 2023. [Online]. Available: https://commons.erau.edu/publication/1217/.

[7] National Transportation Safety Board, "Aircraft accident report: Eastern Air Lines L- 1011, N310EA, Miami, Florida, December 29, 1972," Jun. 1973. Accessed: Dec. 2023. [Online]. Available: https://www.ntsb.gov/investigations/AccidentReports/Reports/AAR7314.pdf.

[8] Taiwan Transportation Safety Board, "Aircraft Accident Report: Crashed on a Partially Closed Runway during Takeoff, Singapore Airlines Flight 006, Boeing 747-400, 9V-SPK, CKS Airport, Taoyuan, Taiwan, October 31, 2000," 2002. Accessed: Dec. 2023. [Online]. Available: https://reports.aviation-safety.net/2000/20001031-0_B744_9V-SPK.pdf.

[9] Swedish Accident Investigation Authority, "Final report RL 2016:11e," Dec. 2016. Accessed: Dec. 2023. [Online]. Available: https://www.havkom.se/assets/reports/RL-2016_11e.pdf.

[10] Interstate Aviation Committee, "Final Report Tatarstan Airlines Flight 363," tech. rep., Nov. 2013. Accessed: Dec. 2023. [Online]. Available: https://mak-iac.org/upload/iblock/459/report_vq-bbn_eng.pdf.

[11] F. Dehais, C. Tessier, L. Christophe, and F. Reuzeau, *The Perseveration Syndrome in the Pilot's Activity: Guidelines and Cognitive Countermeasures*. Springer Science+Business Media, Sep. 2009. doi: 10.1007/978-3-642-11750-3_6.

[12] A. Hamann and N. Carstengerdes, "Investigating mental workload-induced changes in cortical oxygenation and frontal theta activity during simulated flights," *Scientific Reports*, vol. 12, Apr. 2022. doi: 10.1038/s41598-022-10044-y.

[13] F. Dehais, M. Causse, F. Vachon, N. Regis, E. Menant, and S. Tremblay, "Failure to detect critical auditory alerts in the cockpit," *Human Factors*, vol. 56, pp. 631–644, Nov. 2013. doi: 10.1177/0018720813510735.

[14] J. C. Prinet, "Attentional Narrowing: Triggering, Detecting and Overcoming a Threat to Safety," *University of Michigan*, 2016. [Online]. Available: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/135773/jprinet_1.pdf.

[15] Y.-F. Tsai, E. S. Viirre, C. Strychacz, and T.-P. Jung, "Task performance and eye activity: Predicting behavior relating to cognitive workload," *ResearchGate*, Jun. 2007.

[16] K. A. Moore and L. Gugerty, "Development of a Novel Measure of Situation Awareness: The Case for Eye Movement Analysis," *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting*, vol. 54, pp. 1650–1654, Sep. 2010. doi: 10.1177/154193121005401961.

[17] S. D. Young, T. S. Daniels, E. T. Evans, M. U. deHaag, and P. Duan, "Understanding Crew Decision-Making in the Presence of Complexity - A Flight Simulation Experiment," *AIAA Infotech@Aerospace (I@A) Conference*, Aug. 2013. doi: 10.2514/6.2013-4894.

[18] C. D. Wickens, "Attentional Tunneling and Task Management," 2005.

[19] B. V. Syiem, R. Kelly, J. Goncalves, E. Velloso, and T. Dingler, "Impact of Task on Attentional Tunneling in Handheld Augmented Reality," *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 2021. doi: 10.1145/3411764.3445580.

[20] E. M. Rantanen and J. H. Goldberg, "The effect of mental workload on the visual field size and shape," *Ergonomics*, vol. 42, pp. 816–834, Jun. 1999. doi: 10.1080/001401399185315.

[21] B. Reimer, "Impact of cognitive task complexity on drivers' visual tunneling," *Transportation Research Record*, vol. 2138, pp. 13–19, Jan. 2009. doi: 10.3141/2138-03.

[22] B. Cain, "A Review of the Mental Workload Literature," *ResearchGate*, Jul. 2007.

[23] C. D. Wickens, "Multiple Resources and Mental Workload," *Human Factors*, vol. 50, pp. 449–455, Jun. 2008. doi: 10.1518/001872008x288394.

[24] European Union Aviation Safety Agency, "What are 'Sterile Flight Deck Procedures'?," Dec. 2018. Accessed: Dec. 2023. [Online]. Available: https://www.easa.europa.eu/en/faq/19134.

[25] K. Pedret and G. A. Jamieson, *Characterizing Adaptive Display Interventions for Attentional Tunneling*. Jun. 2021. doi: 10.1007/978-3-030-74614-8_51.

[26] G. GlaholtMackenzie, D. R. Toronto, and Development, "Eye Tracking in the Cockpit: a Review of the Relationships between Eye Movements and the Aviators Cognitive State," tech. rep., 2014.

[27] E. Ktistakis, V. Skaramagkas, D. Manousos, N. S. Tachos, E. E. Tripoliti, D. I. Fotiadis, and M. Tsiknakis, "COLET: A dataset for COgnitive workLoad estimation based on eye-tracking," *Computer Methods and Programs in Biomedicine*, vol. 224, p. 106989, Sep. 2022. doi: 10.1016/j.cmpb.2022.106989.

[28] J. C. Prinet and N. Sarter, "The effects of high stress on attention," *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting*, vol. 59, pp. 1530–1534, Sep. 2015. doi: 10.1177/1541931215591331.

[29] J. Engström, E. Johansson, and J. Östlund, "Effects of visual and cognitive load in real and simulated motorway driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 8, pp. 97–120, Mar. 2005. doi: 10.1016/j.trf.2005.04.012.

[30] A. K. Faulhaber, M. Friedrich, and T. Kapol, "Absence of pilot monitoring affects scanning behavior of pilot flying: Implications for the design of Single-Pilot Cockpits," *Human Factors*, vol. 64, pp. 278–290, Jul. 2020. doi: 10.1177/0018720820939691.

[31] Bureau of Enquiry and Analysis for Civil Aviation Safety, "Final Report flight AF 447 Rio de Janeiro - Paris," Jul. 2012. Accessed: Dec. 2023. [Online]. Available: https://bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf.

[32] W. Sullivan-Kwantes, M. N. Cramer, F. Bouak, and L. S. Goodman, *Environmental stress in military settings*. Jan. 2021. doi: 10.1007/978-3-030-02866-4_107-1.

[33] J. L. Szalma and P. A. Hancock, "Noise effects on human performance: A meta-analytic synthesis.," *Psychological Bulletin*, vol. 137, pp. 682–707, Jul. 2011. doi: 10.1037/a0023987.

[34] M. Friedrich, S. Y. Lee, P. R. Bates, W. Martin, and A. K. Faulhaber, "The influence of training level on manual flight in connection to performance, scan pattern, and task load," *Cognition, Technology & Work*, vol. 23, pp. 715–730, Mar. 2021. doi: 10.1007/s10111-020-00663-8.

[35] J. Jarmasz, C. M. Herdman, and K. R. Jóhannsdóttir, "Object-Based attention and cognitive tunneling.," *Journal of Experimental Psychology: Applied*, vol. 11, pp. 3–12, Jan. 2005. doi: 10.1037/1076-898x.11.1.3.

[36] K. D. Kennedy, C. Stephens, R. Williams, and P. C. Schutte, "Automation and inattentional blindness in a simulated flight task," *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting*, vol. 58, pp. 2058–2062, Sep. 2014. doi: 10.1177/1541931214581433.

[37] B. Chase, "Eye tracking and operator attentional state," 2004.

[38] T. Gateau, H. Ayaz, and F. Dehais, "In silico vs. Over the Clouds: On-the-Fly Mental State Estimation of Aircraft Pilots, Using a Functional Near Infrared Spectroscopy Based Passive-BCI," *Frontiers in Human Neuroscience*, vol. 12, May 2018. doi: 10.3389/fnhum.2018.00187.

[39] A. Monk, D. Jackson, D. Nielsen, E. Jefferies, and P. Olivier, "N-backer: An auditory n-back task with automatic scoring of spoken responses," *Behavior Research Methods*, Mar. 2011. doi: 10.3758/s13428-011-0074-z.

[40] A. Meule, "Reporting and Interpreting Working Memory Performance in n-back Tasks," *Frontiers in Psychology*, vol. 8, Mar. 2017. doi: 10.3389/fpsyg.2017.00352.

[41] M. A. Staal, "Stress, Cognition, and Human Performance: A Literature Review and Conceptual Framework," tech. rep., Aug. 2004.

[42] E. Harmon-Jones, T. F. Price, and P. A. Gable, "The influence of affective states on cognitive broadening/narrowing: Considering the importance of motivational intensity," *Social and Personality Psychology Compass*, vol. 6, pp. 314–327, 2012.

[43] L. Thomas and C. D. Wickens, "Eye-tracking and Individual Differences in off-Normal Event Detection when Flying with a Synthetic Vision System Display," *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting*, vol. 48, pp. 223–227, Sep. 2004. doi: 10.1177/154193120404800148.

[44] N. Régis, F. Dehais, E. Rachelson, C. Thooris, S. Pizziol, M. Causse, and C. Tessier, "Formal detection of attentional tunneling in human operator–automation interactions," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 3, 2014.

[45] B. Berthelot, P. Mazoyer, S. Egea, J. André, E. Grivel, and P. Legrand, "Self-Affinity of an aircraft pilot's gaze direction as a marker of visual tunneling," *SAE technical paper series*, Sep. 2019. doi: 10.4271/2019-01-1852.

[46] S. A. Mayo, "Change Blindness in the Synthetic Vision Primary Flight Display: Comparing Eye Tracking Patterns with Pilot Attention," 2009.

[47] A. White and D. O'Hare, "In plane sight: Inattentional blindness affects visual detection of external targets in simulated flight," *Applied Ergonomics*, vol. 98, p. 103578, Jan. 2022. doi: 10.1016/j.apergo.2021.103578.

[48] N. R. Johnson, D. A. Wiegmann, and C. D. Wickens, "Effects of Advanced Cockpit Displays on General Aviation Pilots' Decisions to Continue Visual Flight Rules Flight into Instrument Meteorological Conditions," *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting*, vol. 50, pp. 30–34, Oct. 2006. doi: 10.1177/154193120605000107.

[49] C. Iani and C. D. Wickens, "Factors affecting task management in aviation," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Sep. 2004. doi: 10.1177/154193120404800146.

[50] M. F. Bear, B. W. Connors, and M. A. Paradiso, "Neuroscience: Exploring the brain, 3rd ed.," 2007.

[51] A. T. Duchowski, *Eye tracking methodology: Theory and Practice*. third edition ed., Jan. 2007. doi: 10.1007/978-1-84628-609-4.

[52] S. Peißl, C. D. Wickens, and R. Baruah, "Eye-Tracking Measures in Aviation: A Selective Literature review," *The international journal of aerospace psychology*, vol. 28, pp. 98–112, Sep. 2018. doi: 10.1080/24721840.2018.1514978.

[53] K. Krejtz, A. T. Duchowski, A. Niedzielska, C. Biele, and I. Krejtz, "Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze," *PLOS ONE*, vol. 13, p. e0203629, Sep. 2018. doi: 10.1371/journal.pone.0203629.

[54] K. Holmqvist and R. Andersson, *Eye-tracking: A comprehensive guide to methods, paradigms and measures*. Nov. 2017.

[55] S. Pizziol, F. Dehais, and C. Tessier, "Towards human operator state assessment," pp. 99–106, May 2011.

[56] N. M. Moacdieh, S. P. Devlin, H. Jundi, and S. L. Riggs, "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*, vol. 14, pp. 132–151, Jan. 2020. doi: 10.1177/1555343419892184.

[57] B. Shiferaw, L. A. Downey, and D. P. Crewther, "A review of gaze entropy as a measure of visual scanning efficiency," *Neuroscience Biobehavioral Reviews*, vol. 96, pp. 353–366, Jan. 2019. doi: 10.1016/j.neubiorev.2018.12.007.

[58] F. Di Nocera, M. Terenzi, and M. Camilli, "Another look at scanpath: distance to nearest neighbour as a measure of mental workload," *ResearchGate*, Jan. 2006.

[59] F. Di Nocera, M. Camilli, and M. Terenzi, "A random glance at the flight deck: pilots' scanning strategies and the Real-Time assessment of mental workload," *Journal of Cognitive Engineering and Decision Making*, vol. 1, pp. 271–285, Sep. 2007. doi: 10.1518/155534307x255627.

[60] M. Batty, R. Morphet, A. P. Masucci, and K. Stanilov, "Entropy, complexity, and spatial information," *Journal of Geographical Systems*, vol. 16, pp. 363–385, Sep. 2014. doi: 10.1007/s10109-014-0202-2.

[61] L. L. Di Stasi, C. Díaz-Piedra, H. Rieiro, J. M. S. Carrion, M. M. Berrido, G. Q. Olivares, and A. Catena, "Gaze entropy reflects surgical task load," *Surgical Endoscopy and Other Interventional Techniques*, vol. 30, pp. 5034–5043, Mar. 2016. doi: 10.1007/s00464-016-4851-8.

[62] T. Lu, Z. Lou, F. Shao, Y. Li, and X. You, "Attention and Entropy in Simulated Flight with Varying Cognitive Loads," *Aerospace medicine and human performance*, vol. 91, pp. 489–495, Jun. 2020. doi: 10.3357/amhp.5504.2020.

[63] T. Victor, J. L. Harbluk, and J. Engström, "Sensitivity of eye-movement measures to in-vehicle task difficulty," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 8, pp. 167–190, Mar. 2005. doi: 10.1016/j.trf.2005.04.014.

[64] C. Desmet and K. Diependaele, "An eye-tracking study on the road examining the effects of handsfree phoning on visual attention," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 60, pp. 549–559, Jan. 2019. doi: 10.1016/j.trf.2018.11.013.

[65] S. Hutton, "Visual angle," *Fast, Accurate, Reliable Eye Tracking*, Mar. 2023. Accessed: Dec. 2023. [Online]. Available: https://www.sr-research.com/eye-tracking-blog/background/visual-angle/.

[66] R. A. McKinley, L. K. McIntire, R. Schmidt, D. Repperger, and J. A. Caldwell, "Evaluation of eye metrics as a detector of fatigue," *Human Factors*, vol. 53, pp. 403–414, Jul. 2011. doi: 10.1177/0018720811411297.

[67] M. Kaczorowska, M. Plechawska–Wójcik, and M. Tokovarov, "Interpretable machine learning models for Three-Way classification of cognitive workload levels for Eye-Tracking features," *Brain Sciences*, vol. 11, p. 210, Feb. 2021. doi: 10.3390/brainsci11020210.

[68] O. V. Bitkina, J. Park, and H. K. Kim, "The ability of eye-tracking metrics to classify and predict the perceived driving workload," *International Journal of Industrial Ergonomics*, vol. 86, p. 103193, Nov. 2021. doi: 10.1016/j.ergon.2021.103193.

[69] Dansbecker, "XGBoost," Jul. 2018. https://www.kaggle.com/code/dansbecker/xgboost.

[70] Vipulgandhi, "How to choose right metric for evaluating ML Model," Jan 2020. Accessed: Dec. 2023. [Online]. Available: https://www.kaggle.com/code/vipulgandhi/how-to-choose-right-metric-for-evaluating-ml-model.

[71] R. Hendricks, "What is a good accuracy score in Machine Learning?," Nov. 2022. Accessed: Dec. 2023. [Online]. Available: https://deepchecks.com/question/what-is-a-good-accuracy-score-in-machine-learning/.

[72] M. J. Kleiner, L. Wong, A. Dubé, K. Wnuk, S. Hunter, and L. Graham, "Dual-Task Assessment Protocols in Concussion Assessment: A Systematic Literature review," *Journal of Orthopaedic Sports Physical Therapy*, vol. 48, pp. 87–103, Feb. 2018. doi: 10.2519/jospt.2018.7432.

[73] C. Leone, L. Moumdjian, F. Patti, E. Vanzeir, I. Baert, R. Veldkamp, B. Van Wijmeersch, and P. Feys, "Comparing 16 different Dual–Tasking paradigms in individuals with multiple sclerosis and healthy controls: working memory tasks indicate Cognitive–Motor interference," *Frontiers in Neurology*, vol. 11, Aug. 2020. doi: 10.3389/fneur.2020.00918.

[74] W. P. Berg, E. D. Berglund, A. J. Strang, and M. J. Baum, "Attention-capturing properties of high frequency luminance flicker: Implications for brake light conspicuity," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 10, pp. 22–32, Jan. 2007. doi: 10.1016/j.trf.2006.03.006.

[75] J. Cass, E. Van Der Burg, and D. Alais, "Finding Flicker: Critical differences in temporal frequency capture attention," *Frontiers in Psychology*, vol. 2, Jan. 2011. doi: 10.3389/fpsyg.2011.00320.

[76] G. Intelligence, "SMI Porduct Manuals." Accessed: Dec. 2023. [Online]. Available: https://gazeintelligence.com/smi-product-manual.

[77] M. Friedrich, N. Rußwinkel, and C. Möhlenbrink, "A guideline for integrating dynamic areas of interests in existing set-up for capturing eye movement: Looking at moving aircraft," *Behavior Research Methods*, vol. 49, pp. 822–834, Jun. 2016. doi: 10.3758/s13428-016-0745-x.

[78] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, p. 71–78, Association for Computing Machinery, 2000. doi: 10.1145/355017.355028.

[79] S. G. Hart, "NASA-Task Load Index (NASA-TLX); 20 years later," *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting*, vol. 50, pp. 904–908, Oct. 2006. doi: 10.1177/154193120605000909.

[80] V. Battiste and M. R. Bortolussi, "Transport pilot workload: A comparison of two subjective techniques," *Proceedings of the Human Factors Society annual meeting*, vol. 32, pp. 150–154, Oct. 1988. doi: 10.1177/154193128803200232.

[81] H. Devos, K. M. Gustafson, P. Ahmadnezhad, K. Liao, J. D. Mahnken, W. M. Brooks, and J. M. Burns, "Psychometric properties of NASA-TLX and index of cognitive activity as measures of cognitive workload in older adults," *Brain Sciences*, vol. 10, p. 994, Dec. 2020. doi: 10.3390/brainsci10120994.

[82] J. G. O'Connell, "Comparison of three subjective workload metrics for a free flight enviroment," *2007 International Symposium on Aviation Psychology*, pp. 481–485, 2007.

[83] A. Tattersall and P. S. Foord, "An experimental evaluation of instantaneous self-assessment as a measure of workload," *Ergonomics*, vol. 39, pp. 740–748, May 1996. doi: 10.1080/00140139608964495.

[84] P. Keating and U. of California, "Lecture: D-prime (signal detection) analysis," 2004. Accessed: Dec. 2023. [Online]. Available: http://phonetics.linguistics.ucla.edu/facilities/statistics/dprime.htm.

[85] A. Meule, A. K. Skirde, R. Freund, C. Vögele, and A. Kübler, "High-calorie food-cues impair working memory performance in high and low food cravers," *Appetite*, vol. 59, pp. 264–269, Nov. 2012. doi: 10.1016/j.appet.2012.05.010.

[86] L. Meteyard and R. Davies, "Best practice guidance for linear mixed-effects models in psychological science," *Journal of Memory and Language*, vol. 112, p. 104092, Jun. 2020. doi: 10.1016/j.jml.2020.104092.

[87] German Aerospace Center (DLR), "DLR - Institute of Aerospace Medicine - zertifikat." Accessed: Dec. 2023. [Online]. Available: https://www.dlr.de/me/en/desktopdefault.aspx/tabid-5054/.

[88] Wikipedia-Autoren, "DLR-Test," Jan. 2006. Accessed: Dec. 2023. [Online]. Available: https://de.wikipedia.org/wiki/DLR-Test.

[89] P. J. Rosopa, M. M. Schaffer, and A. N. Schroeder, "Managing heteroscedasticity in general linear models.," *Psychological Methods*, vol. 18, pp. 335–351, Jan. 2013. doi: 10.1037/a0032553.

[90] Zach, "Understanding heteroscedasticity in regression analysis," Nov. 2020. Accessed: Dec. 2023. [Online]. Available: https://www.statology.org/heteroscedasticity-regression/.

[91] C. Ford and University of Virginia Library, "Interpreting log transformations in a linear model — UVA Library," Aug. 2018. Accessed: Dec. 2023. [Online]. Available: https://library.virginia.edu/data/articles/interpreting-log-transformations-in-a-linear-model.

[92] Z. Hailat, "Part 09 - Constructing Multi-Class Classifier Using SVM with Python," Aug. 2020. Accessed: Dec. 2023. [Online]. Available: https://www.youtube.com/watch?v=Zj1CoJk2feE.

[93] J. Brownlee, "A Gentle Introduction to k-fold Cross-Validation," Oct. 2023. Accessed: Dec. 2023. [Online]. Available: https://machinelearningmastery.com/k-fold-cross-validation/.

[94] A. Singh, T. Tiwari, and I. Singh, "Performance feedback, mental workload and monitoring efficiency," *Journal of the Indian Academy of Applied Psychology*, vol. 36, Jan. 2010.

[95] S. Scannella, V. Peysakhovich, F. Ehrig, E. Lepron, and F. Dehais, "Assessment of ocular and physiological metrics to discriminate flight phases in real light aircraft," *Human Factors*, vol. 60, pp. 922–935, Jul. 2018. doi: 10.1177/0018720818787135.

[96] SKYBrary Aviation Safety, "Pilot Workload." Accessed: Dec. 2023. [Online]. Available: https://skybrary.aero/articles/pilot-workload.

A Appendix: Preparational Materials

A.1 Ethics Review

Geschäftsstelle „Forschungsethik“

Deutsches Zentrum für Luft- und Raumfahrt

DLR e. V. Geschäftsstelle „Forschungsethik“
51170 Köln

Elena Rankova
Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR)
Institut für Verkehrssystemtechnik
Lilienthalplatz 7
38106 Braunschweig

Ihr Zeichen
Ihr Schreiben 01.06.2023
Unser Zeichen 08/23

Ihr Gesprächspartner Philipp Bergeron

Telefon +492203 601 4002
Telefax +492203 601 3906
E-Mail ethik@dlr.de

27. Juni 2023

Ethikantrag “ LOKI - Studie „Workload and Attentional Tunneling““ vom 01.06.2023

Sehr geehrte Frau Rankova,

wir beziehen uns auf Ihren Antrag auf Stellungnahme der Geschäftsstelle „Forschungsethik“ im Deutschen Zentrum für Luft- und Raumfahrt e.V. vom 01.06.2023.

Die Durchführung Ihres Vorhabens scheint ethisch unbedenklich, jedoch bleibt die Verantwortung des Vorhabens selbstverständlich beim Antragssteller und bleibt auch durch dieses Schreiben unberührt.

Die Geschäftsstelle „Forschungsethik“ weist dringend darauf hin, dass bei Probanden – nicht deutscher Staatbürgerschaft“ vorab eine interne Sicherheitsprüfung eingeleitet werden sollte. Sanktionierten Personen dürfen keine wirtschaftlichen Ressourcen zur Verfügung gestellt werden und der Zutritt zum Gelände muss ihnen verweigert werden.

Weiterhin sollte auch mit der Einwilligungserklärung die Löschung aller personenbezogener Daten (wie z.B. Kontoverbindungen, Videomaterial und sonstige medizinische Daten) geklärt werden.

Für Ihre Studie wünschen wir viel Erfolg!

Mit freundlichen Grüßen

i.A. Philipp Bergeron
Leiter der Geschäftsstelle „Forschungsethik“

A.2 Study Invitation Brochure

Einladung von Pilotinnen und Piloten

Das DLR sucht Pilotinnen und Piloten mit Airbus-Type-Rating zur Teilnahme an einer Studie zur Arbeitsbelastung im Cockpit im August 2023 in Braunschweig.

Motivation
Am Institut für Flugführung des DLR am Standort Braunschweig werden neue Methoden zur Erkennung hoher Arbeitsbelastung und gefährlicher kognitiver Zustände erforscht. Für den Zweck wird in August 2023 eine Studienreihe durchgeführt, bei der Daten über die Augenbewegungen und Leistungen von Piloten unter variierenden Arbeitsbelastungen im Cockpit gesammelt werden.

Sie als Versuchsperson werden bei der Studie mehrere Anflüge mit Landung in einem Full Motion A320 Simulator fliegen. Dabei werden Ihre Augenbewegungen und Flugdaten gesammelt und ausgewertet. Die Teilnahme nimmt etwa drei Stunden in Anspruch. Innerhalb dieser Zeit werden eine Einführung zum Simulator und vier Szenarien durchgespielt. Durch Ihre Teilnahme helfen Sie uns, neue Ansätze für die Sicherheit beim Fliegen zu entwickeln.

Wer?
Piloten mit Lizenz (ATPL, CPL, PPL)

Wann?
Die Studie findet in August, an folgenden Tagen statt:
09.08, 10.08, 11.08, 17.08, 18.08, 21.08, 22.08, 23.08

Wo?
Am DLR-Standort in Braunschweig (Lilienthalplatz 7, 38108 Braunschweig)

Aufwandsentschädigung
Für die Teilnahme erhalten Sie eine Aufwandsentschädigung von insgesamt 120 €.
Für Ihre Reise- und Hotelkosten erstatten wir Ihnen bis zu 200 €.

Kurzbeschreibung
An dem vereinbarten Studientag werden Sie an einer Studie zum oben genannten Thema teilnehmen. Die Studie wird zwei bis drei Stunden dauern. Die Versuche finden am Air Vehicle motion Simulator statt.

Wenn Sie Interesse an der Teilnahme oder weitere Fragen haben, melden Sie sich bitte bei Julia Höver unter der Mailadresse: julia.hoever@dlr.de

Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR) • Institut für Flugführung, Abteilung Systemergonomie
Lilienthalplatz 7, 38108 Braunschweig • DLR.de

DLR

B Appendix: Experiment Materials

B.1 Participant Consent Form

Deutsches Zentrum
DLR für Luft- und Raumfahrt
German Aerospace Center

LOKI 2023, Arbeitsbelastung im Cockpit

Dates	August 2023
Place	German Aerospace Center (DLR) Lilienthalplatz 7 38108 Braunschweig
Project Leader	Maik Friedrich, DLR
Exercise Leader	Elena Rankova, DLR

Participant Agreement Form

I have been informed by the exercise leader about the purpose, course and meaning of the experiment, as well as about the benefits and risks that may be associated with it. I completely understood this information. All my questions have been answered to my satisfaction. I had enough time to reconsider my decision to participate at my own will.

Throughout the experiment, the following data will be collected:

- Eye tracking data
- Simulator data, including flight strips and radio communication data
- Video and audio data
- Questionnaire data

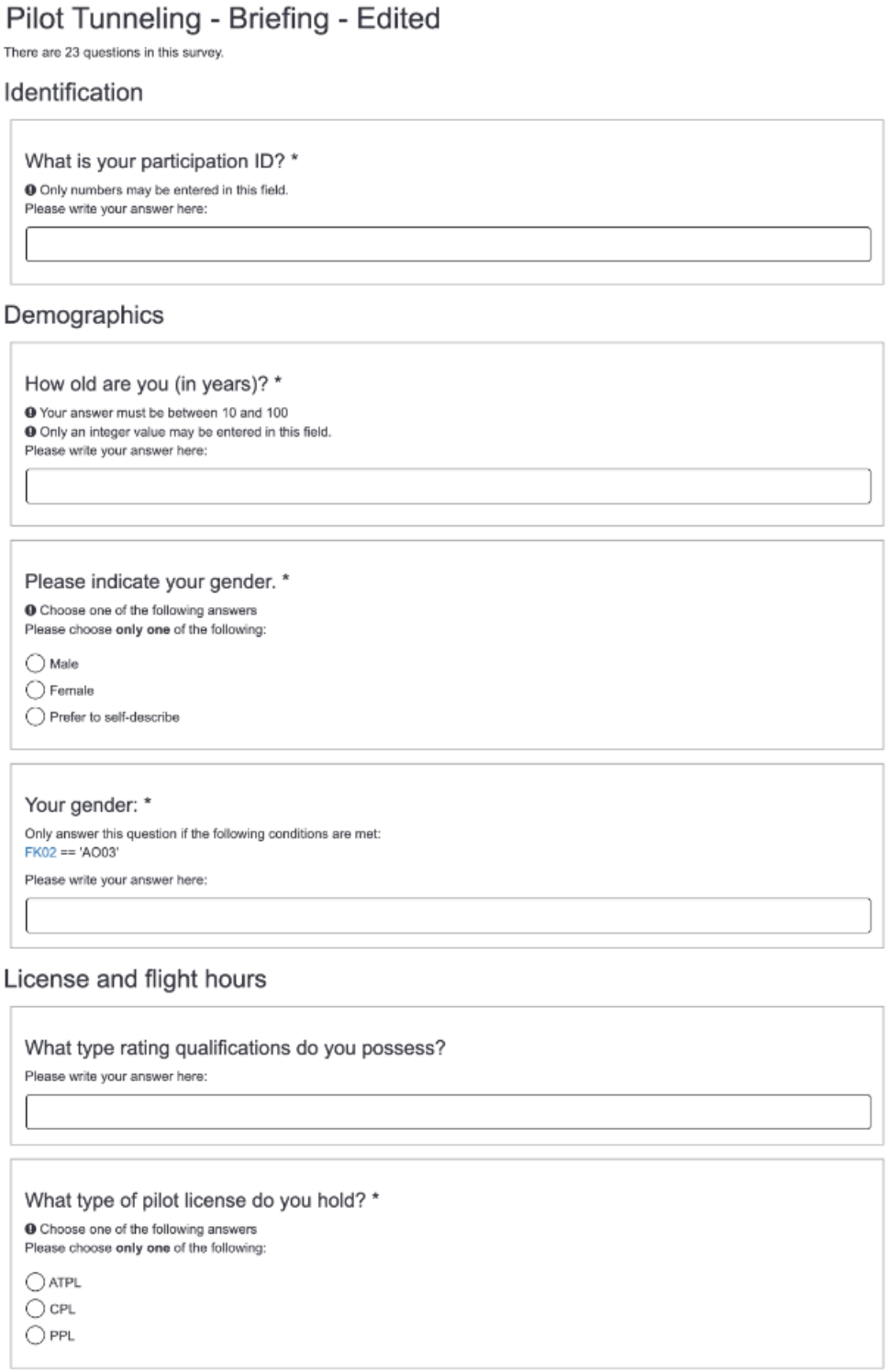
Please read carefully the following statements and tick on the side bar (✓or ✗) if you agree with it. Leave the side bar of a statement empty, if you don't agree with it.	✓ ✗
I am aware of the main aspects of the validation plan for the planned LOKI activity.	
I confirm that I had the opportunity to ask questions.	
I understand that my participation is entirely voluntary. I can refrain from participating at any time, without penalty or prejudice.	
I understand that my answers to any questionnaire will remain anonymous.	
Should I not wish to answer any particular question(s), I am free to decline without any penalty or prejudice.	
I give permission for members of the research team to have access to my anonymized responses. I understand that my name will not be linked to the research materials and that I will not be identified or identifiable in the outputs that result from the research without my agreement. Any data will be transferred will be anonymous.	
I have the right to request to have my personal data deleted at any time by contacting the Data Protection Officer. I understand that the retention period for all personal data related to the project is 5 years after the end of the project. After this 5-year period, all personal data concerning the volunteer participants will be destroyed.	
Some picture/video could be taken during the validation and may be published in the project website for communication and dissemination purposes. I give authorization to use my image only for these purposes.	
I will receive a compensation of 120,00 € incentive for having taken part in this validation activity. Additionally, I can receive up to 200 € reimbursement for travel and accommodation costs.	
I agree to take part in the validation activity.	

_____ _____
Name of Participant Date, Signature

Elena Rankova
_____ _____
Name of Exercise Leader Date, Signature

B.2 Briefing Questionnaire

The questionnaire was presented to participants once at the beginning of the experiment, following the briefing presentation. It was executed on a tablet, using the LimeSurvey software and consisted of multiple pages. The experimenter filled in the first question.



Pilot Tunneling - Briefing - Edited

There are 23 questions in this survey.

Identification

What is your participation ID? *

Only numbers may be entered in this field.
Please write your answer here:

Demographics

How old are you (in years)? *

Your answer must be between 10 and 100
Only an integer value may be entered in this field.
Please write your answer here:

Please indicate your gender. *

Choose one of the following answers
Please choose **only one** of the following:

- Male
- Female
- Prefer to self-describe

Your gender: *

Only answer this question if the following conditions are met:
FK02 == 'AO03'

Please write your answer here:

License and flight hours

What type rating qualifications do you possess?

Please write your answer here:

What type of pilot license do you hold? *

Choose one of the following answers
Please choose **only one** of the following:

- ATPL
- CPL
- PPL

Briefing Questionnaire - Part 1

What type of pilot license do you hold? *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ ATPL
- ◯ CPL
- ◯ PPL

Approximately how many flight hours have you completed? *

Only numbers may be entered in this field.
Please write your answer here:

Approximately how many flight hours have you accumulated in the past year?

Only numbers may be entered in this field.
Please write your answer here:

NASA-TLX Pairwise Comparisons

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Frustration
- ◯ Temporal Demand

Temporal Demand refers to how much time pressure you felt due to the rate or pace at which the task occurred.

Frustration refers to how insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent you felt during the task.

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Mental Demand
- ◯ Physical Demand

Mental Demand refers to how much mental and perceptual activity was required to complete the task.

Physical Demand refers to how much physical activity was required to complete the task.

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Mental Demand
- ◯ Temporal Demand

Temporal Demand refers to how much time pressure you felt due to the rate or pace at which the task occurred.

Mental Demand refers to how much mental and perceptual activity was required to complete the task.

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Effort
- ◯ Mental Demand

Mental Demand refers to how much mental and perceptual activity was required to complete the task.

Effort refers to how hard you had to work mentally and physically to accomplish your level of performance.

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Effort
- ◯ Temporal Demand

Temporal Demand refers to how much time pressure you felt due to the rate or pace at which the task occurred.

Effort refers to how hard you had to work mentally and physically to accomplish your level of performance.

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Frustration
- ◯ Performance

Performance refers to how successful you think you were in accomplishing the goals of the task.

Frustration refers to how insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent you felt during the task.

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Effort
- ◯ Performance

Effort refers to how hard you had to work mentally and physically to accomplish your level of performance.

Performance refers to how successful you think you were in accomplishing the goals of the task.

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Effort
- ◯ Physical Demand

Effort refers to how hard you had to work mentally and physically to accomplish your level of performance.

Physical Demand refers to how much physical activity was required to complete the task.

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Frustration
- ◯ Physical Demand

Physical Demand refers to how much physical activity was required to complete the task.

Frustration refers to how insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent you felt during the task.

Briefing Questionnaire Part - 3

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Effort
- ◯ Frustration

Frustration refers to how insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent you felt during the task.

Effort refers to how hard you had to work mentally and physically to accomplish your level of performance.

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Physical Demand
- ◯ Performance

Physical Demand refers to how much physical activity was required to complete the task.

Performance refers to how successful you think you were in accomplishing the goals of the task.

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Performance
- ◯ Temporal Demand

Performance refers to how successful you think you were in accomplishing the goals of the task.

Temporal Demand refers to how much time pressure you felt due to the rate or pace at which the task occurred.

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Physical Demand
- ◯ Temporal Demand

Physical Demand refers to how much physical activity was required to complete the task.

Temporal Demand refers to how much time pressure you felt due to the rate or pace at which the task occurred.

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Frustration
- ◯ Mental Demand

Frustration refers to how insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent you felt during the task.

Mental Demand refers to how much mental and perceptual activity was required to complete the task.

Select the factor that represents the more important contributor to workload based on the specific tasks and duties of a pilot. *

Choose one of the following answers
Please choose **only one** of the following:

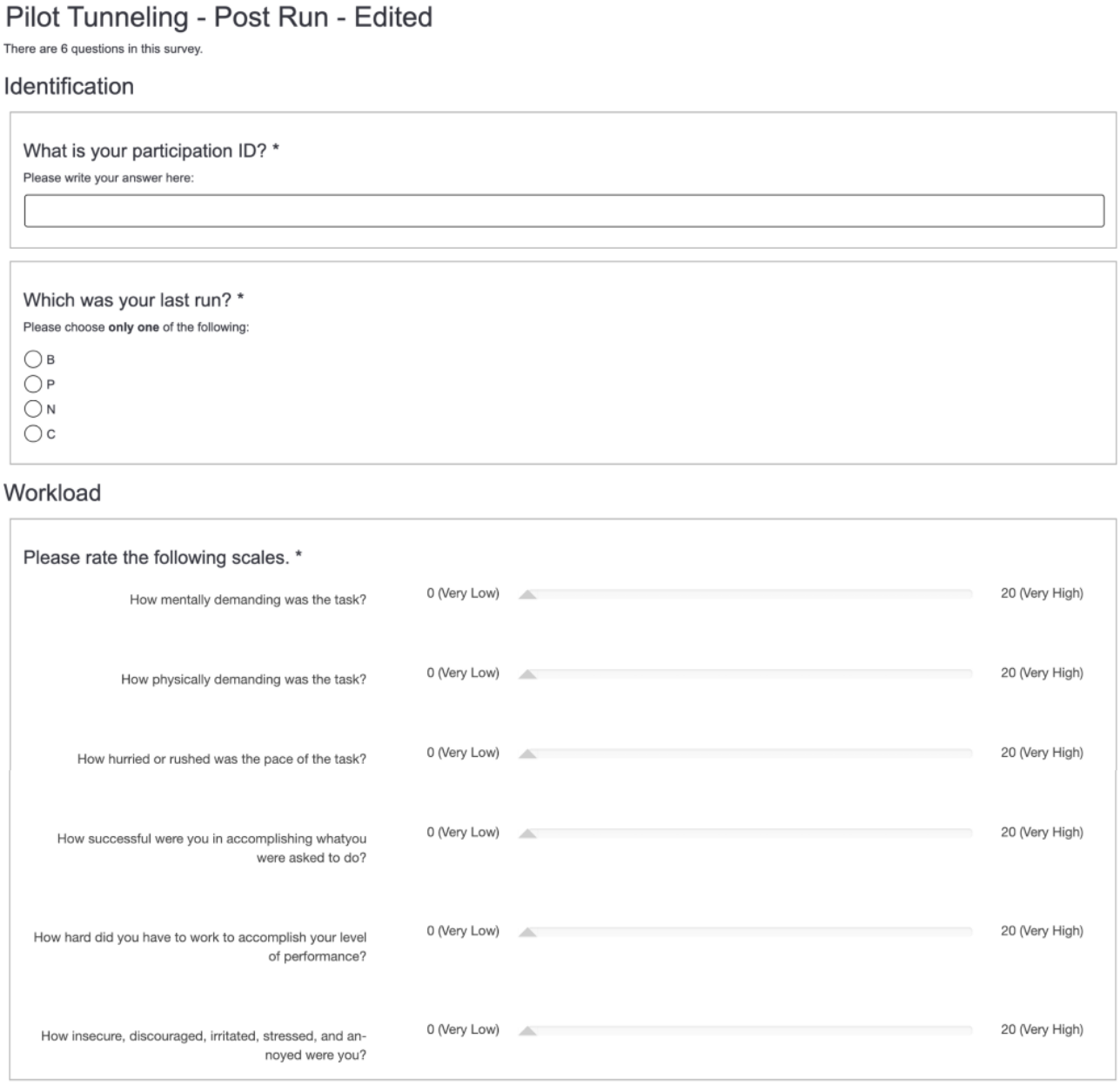
- ◯ Mental Demand
- ◯ Performance

Performance refers to how successful you think you were in accomplishing the goals of the task.

Mental Demand refers to how much mental and perceptual activity was required to complete the task.

B.3 Post-Run Questionnaire

The questionnaire was presented to participants at the end of each run. It was executed on a tablet, using the LimeSurvey software, and consisted of multiple pages. The experimenter filled in the first two questions.



Pilot Tunneling - Post Run - Edited

There are 6 questions in this survey.

Identification

What is your participation ID? *

Please write your answer here:

Which was your last run? *

Please choose **only one** of the following:

- B
- P
- N
- C

Workload

Please rate the following scales. *

How mentally demanding was the task?	0 (Very Low)	20 (Very High)
How physically demanding was the task?	0 (Very Low)	20 (Very High)
How hurried or rushed was the pace of the task?	0 (Very Low)	20 (Very High)
How successful were you in accomplishing whatyou were asked to do?	0 (Very Low)	20 (Very High)
How hard did you have to work to accomplish your level of performance?	0 (Very Low)	20 (Very High)
How insecure, discouraged, irritated, stressed, and annoyed were you?	0 (Very Low)	20 (Very High)

Post-Run Questionnaire Part - 1

Tunneling and Performance

Please briefly describe any abnormal situations you encountered during the last run. If nothing unusual happened, type "no". *

Please write your answer here:

How would you rate your performance during the last run?

Please choose **only one** of the following:

- ◯ Poor
- ◯ Fair
- ◯ Good
- ◯ Very good
- ◯ Excellent

Simulation Quality

Please choose the appropriate response for each item:

	Very Satisfied	Satisfied	Neutral	Dissatisfied	Very Dissatisfied
How would you rate the realism in comparison to your ususal cockpit situations?	◯	◯	◯	◯	◯

08-22-2023 – 13:37

Submit your survey.
Thank you for completing this survey.

Post-Run Questionnaire Part - 2

B.4 Protocol and Timing Sheet

This protocol was utilized as a checklist to ensure the proper implementation of the scenarios. Additionally, the reported ISA workload scores and the notice times were documented using this sheet.

Müssen Koordinaten Daten neu gespeichert werden?							
Step	**Minuten**		**Done?**				
1		Meet VP at the entrance			08:50 – 09:10	12:20 – 12:40	15:50 – 16:10
2	10	Briefing und Fragebogen			09:10 – 09:20	12:40 – 12:50	16:10 – 16:20
3	10	Sicherheitsanweisung – David			09:20 – 09:30	12:50 – 13:00	16:20 – 16:30
4	5	NBack training			09:35 – 09:40	13:00 – 14:05	16:30 – 16:35
5	10	Testflug			09:35 – 09:45	13:05 – 13:15	16:35 – 16:45
6		Eye-tracking on?					
7		AVES Data on?					
8		Audio on?					
9		Timer on?					
10	15	Versuch 1			09:45 – 10:00	13:15 – 13:30	16:45 – 17:00
11		ab Magdeburg N-Back on?					
		When did he notice it?	ST:	NT:			
12	10	Fragebogen und Pause			10:00 – 10:10	13:30 – 13:40	17:00 – 17:10
13		Eye-tracking on?					
14		AVES Data on?					
15		Audio on?					
16		Timer on?					
17	15	Versuch 2			10:10 – 10:25	13:40 – 13:55	17:10 – 17:25
18		ab Magdeburg N-Back on?					
		When did he notice it?	ST:	NT:			
19	10	Fragebogen und Pause			10:25 – 10:35	14:05 – 14:15	17:25 – 17:35
20		Eye-tracking on?					
21		AVES Data on?					
22		Audio on?					
23		Timer on?					
24	15	Versuch 3			10:35 – 10:50	14:15 – 14:30	17:35 – 17:50
25		ab Magdeburg N-Back on?					
		When did he notice it?	ST:	NT:			
26	10	Fragebogen und Pause			10:50 – 11:00	14:30 – 14:40	17:50 – 18:05
27		Eye-tracking on?					
28		AVES Data on?					
29		Audio on?					
30		Timer on?					
31	15	Versuch 4			11:00 – 11:15	14:40 – 14:55	18:05 – 18:25
32		ab Magdeburg N-Back on?					
		When did he notice it?	ST:	NT:			
33	10	Fragebogen und Pause			11:15 – 11:25	14:55 – 15:05	18:25 – 18:35
34	5	Debriefing			11:25 – 11:30	15:05 – 15:10	18:35 – 18:40

B.5 N-back Response Sheet

Participants' responses were documented using this sheet. The numbers enclosed in brackets represent the correct n-back targets.

VP Nr:																														
Aufgabe 2 Training																														
2	10	7	5	9	[5]	[9]	8	[9]	9	[9]	[9]	[9]	4	3	3	2	[3]	[2]	[3]	10										
2	9	[2]	[9]	10	8	8	3	[8]	8	10	[8]	[10]	9	[10]	[9]	[10]	6	[10]	2	1	[2]	[1]	[2]		14 = 60%					
Testflug																														
10	10	[10]	3	[10]	7	[10]	[7]	[10]	4	9	[4]	2	[4]	3	[4]	6	[4]	8	5	[8]	[5]	3	[5]							
9	10	6	1	[6]	[1]	2	8	[2]	[8]	[2]	9	[2]	[9]	[2]	2	[2]	10	[2]	8	6	[8]	[6]	10							
Run N																														
5	7	6	6	2	8	7	[8]	[7]	[8]	[7]	[8]	[7]	[8]	[7]	[8]	[7]	[7]	1	[7]	[1]	6	2	1	5						
3	10	[3]	[10]	4	2	5	[2]	9	[2]	4	[2]	[4]	[2]	3	[2]	1	[2]	3	1	[3]	7	[3]	[7]							
5	9	3	[9]	[3]	6	[3]	[6]	4	7	[4]	[7]	6	[7]	3	5	[3]	[5]	8	[5]	2	[5]	6	[5]							
4	3	2	2	[2]	[2]	6	1	9	[1]	[9]	4	[9]	[7]	5	4	[5]	3	7	[8]	3	[7]	[3]	[7]							
9	10	[9]	[10]	[9]	6	3	[6]	[3]	9	[3]	7	[3]	[7]	5	4	[5]	3	7	9	[7]	[9]	[7]	[7]							
10	3	[10]	[3]	9	10	[9]	[10]	1	9	[1]	[9]	[1]	[9]	9	[9]	3	[9]	4	[9]	[4]	1	10	8							
4	7	1	9	2	[9]	[2]	8	[2]	[8]	9	[8]	8	[8]	[8]	[8]	[8]	[8]	5	4	8	[4]	1	[4]							
9	3	4	[3]	[4]	[3]	8	9	10	4	[10]	[4]	[10]	[4]	[10]	8	2	[8]	[2]	[8]	[2]	10	6	6							
Run C																														
3	1	1	6	[1]	[6]	[1]	[6]	7	[6]	[7]	9	9	[9]	[9]	2	5	[2]	[5]	[2]	7	3	2	[3]							
2	9	[2]	[9]	9	1	[9]	10	[9]	4	[9]	6	[9]	[6]	[9]	8	[9]	5	1	[5]	[1]	7	[1]	9							
5	3	3	8	[3]	1	[3]	[1]	[3]	10	7	[10]	2	5	9	[5]	[9]	10	[9]	[10]	[9]	[10]	1	[10]							
7	10	3	6	[3]	[6]	[3]	4	[3]	[4]	[3]	[4]	[3]	3	1	6	10	9	8	10	[8]	[10]	[8]	[10]							
8	8	[8]	[8]	[8]	6	[8]	2	[8]	[2]	2	8	6	[8]	[6]	1	[6]	[1]	[6]	4	9	7	[9]	1							
7	10	10	[10]	[10]	3	3	[3]	10	10	6	[10]	[6]	[10]	2	5	[2]	7	[2]	[7]	[2]	6	[2]	[6]							
7	6	9	4	3	5	[3]	6	[3]	[6]	[3]	2	[3]	[2]	8	[2]	[8]	[2]	[8]	[2]	2	[2]	4	4							
1	8	7	7	[7]	8	[7]	[8]	[7]	[8]	[7]	6	[7]	2	1	[2]	[1]	10	[1]	8	[1]	9	2	[9]							

C Appendix: Code Samples

C.1 Transition Frequency Calculation

A snippet from the code calculating, in this case, the transition frequency from the tunneling data.

```
import pandas as pd
import numpy as np
import sys, os, glob
import re

noticed_files = []
unnoticed_files = []

for filename in os.listdir(l_ErgebniseOrdner_Eye):
    f = os.path.join(l_ErgebniseOrdner_Eye, filename)
    # checking if it is a file
    if os.path.isfile(f) and os.path.getsize(f) == 0:
        os.remove(f)
    elif "_noticed" in f and os.path.getsize(f) > 0:
        noticed_files.append(f)
    elif "_unnoticed" in f and os.path.getsize(f) > 0:
        unnoticed_files.append(f)

print(len(noticed_files), len(unnoticed_files))

noticed_data= {}
unnoticed_data= {}

for i in noticed_files:
    match = re.search(r'VP(\d{2})_(.*?)_(.*?)_(.*?)_', i)
    if match:
        dfName = f"VP{match.group(1)}_{match.group(3)}_{match.group(4)}"
    noticed_data[dfName] = pd.read_csv(i, sep=";")

for j in unnoticed_files:
    match = re.search(r'VP(\d{2})_(.*?)_(.*?)_(.*?)_', j)
    if match:
        dfName = f"VP{match.group(1)}_{match.group(3)}_{match.group(4)}"
    unnoticed_data[dfName] = pd.read_csv(j, sep=";")

def calc_transitions(noticed, transition_rates, session, noticed_data):
    transitions_counter = 0
    vp_nr = session
    determinator, condition = session.split("_")[2], session.split("_")[1]

    for index,row in noticed_data[session].iloc[1:,:].iterrows():
```

```
        if index == len(noticed_data[session]["Time"])-1:
            new_row = pd.DataFrame({"ParticipantID": vp_nr,"Session":session,
                "Transition_frequency":transitions_counter, "Determinator": determinator,
                "Condition": condition, "Noticed": noticed}, index=[0])
            transition_rates=
                pd.concat([transition_rates.loc[:], new_row]).reset_index(drop=True)
        elif noticed_data[session].at[index,"StaticMask"]!=
            noticed_data[session].at[index-1, "StaticMask"]:
            transitions_counter += 1
        elif noticed_data[session].at[index,"StaticMask"]
            == noticed_data[session].at[index-1, "StaticMask"]:
            continue
        else:
            print(f"Something's wrong, please check noticed transition rate for {session}")

    return transition_rates

transition_rates = pd.DataFrame(columns=["ParticipantID", "Session", "Determinator",
    "Condition" , "Noticed" , "Transition_frequency"])

for session in noticed_data:
    noticed = True
    transition_rates =
        calc_transitions(noticed, transition_rates, session, noticed_data)

for session in unnoticed_data:
    noticed = False
    transition_rates =
        calc_transitions(noticed, transition_rates, session, unnoticed_data)

transition_path =
    os.path.join(l_Export_Ergebnis + "determinator_transition_airb_only.csv")
transition_airb_only.to_csv(transition_path, sep=";", decimal =",", index=False)

print("noticed frequency", len(transition_rates[transition_rates["Noticed"] == True]))
print("not noticed frequency", len(transition_rates[transition_rates["Noticed"] == False]))
```

C.2 Saccade calculation

A snippet from the code calculating, in this case, the mean saccade length for the tunneling data.

```
import pandas as pd
import numpy as np
import math
import sys, os, glob
import re


noticed_files = []
unnoticed_files = []

for filename in os.listdir(l_ErgebniseOrdner_Eye):
    f = os.path.join(l_ErgebniseOrdner_Eye, filename)
    # checking if it is a file
    if os.path.isfile(f) and os.path.getsize(f) == 0:
        os.remove(f)
    elif "_noticed" in f and os.path.getsize(f) > 0:
        noticed_files.append(f)
    elif "_unnoticed" in f and os.path.getsize(f) > 0:
        unnoticed_files.append(f)

print(len(noticed_files), len(unnoticed_files))

noticed_data= {}
unnoticed_data= {}

for i in noticed_files:
    match = re.search(r'VP(\d{2})_(.*?)_(.*?)_(.*?)_', i)
    if match:
        dfName = f"VP{match.group(1)}_{match.group(3)}_{match.group(4)}"
    noticed_data[dfName] = pd.read_csv(i, sep=";")

for j in unnoticed_files:
    match = re.search(r'VP(\d{2})_(.*?)_(.*?)_(.*?)_', j)
    if match:
        dfName = f"VP{match.group(1)}_{match.group(3)}_{match.group(4)}"
    unnoticed_data[dfName] = pd.read_csv(j, sep=";")

len(noticed_data), len(unnoticed_data)

### Calculate saccades function
def calculate_saccades(AOI_dimensions, saccade_dict, i, eye_saccade_dict, noticed):

    if len(eye_saccade_dict) > 1:
        saccade_dict[i] = eye_saccade_dict[["Session", "StaticMask"]].copy()
        saccade_dict[i]["Session"] = saccade_dict[i]["Session"].apply(lambda x: i)
        # Where a fixation Ends, a saccade Starts, so Endtime in the
        # -> fixation data library = saccade start
        saccade_dict[i]["StartTime"] =
            eye_saccade_dict['EndTime'].str.replace(",", ".").astype(float).copy()
        saccade_dict[i]["StartX"] =
```

```
            eye_saccade_dict['CenterX'].str.replace(",", ".").astype(float).copy()
        saccade_dict[i]["StartY"] =
            eye_saccade_dict['CenterY'].str.replace(",", ".").astype(float).copy()
        saccade_dict[i]["EndX"] =
            eye_saccade_dict['CenterX'].str.replace(",", ".").astype(float).copy().shift(-1)
        saccade_dict[i]["EndY"] =
            eye_saccade_dict['CenterY'].str.replace(",", ".").astype(float).copy().shift(-1)

        # Check if the Statitc Mask of the current row is the same as the next row (-1)
        # -> means no transition happened
        # rewrite saccade_dict i to
        mask = saccade_dict[i]['StaticMask'] != saccade_dict[i]['StaticMask'].shift(-1)
        saccade_dict[i].loc[mask, ["StartX", "StartY" ,"EndX", "EndY"]] = 'transition'
        saccade_dict[i] = saccade_dict[i][~mask]
        saccade_dict[i] = saccade_dict[i].dropna()

        for index,row in saccade_dict[i].iterrows():
            ## Take the dimensions from the config file and multiply
            ## percent position values to actual dimensions
            aoi_index = AOI_dimensions.index[AOI_dimensions["AOI"]==
                saccade_dict[i].at[index,"StaticMask"]][0]
            AOI_x = float(AOI_dimensions.at[aoi_index,"SizeX"])
            AOI_y = float(AOI_dimensions.at[aoi_index,"SizeY"])

            StartX = saccade_dict[i].at[index,"StartX"]*AOI_x
            saccade_dict[i].at[index,"StartX"] = StartX

            StartY = saccade_dict[i].at[index,"StartY"]*AOI_y
            saccade_dict[i].at[index,"StartY"] = StartY

            EndX = saccade_dict[i].at[index,"EndX"]*AOI_x
            saccade_dict[i].at[index,"EndX"] = EndX

            EndY = saccade_dict[i].at[index,"EndY"]*AOI_y
            saccade_dict[i].at[index,"EndY"] = EndY

            ## Calculate Euclidean dist
            start = StartX - EndX
            end = StartY - EndY
            euclidean = math.sqrt(start**2 + end**2)
            saccade_dict[i].at[index,"Euclidean"] = euclidean
            saccade_dict[i].at[index,"Noticed"] = noticed

    return saccade_dict[i]

AOI_config_path = os.path.join(l_Config, "Planes_AVES0_11082023_dimensions.csv")
AOI_dimensions = pd.read_csv(AOI_config_path, sep=";", decimal =",")
saccade_dict = {}

for i,eye_df in noticed_data.items():
    noticed = True
    saccade_dict[i] = calculate_saccades(AOI_dimensions, saccade_dict, i,
        eye_df, noticed)
```

```
for i,eye_df in unnoticed_data.items():
    noticed = False
    saccade_dict[i] = calculate_saccades(AOI_dimensions, saccade_dict, i,
        eye_df, noticed)
```

C.3 Entropy Calculation

A snippet from the code calculating, in this case, the entropy values from the tunneling data. As previously stated, the code is based on the example[1] provided in a paper by Shiferaw et al. [57].

```
import pandas as pd
import numpy as np
import math
import sys, os, glob


noticed_files = []
unnoticed_files = []

for filename in os.listdir(l_ErgebnisOrdner_Eye):
    f = os.path.join(l_ErgebnisOrdner_Eye, filename)
    # checking if it is a file
    if os.path.isfile(f) and os.path.getsize(f) == 0:
        os.remove(f)
    elif "_noticed" in f and os.path.getsize(f) > 0:
        noticed_files.append(f)
    elif "_unnoticed" in f and os.path.getsize(f) > 0:
        unnoticed_files.append(f)

print(len(noticed_files), len(unnoticed_files))

noticed_data= {}
unnoticed_data= {}

for i in noticed_files:
    match = re.search(r'VP(\d{2})_(.*?)_(.*?)_(.*?)_', i)
    if match:
        dfName = f"VP{match.group(1)}_{match.group(3)}_{match.group(4)}"
    noticed_data[dfName] = pd.read_csv(i, sep=";")

for j in unnoticed_files:
    match = re.search(r'VP(\d{2})_(.*?)_(.*?)_(.*?)_', j)
    if match:
        dfName = f"VP{match.group(1)}_{match.group(3)}_{match.group(4)}"
    unnoticed_data[dfName] = pd.read_csv(j, sep=";")

## The entropy code is adjusted from Book Shiferaw's paper / Github:
## https://github.com/BrookShiferaw/entropy/blob/master/stationary_gaze_entropy.ipynb
minn = 1
sby = 0.1 # bin size in percent

entropy_results=
    pd.DataFrame(columns=["ParticipantID","Session", "Determinator",
        "Noticed", "Observed_H"])
index = 0
```

[1] https://github.com/BrookShiferaw/entropy/blob/master/stationary_gaze_entropy.ipynb, last accessed: Dec. 2023

```
for file_path,entropy in noticed_data.items():
    vp = file_path.split("_")[0]
    determiner = file_path.split("_")[-1]
    condition = file_path.split("_")[-2]

    pfd_entropy = entropy[entropy["StaticMask"] == "PFD_FO"].copy()
    pfd_entropy.reset_index(drop=True, inplace=True)
    pfd_entropy["CenterX"] =
        pfd_entropy["CenterX"].str.replace(",", ".").astype(float)
    pfd_entropy["CenterY"] =
        pfd_entropy["CenterY"].str.replace(",", ".").astype(float)

    N = len(pfd_entropy)
    pfd_entropy["CenterX_range"] =
        pd.cut(pfd_entropy.CenterX, np.arange(0, s, sby), right=False)
    pfd_entropy["CenterY_range"] =
        pd.cut(pfd_entropy.CenterY, np.arange(0, s, sby), right=False)

    pfd_entropy_grouped= pfd_entropy.groupby
        (['CenterX_range','CenterY_range']).size().reset_index()
            .rename(columns={0:'count'})
    pfd_entropy_grouped =
        pfd_entropy_grouped[pfd_entropy_grouped['count'] != 0].copy().reset_index()
    pfd_entropy_grouped['p']=
        pfd_entropy_grouped['count']/pfd_entropy_grouped['count'].sum()
    p_by_log = list()

    for i in np.array(pfd_entropy_grouped['p']):
        p_by_log.append(math.log2(i)*i)

    pfd_entropy_grouped['p*log(p)']= p_by_log
    observed_h= abs(pfd_entropy_grouped['p*log(p)'].sum())
    normalised_h= abs(pfd_entropy_grouped['p*log(p)'].sum()/math.log2(s/sby*s/sby))
    new_row = pd.DataFrame({"ParticipantID":vp, "Session":file_path, "Noticed":True,
        "Determinator":determiner ,"Condition":condition,
            "Normalised_H": normalised_h ,"Observed_H":observed_h}, index=[0])
    entropy_results= pd.concat([entropy_results, new_row]).reset_index(drop=True)


for file_path,entropy in unnoticed_data.items():
    vp = file_path.split("_")[0]
    determiner = file_path.split("_")[-1]
    condition = file_path.split("_")[-2]

    pfd_entropy = entropy[entropy["StaticMask"] == "PFD_FO"].copy()
    pfd_entropy.reset_index(drop=True, inplace=True)
    pfd_entropy["CenterX"] =
        pfd_entropy["CenterX"].str.replace(",", ".").astype(float)
    pfd_entropy["CenterY"] =
        pfd_entropy["CenterY"].str.replace(",", ".").astype(float)

    N = len(pfd_entropy)
    pfd_entropy["CenterX_range"] =
```

```
    pd.cut(pfd_entropy.CenterX, np.arange(0, s, sby), right=False)
pfd_entropy["CenterY_range"] =
    pd.cut(pfd_entropy.CenterY, np.arange(0, s, sby), right=False)

pfd_entropy_grouped= pfd_entropy.groupby
    (['CenterX_range','CenterY_range']).size().reset_index()
        .rename(columns={0:'count'})
pfd_entropy_grouped =
    pfd_entropy_grouped[pfd_entropy_grouped['count'] != 0].copy().reset_index()
pfd_entropy_grouped['p']=
    pfd_entropy_grouped['count']/pfd_entropy_grouped['count'].sum()
p_by_log = list()

for i in np.array(pfd_entropy_grouped['p']):
    p_by_log.append(math.log2(i)*i)

pfd_entropy_grouped['p*log(p)']= p_by_log
observed_h= abs(pfd_entropy_grouped['p*log(p)'].sum())
normalised_h= abs(pfd_entropy_grouped['p*log(p)'].sum()/math.log2(s/sby*s/sby))
new_row = pd.DataFrame({"ParticipantID":vp, "Session":file_path, "Noticed":False,
    "Determinator":determiner ,"Condition":condition,
        "Normalised_H": normalised_h ,"Observed_H":observed_h}, index=[0])
entropy_results= pd.concat([entropy_results, new_row]).reset_index(drop=True)
```

C.4 SVM Pipeline

An overview of the SVM pipeline utilized, in this case, for the tunneling data. As mentioned earlier, the pipeline has been developed based on examples from two tutorials[2,3].

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from sklearn import svm
from sklearn.model_selection import KFold, cross_val_score, cross_val_predict
from sklearn import metrics
from sklearn.metrics import confusion_matrix, precision_score

df =  pd.read_csv(l_EyeAnalysis + "\\merged_ISA_airbus_only.csv", sep=";", decimal =",")
df["Mean_Euclidean"] = df["Mean_Euclidean"].astype(float)

# Drop unnecessary columns + make sure there are no missing values
df.drop(["Condition", "ISA", "Normalised_H", "Determinator", "ParticipantID"],
    axis=1, inplace=True)
df = df.dropna()

# Randomize the sample order
df = df.sample(frac=1)
# Create a list of the actual classes
tunnel = list(df["Tunnel"])
df.drop("Tunnel", axis=1, inplace=True)

# Centralize the data
for c in df.columns:
    mean = df[c].mean()
    std = np.std(df[c])
    df[c] = (df[c] - mean)/std
    print(df[c])

# List of the train/test eye-tracking data
list_val = df.values.tolist()

# Create kFold parameters
cv = KFold(n_splits=5)
# Fit model
svm_classifier = svm.SVC(kernel='rbf', C=1)

# Evaluate model
acc_score = cross_val_score(svm_classifier, list_val, tunnel,
    scoring='accuracy', cv=cv, n_jobs=-1, error_score="raise")
prec_score = cross_val_score(svm_classifier, list_val, tunnel,
    scoring='precision', cv=cv, n_jobs=-1, error_score="raise")

print("Cross Validation Accuracy Scores: ", acc_score)
```

[2] https://github.com/zhailat/Introduction-to-machine-learning-Python/tree/b6eddb8ff52797e318afb07686cc53e59b443890/Part%2009%20-%20Constructing%20Multi-Class%20Classifier%20Using%20SVM%20with%20Python, last accessed: Dec. 2023

[3] https://machinelearningmastery.com/k-fold-cross-validation/, last accessed: Dec. 2023

```
print("Cross Validation Precision Scores: ", prec_score)

print("Average Acuracy CV Score: ", acc_score.mean())
print("Average Precision CV Score: ", prec_score.mean())


# Plot confusion matrix
plt.rcParams.update({'font.size':12})
plt.rcParams.update({'axes.labelsize':14})

predictions = cross_val_predict(svm_classifier, list_val, tunnel, cv=cv)
prec = precision_score(tunnel, predictions, average='macro')

confusies = metrics.confusion_matrix(tunnel, predictions)

cm_display = metrics.ConfusionMatrixDisplay(confusion_matrix = confusies,
    display_labels = [False, True])
plotsie= cm_display.plot(cmap = plt.cm.BuPu, colorbar=False)
plt.xlabel("Predicted label",labelpad=10)
plt.ylabel("True label",labelpad=10)

plt.show()
```

C.5 LR Pipeline

An overview of the LR pipeline utilized, in this case, for the tunneling data. As mentioned earlier, the pipeline has been developed based on examples from two tutorials[4,5].

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import KFold, cross_val_score, cross_val_predict
from sklearn import metrics
from sklearn.metrics import confusion_matrix, precision_score

df =  pd.read_csv(l_EyeAnalysis + "\\merged_ISA_airbus_only.csv", sep=";", decimal =",")
df["Mean_Euclidean"] = df["Mean_Euclidean"].astype(float)

# Drop unnecessary columns + make sure there are no missing values
df.drop(["Condition", "ISA", "Normalised_H", "Determinator",
    "ParticipantID"], axis=1, inplace=True)
df = df.dropna()

# Randomize the sample order
df = df.sample(frac=1)
# Create a list of the actual classes
tunnel = list(df["Tunnel"])
df.drop("Tunnel", axis=1, inplace=True)

# Centralize the data
for c in df.columns:
    mean = df[c].mean()
    std = np.std(df[c])
    df[c] = (df[c] - mean)/std
    print(df[c])

# List of the train/test eye-tracking data
list_val = df.values.tolist()

# Create kFold parameters
cv = KFold(n_splits=5, random_state=1, shuffle=True)
# Fit model
logistic_model = LogisticRegression(C=1, penalty='l2',solver='lbfgs', random_state=0)

# Evaluate model
acc_score = cross_val_score(logistic_model, list_val, tunnel,
    scoring='accuracy', cv=cv, n_jobs=-1, error_score="raise")
prec_score = cross_val_score(logistic_model, list_val, tunnel,
    scoring='precision', cv=cv, n_jobs=-1, error_score="raise")

print("Cross Validation Accuracy Scores: ", acc_score)
```

[4] https://github.com/zhailat/Introduction-to-machine-learning-Python/tree/b6eddb8ff52797e318afb07686cc53e59b443890/Part%2009%20-%20Constructing%20Multi-Class%20Classifier%20Using%20SVM%20with%20Python, last accessed: Dec. 2023

[5] https://machinelearningmastery.com/k-fold-cross-validation/, last accessed: Dec. 2023

```
print("Cross Validation Precision Scores: ", prec_score)

print("Average Acuracy CV Score: ", acc_score.mean())
print("Average Precision CV Score: ", prec_score.mean())
print("Number of CV Scores used in Average: ", len(prec_score))

# Plot confusion matrix
plt.rcParams.update({'font.size':12})
plt.rcParams.update({'axes.labelsize':14})

predictions = cross_val_predict(logistic_model, list_val, tunnel, cv=cv)
prec = precision_score(tunnel, predictions, average='macro')

confusies = metrics.confusion_matrix(tunnel, predictions)

cm_display = metrics.ConfusionMatrixDisplay(confusion_matrix = confusies,
    display_labels = [False, True])
plotsie= cm_display.plot(cmap = plt.cm.BuPu, colorbar=False)
plt.xlabel("Predicted label",labelpad=10)
plt.ylabel("True label",labelpad=10)

plt.show()
```

C.6 TPOT Pipeline

An overview of the TPOT pipeline utilized, in this case, for the tunneling data. As mentioned earlier, the pipeline has been developed based on examples from two tutorials[6,7].

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from tpot import TPOTClassifier
from sklearn.naive_bayes import BernoulliNB
from sklearn.pipeline import make_pipeline
from sklearn.kernel_approximation import RBFSampler
from sklearn.model_selection import KFold, cross_val_score, cross_val_predict
from sklearn import metrics
from sklearn.metrics import confusion_matrix, precision_score
from sklearn.model_selection import train_test_split, RepeatedStratifiedKFold

## First the tpot tool was run, then the cross-validation:
df_tpot =  pd.read_csv(l_EyeAnalysis + "\\merged_ISA_airbus_only.csv", sep=";",
    decimal =",")
df_tpot["Mean_Euclidean"] = df_tpot["Mean_Euclidean"].astype(float)

# Drop unnecessary columns + make sure there are no missing values
df_tpot.drop(["Condition", "ISA", "Normalised_H", "Determinator",
    "ParticipantID"], axis=1, inplace=True)
df_tpot = df_tpot.dropna()

# Randomize the sample order
df_tpot = df_tpot.sample(frac=1)
# Copy the data for the later k-fold validation
df = df_tpot

# Separate in two datasets
train_dataset, test_dataset = train_test_split(df_tpot, test_size=0.2)
print(train_dataset.shape)

# Generate overview
train_stats = train_dataset.describe()
train_stats = train_stats.transpose()

# Take out the correct labels
test_labels = test_dataset.pop("Tunnel")
train_labels = train_dataset.pop("Tunnel")

# Center the data
def center_dat(val):
    val = (val - train_stats["mean"]) / train_stats["std"]
    return val
```

[6] https://github.com/zhailat/Introduction-to-machine-learning-Python/tree/b6eddb8ff52797e318afb07686cc53e59b443890/Part%2009%20-%20Constructing%20Multi-Class%20Classifier%20Using%20SVM%20with%20Python, last accessed: Dec. 2023

[7] https://machinelearningmastery.com/k-fold-cross-validation/, last accessed: Dec. 2023

```
center_test = center_dat(test_dataset)
center_test.drop("Tunnel", axis=1, inplace=True)
center_train = center_dat(train_dataset)
center_train.drop("Tunnel", axis=1, inplace=True)

## First the TPOT tool was run, then the pipeline was continued based on the suggestion
cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
model = TPOTClassifier(generations=5, population_size=100, verbosity=2)
model.fit(center_test, test_labels)
model.export('tpot_sonar_best_model.py')

## Cross-Validation:
# Prepare the data for the cross-validation
# Create a list of the actual classes
tunnel = list(df["Tunnel"])
df.drop("Tunnel", axis=1, inplace=True)

# Centralize the data
for c in df.columns:
mean = df[c].mean()
std = np.std(df[c])
df[c] = (df[c] - mean)/std
print(df[c])

# List of the train/test eye-tracking data
list_val = df.values.tolist()

# Set kFold parameters
cv = KFold(n_splits=5, random_state=1, shuffle=True)

# Best pipeline: BernoulliNB(RBFSampler(input_matrix, gamma=0.1),
#    alpha=0.01, fit_prior=False)
bern = BernoulliNB(alpha=0.01, fit_prior=False)

# Fit model
exported_pipeline = make_pipeline(
    RBFSampler(gamma=0.1),
    BernoulliNB(alpha=0.01, fit_prior=False),
)

exported_pipeline.fit( list_val, tunnel)

# Evaluate model
acc_score = cross_val_score(exported_pipeline, list_val, tunnel,
    scoring='accuracy', cv=cv, n_jobs=-1, error_score="raise")
prec_score = cross_val_score(exported_pipeline, list_val, tunnel,
    scoring='precision', cv=cv, n_jobs=-1, error_score="raise")

print("Cross Validation Accuracy Scores: ", acc_score)
print("Cross Validation Precision Scores: ", prec_score)

print("Average Acuracy CV Score: ", acc_score.mean())
print("Average Precision CV Score: ", prec_score.mean())
```

```
print("Number of CV Scores used in Average: ", len(prec_score))

# Plot confusion matrix
plt.rcParams.update({'font.size':12})
plt.rcParams.update({'axes.labelsize':14})

predictions = cross_val_predict(exported_pipeline, list_val, tunnel, cv=cv)
prec = precision_score(tunnel, predictions, average='macro')

confusies = metrics.confusion_matrix(tunnel, predictions)

cm_display = metrics.ConfusionMatrixDisplay(confusion_matrix = confusies,
    display_labels = [False, True])
plotsie= cm_display.plot(cmap = plt.cm.BuPu, colorbar=False)
plt.xlabel("Predicted label",labelpad=10)
plt.ylabel("True label",labelpad=10)
```

C.7 Example of the Individual Test and Train Sessions

An example of the data preparation, fitting and testing for the exploratory individual examinations. As mentioned earlier, the pipeline has been developed based on an example[8] from a tutorial by Zeyad Hailat [92].

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from tpot import TPOTClassifier
from sklearn.naive_bayes import BernoulliNB
from sklearn.pipeline import make_pipeline
from sklearn.kernel_approximation import RBFSampler
from sklearn.metrics import confusion_matrix, precision_score
from sklearn.model_selection import train_test_split, RepeatedStratifiedKFold

df =  pd.read_csv(l_EyeAnalysis + "\\merged_ISA_airbus_only.csv", sep=";", decimal =",")
df["Mean_Euclidean"] = df["Mean_Euclidean"].astype(float)

# Drop unnecessary columns + make sure there are no missing values
df.drop(["Condition", "ISA", "Normalised_H", "Determinator",
    "ParticipantID"], axis=1, inplace=True)
df = df.dropna()

# Randomize the sample order
df = df.sample(frac=1)

## Prepare the data for the TPOT tool:
# Separate in two datasets
train_dataset, test_dataset = train_test_split(df, test_size=0.2)
print(train_dataset.shape)

# Generate overview
train_stats = train_dataset.describe()
train_stats = train_stats.transpose()

# Take out the correct labels
test_labels = test_dataset.pop("Tunnel")
train_labels = train_dataset.pop("Tunnel")

# Center the data
def center_dat(val):
    val = (val - train_stats["mean"]) / train_stats["std"]
    return val

center_test = center_dat(test_dataset)
center_test.drop("Tunnel", axis=1, inplace=True)
center_train = center_dat(train_dataset)
center_train.drop("Tunnel", axis=1, inplace=True)
```

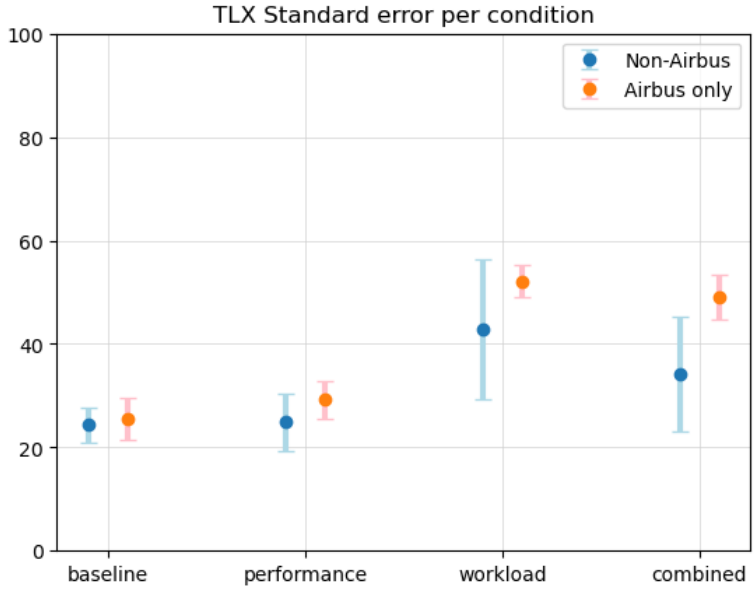
[8] https://github.com/zhailat/Introduction-to-machine-learning-Python/tree/b6eddb8ff52797e318afb07686cc53e59b443890/Part%2009%20-%20Constructing%20Multi-Class%20Classifier%20Using%20SVM%20with%20Python, last accessed: Dec. 2023

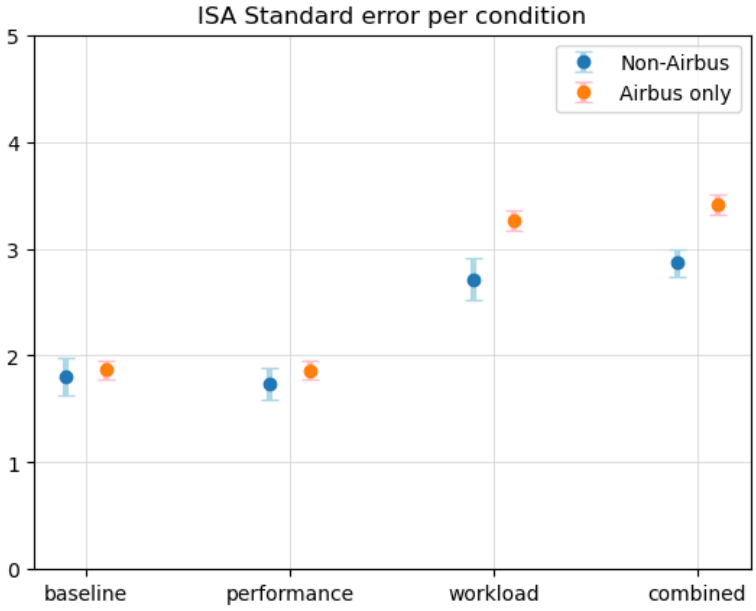
```
## Firs the TPOT tool was run, then the pipeline was continued based on the suggestion
cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
model = TPOTClassifier(generations=5, population_size=100, verbosity=2)
model.fit(center_test, test_labels)
model.export('tpot_sonar_best_model.py')


# Fit the model
# Best pipeline: BernoulliNB(RBFSampler(input_matrix, gamma=0.1),
#    alpha=0.01, fit_prior=False)

exported_pipeline = make_pipeline(
    RBFSampler(gamma=0.1),
    BernoulliNB(alpha=0.01, fit_prior=False),
)

exported_pipeline.fit(center_train, train_labels)

train_pred =  exported_pipeline.predict(center_train)
test_pred = exported_pipeline.predict(center_test)


# Print performance
print('Accuracy of the classifier on train set: {:.2f}'
    .format(exported_pipeline.score(center_train, train_labels)))
print('Accuracy of the classifier on test set: {:.2f}'
    .format(exported_pipeline.score(center_test, test_labels)))
print('Precision of the classifier on test set: {:.2f}'
    .format(precision_score(test_labels, test_pred, average='macro')))

# Plot confusion matrix
ax= plt.subplot()
cm = confusion_matrix(test_labels, test_pred)
sns.heatmap(cm, annot=True, ax = ax);
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
ax.set_xlabel('Predicted labels', fontsize = 14, labelpad=10);
ax.set_ylabel('True labels', fontsize = 14, labelpad = 10);
```

D Appendix: Comparison Airbus vs. non-Airbus

D.1 Airbus vs. non-Airbus Workload

An overview of the workload self-assessment responses per condition between participants with an Airbus certification and without.





D.2 Airbus vs. non-Airbus Entropy

An overview of the calculated entropy per condition between participants with an Airbus certification and without.