Deliverable D6.4

Second validation report (VAL2) and demonstrator (DEM2)

Author: Carl Westin (LiU)

Major contributors: Roberto Venditti (DBL), Anaisa Villani (EMBRT), Miguel Villegas (Skyway), Philip Wright (ENG), Evangelos Spyrou (CERTH), Anaisa Villani (EMBRT), Turkan Hentati (CATIE), Charles Dormoy (CATIE), Jaime, Diaz Pineda (THAL), Jean-Paul Imbert (ENAC), Mansi Sharma (DFKI), Florian Daiber (DFKI), Maurice Rekrut (DFKI).

Reviewed by: Brian Hilburn (CHPR), Vanessa Arrigoni (DBL), Simone Pozzi (DBL)

Abstract:

This document provides the fourth and final deliverable in WP6 – *D6.4: Second validation report (VAL2) and demonstrator (DEM2).* The document contains the results from the second validation activities for all six HAIKU use cases. This includes descriptions of the finalized concepts for each of the six HAIKU Use Cases (UCs), a description of the TRL progression, validation objectives, methods and results. To harmonize use cases validation approaches, and to be able to relate them to the HAIKU overall research questions, the EASA Human-AI Teaming requirements have been used as a common reference that each use cases validation objectives are cross referenced against. Each UC has conducted validation exercises involving target end users in their respective aviation domain. UC results are discussed in relation to the HAIKU high level research questions, with this report focusing on human factors requirements for HAT and lessons learned from building AI intelligent assistants using small data samples.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Information Table

Deliverable Number	D6.4
Deliverable Title	Second validation report (VAL2) and demonstrator (DEM2)
Version	1.0
Status	Final
Responsible Partner	LiU
Contributors	UC1: ENAC, DFKI UC2: TAVS, CATIE, Bordeaux INPS, EMBRT UC3: LiU, LFV UC4: Skyway, DBL UC5: ECTL, ENG, Suite5 UC6: CERTH Whole document: LiU, DBL
Contractual Date of Delivery	May 31st, 2025
Actual Date of Delivery	June 6th, 2025
Dissemination Level	Public

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Document History

Version	Date	Status	Author	Description
0.1	Nov 5th, 2025	Repository Template & Deliverable outline	DBL and LiU	Template for all partners
0.2	Jan 09th-Apr 18th, 2025	Draft contribution in repository	UC Leaders, DBL	UC contributions in Repository with iterative reviews to ensure consistency
0.3	April 30th, 2025	First draft	UC Leaders	Contributions from all partners
0.4	May 16th, 2025	Second draft	Venditti R. (DBL), Westin C. (LiU)	Partner contributions reviewed with structure fine-tuning to improve consistency. Introduction added
0.5	May 19th, 2025	Third draft	Westin C. (LiU)	Document sent for internal review.
0.6	May 21st, 2025	Fourth draft	Hilburn B. (CHPR)	Review with comments
0.7	May 28th, 2025	Fifth draft	UC Leaders, Westin C. (LiU)	UC section fine-tuning according to received comments. Conclusions added
0.8	May 29th, 2025	Sixth draft	Arrigoni V. (DBL)	Final quality check with minor comments
1.0	June 5th, 2025	Final version	Westin C. (LiU)	Final version for submission

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

List of Acronyms

Acronym	Definition
AAM	Advanced Air Mobility
AI	Artificial Intelligence
AltMOC	Alternative Means of Compliance
AMC	Acceptable Means of Compliance
ASW	Airport Safety Watch
ATCO	Air Traffic Control Operator
ATM	Air Traffic Management
CISP	Common information service provider
CLT	Construal Level Theory
CRM	Crew Resource Management
DA	Digital Assistant
DM	Decision-making
DUC	Digital Assistant for UAM Coordinator
EASA	European Union Aviation Safety Agency
FIFO	First-In-First-Out
HAIKU	Human AI Knowledge and Understanding
HAIT	Human AI Teaming
HAT	Human Autonomy Teaming
HATR	Human Autonomy Teaming Requirements
HAZOP	Hazard and Operability Study
HF	Human Factors

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

HITL	Human-In-The-Loop
HLR	High Level Requirements
HMI	Human Machine Interface
HPC	High-Performance Computing
IA	Intelligent Assistant
IoT	Internet of Things
IR	Implementing Rules
ISA	Intelligent Sequence Assistant
JRCC	Joint Rescue Coordination Center
KPA	Key Performance Area
KPI	Key Performance Index
LACC	Levels-of-autonomy-in-cognitive-control
LIFO	Last-In First-Out
LLA	London Luton Airport
LOA	Levels Of Automation
LOC-I	Loss Of Control In Flight
M	Mean
MbC	Management by Consent
MbE	Management by Exception
ML	Machine Learning
MOC	Means of Compliance
MoE	Measures of Effectiveness
MoP	Measures of Performance

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

NLP	Natural Language Processing
OliviA	OperationaL Intentions adVIser for Aviation
P	p-value
r	Pearson's correlation
R&D	Research and Development
SD	Standard Deviation
SA	Situation Awareness
SHAP	Shapley Additive Explanations
TEM	Threat and Error Management
TRL	Technology Readiness Level
UAM	Urban Air Mobility
UAS	Unmanned Aircraft System
UC	Use Case
UPRT	Upset Prevention and Recovery Training
UTM	Unmanned Aircraft System Traffic Management
VAL1	First Validation
VAL2	Second Validation
VTOL	Vertical Take-off and Landing
W	Wilcoxon test statistic
WP	Work Package
XAI	Explainable Artificial Intelligence

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Table of contents

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Executive Summary

The HAIKU project aims to pave the way for human-centric Intelligent Assistants (IAs) in the aviation domain by developing AI enabled prototypes for six aviation-related use cases.

- Use Case #1 – Flight Deck Startle Response
- Use Case #2 – Flight Deck Route Planning/Replanning
- Use Case #3 – Urban Air Mobility
- Use Case #4 – Digital and Remote Tower
- Use Case #5 – Airport Safety Watch
- Use Case #6 – Airport Spreading Virus Prevention

WP6 has two primary objectives: to evaluate the progress of the Intelligent Assistant concepts and prototypes, and to assess the final prototypes by providing empirical evidence of their operational benefits. Deliverable D6.4, *Second validation report (VAL2) and demonstrator (DEM2)*, supports these objectives by detailing the results from the final VAL2 activities.

This report, following the previous deliverables D6.1, D6.2, and D6.3, covers an update of the following activities:

- **Technology Readiness Level.** Update for all use cases, demonstrating the advancement of prototype readiness for VAL2 and the remainder of the project.
- **Validation plan and strategy** as performed by each use case, with specific validation objectives, activities, and metrics based on the scope and goals of each case.
- **Validation results.** Presenting the results from all use cases VAL2 activities, in relation to the shared validation strategy, harmonised against the EASA Human-AI Teaming requirements.
- **HAIKU high-level research questions.** Relating results and lessons learned across the six use cases to answer the HAIKU high-level research questions.

The structure of this deliverable builds on its predecessor D6.3: Updated validation strategy and plan.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Introduction

This report describes the work conducted under HAIKU Task 6.7: Second Validation (VAL2). The objective of this task is to assess the proposed Intelligent Assistant (IA) concepts for each use case (UC) at a higher Technology Readiness Level (TRL) than previously explored in VAL1. The goal has been to test IAs at TRL 4–6 with target end-users and realistic scenarios, through prototype demonstrations in relevant operational environments, such as high-fidelity simulators or real-world settings. This report presents the results of all VAL2 activities carried out following VAL1 and is structured to provide dissemination-ready sections for each UC, highlighting key findings.

The document is organized into 9 sections:

- **Section 1** summarises the theoretical framework used to harmonize and guide the UCs' validation strategies and plans (originally presented in detail in D6.3)[1], outlines the HAIKU research questions, introduces the validation strategy, and provides URL links to the UC validation demonstrators.
- **Sections 2 to 7** describe the six use cases:
  - **Section 2**: UC#1 – Flight Deck Startle Response
  - **Section 3**: UC#2 – Flight Deck Route Planning/Replanning
  - **Section 4**: UC#3 – Urban Air Mobility
  - **Section 5**: UC#4 – Digital and Remote Tower
  - **Section 6**: UC#5 – Airport Safety Watch
  - **Section 7**: UC#6 – Airport Spreading Virus Prevention

  Each UC section presents the status of the IA prototype on the TRL scale, the validation strategy and methods, including validation objectives, research questions, VAL2 activities, participants, scenarios, measures, and data analysis approach. Results are presented in alignment with the shared validation strategy and harmonized against the EASA HAT requirements. Each UC concludes with a discussion linking their VAL2 results to answer their own research questions, the HAIKU high-level research questions, and provide recommendations for future research. While UC sections follow this common structure to maintain consistency, unique requirements and variations across the UCs necessitated some deviations from the standard format. More specifically:

  - UC1 includes three separate VAL2 activities, each tied to a distinct experiment exploring different facets of the UC1 IA prototype, FOCUS. Each activity is described in terms of its objectives, methods, results, and discussion.
  - UC2, UC3, and UC4 each report on a single validation activity and follow a similar structure.

[1] This framework builds on the EASA Human-AI Teaming (HAT) requirements (2024) and aligns them with HAIKU's high-level research questions and objectives.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- UC5 focuses on validation objectives rooted in functional, real-world applications rather than controlled experiments, aligning with a Technology Readiness Level (TRL) progression. As such, its structure reflects the steps taken to advance the IA prototype toward TRL9.
- UC6 also reports on a single validation activity. However, UC6 validation approach differs since the IA prototype, being designed for use by the general public (e.g., airport passengers), differs from those in other UCs, which are intended for professional aviation stakeholders such as pilots, air traffic controllers, U-space traffic managers (UAM Coordinators), and airport safety personnel.

- **Section 8** provides a general conclusion, synthesizing insights from all six UCs in relation to the EASA HAT requirements and the HAIKU research questions.
- **Section 9** contains the Annexes, which provide more detailed information on the UC activities for readers interested in exploring the work in greater depth. Only UC2, UC3, and UC4 has provided mate

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

1. Validation Strategy

To create a unified validation strategy across UCs, a validation framework was designed building on the EASA HAT requirements. As outlined in D4.1 (HAIKU, 2024), each UC was aligned with the EASA HAT requirements derived from EASA's *Roadmap for Trustworthy AI* (EASA, 2020) and its *Initial Guidance for AI in Aviation Applications, Version 1* (EASA, 2023). In D6.3, UCs were asked to identify which of these EASA HAT requirements (initially covered at a conceptual level) would be taken forward for practical validation in VAL2.

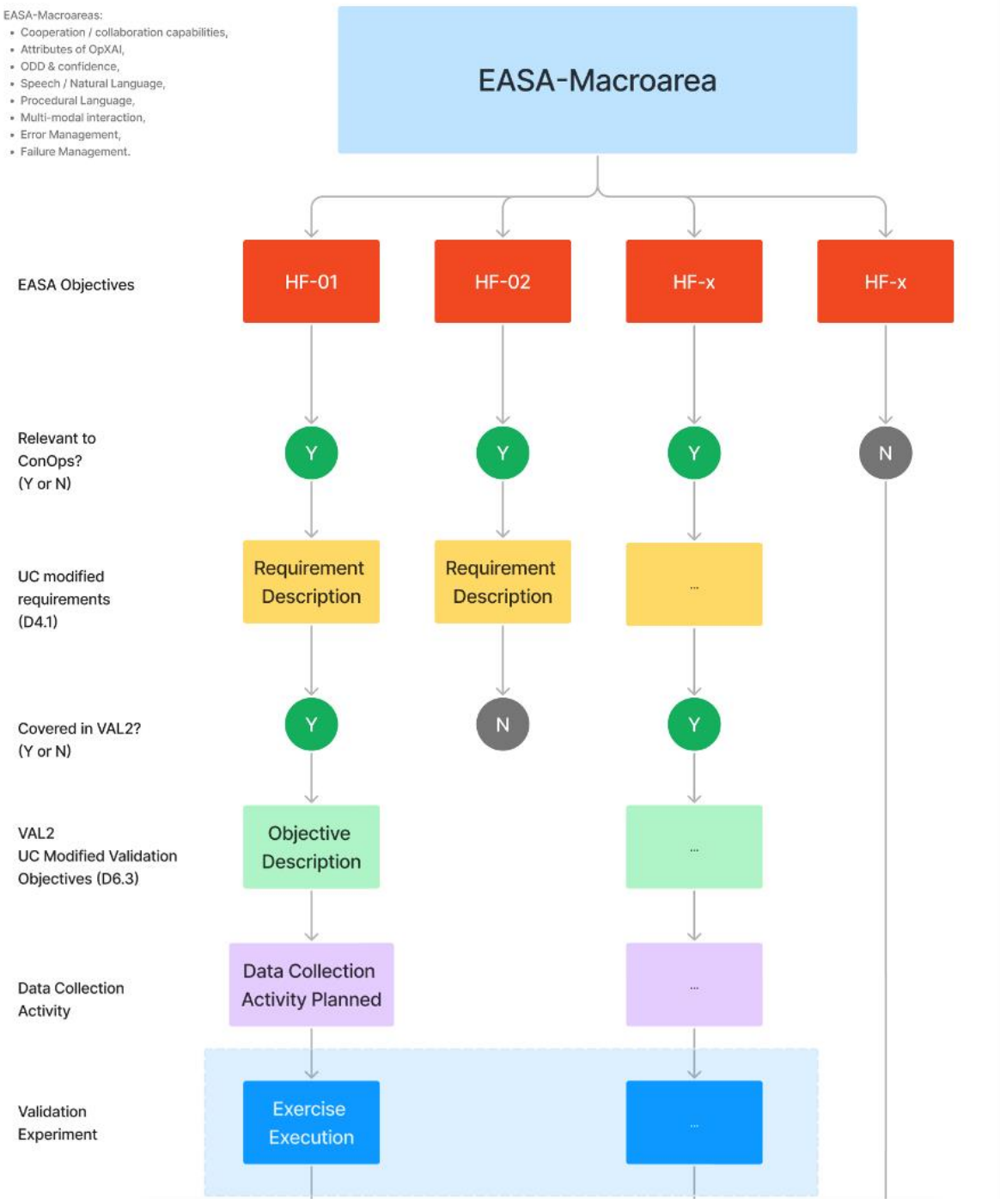
Figure 1 illustrates how each UC developed its validation approach by following the validation framework, progressing from the conceptual validation requirements (expressed as high-level requirements and HAT requirements in D4.1) to defining concrete validation objectives, activities, and associated metrics for VAL2 (presented in D6.3). The framework begins with the EASA HAT requirements and traces how these informed the conceptual design of the Intelligent Assistants (IAs) within each UC. The HAIKU project focused primarily on requirements within the EASA macro-areas of Human Factors and Explainability. As part of the validation design process documented in D6.3, each UC was asked to indicate, via a two-step yes/no decision process, if:

1. A given requirement was conceptually addressed in the IA design,
2. That requirement would be empirically validated during VAL2.

For all requirements selected for validation, UCs specified the corresponding validation objectives, data collection activities, and performance metrics.

This deliverable presents how each UC implemented this validation strategy, reporting results in direct relation to the EASA HAT requirements (indicated by the last two boxes in Figure 1). In addition, each UC discusses how its findings contribute to answering the HAIKU high-level research questions. These individual insights are then aggregated across all UCs to support a project-wide synthesis, drawing broader conclusions on HAT recommendations in aviation.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332



This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

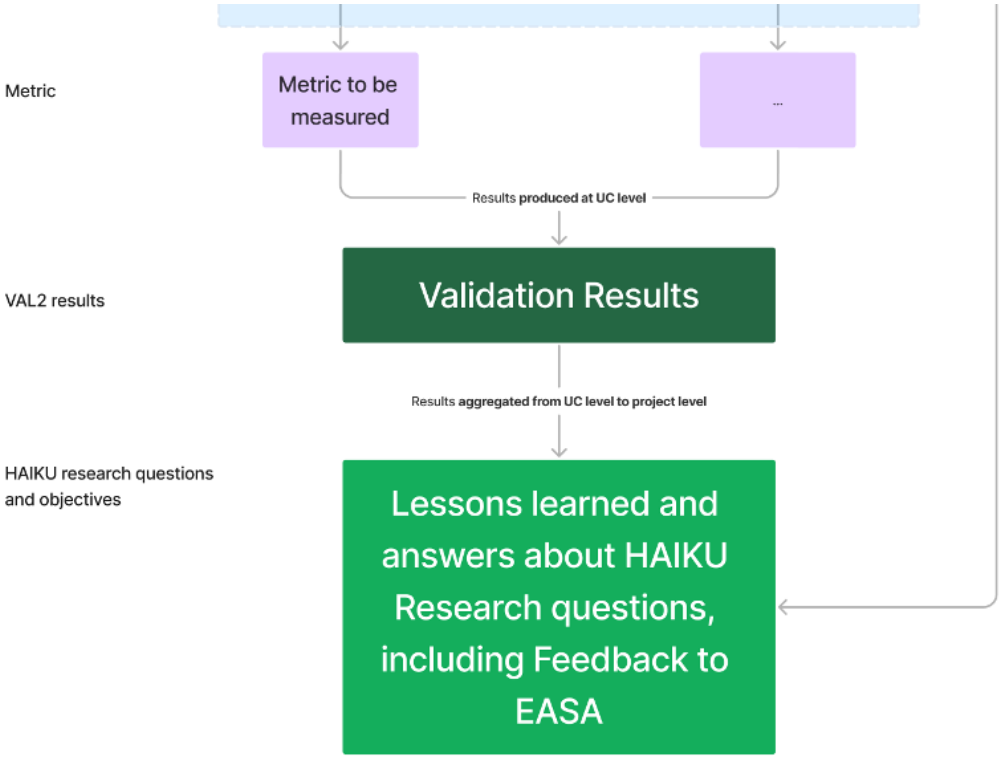


Figure 1. HAIKU validation framework

Each UC conducted one or more data collection activities as part of their VAL2, involving target stakeholders or end users to evaluate the IA prototypes. Table 1 provides an overview of the timing and the number of participants involved in the experiments conducted across the different UCs.

Table 1. UC VAL2 experiment schedule and recruited participants

UC	Experiment dates	Stakeholder and nr. of participants
#1	Q3-Q4, 2024	87 general population and 12 Commercial Airline Pilots
#2	Q1, 2025	10 Commercial Airline Pilots
#3	Q1, 2025	9 Air Traffic Controllers
#4	Q1-Q2, 2025	8 Air Traffic Controllers
#5	Q3-Q4, 2024	22 participants from national airport authorities, ground handling providers, aviation consultants, regulatory bodies, airlines, and technology service providers
#6	Q1-Q3, 2025	10 general population

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

1.1.1. HAIKU high-level research questions

The HAIKU project is guided by three overarching high-level research questions that provide a common framework across all use cases:

HAIKU Research Question 1 (HAIKU Q1): *What is the recommended human-AI relationship for each of the different AI aviation applications?*

> This question explores the appropriate role and level of autonomy for AI systems, its intended purposes, expected operational benefits, and the underlying values and design principles across the six different aviation UCs. It also explores how AI systems should interact with human operators, ensuring alignment with operational goals and user expectations, to support safe, effective, and trusted operations.

HAIKU Research Question 2 (HAIKU Q2): *What does it mean for AI to be explainable?*

> This question explores the form, depth, and necessity of explainability in building operator trust and understanding. In some contexts, explainability may mean that the AI's behaviour is intuitive and aligns with human reasoning, not altering the task's nature. In more complex scenarios, AI reasoning might diverge from human logic due to the complexity and breadth of variables involved.

HAIKU Research Question 3 (HAIKU Q3): *How to train AI to assist humans in safety critical tasks when training data are insufficient?*

> This question addresses the critical challenge of data availability in AI system development, particularly in safety-critical environments where access to high-quality, representative data is limited. Rare or unforeseen events are especially problematic, as they occur infrequently and therefore provide little training data. Yet, these events often involve high uncertainty and demand critical, time-sensitive decisions.

Each UC explores its own specific research questions, but all are guided by the three overarching HAIKU questions that provide a common investigative framework. UCs contribute with context-specific insights based on their individual IA concepts, prototypes, and evaluation focus. While their engagement with each high-level question varies, their contributions are complementary and together help build a more comprehensive understanding of HAT in aviation.

Finally, findings related to the second high-level research question (HAIKU Q2), focused on explainability, are addressed in detail in Work Package 5, and reported in deliverable D5.2: *Case Studies and Results of Validation Activities.*

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

1.1.2. EASA Validation Requirements

Table 2 below overviews the UC IA prototype requirements mapped against the EASA HAT requirements. Note that VAL2 requirements are a subset of the IA concept requirements. As such, a green box indicates that the EASA objective is relevant to both the UC IA concept (check mark in light grey box) and VAL2 requirements (checkmark in green box). An empty (white) box indicates that the EASA objective is not addressed by UCs at all.

Within the EASA macro areas, requirements HF-01 – HF-10 focus on the IA's ability to engage in human-AI teaming through cooperation and collaboration. These include situational awareness, situation identification and diagnostics, and decision-making processes such as evaluating user proposals, negotiation, and adaptiveness. The HAIKU project demonstrates particularly strong coverage of situational awareness requirements (HF-01– HF-03) and selected aspects of decision-making (HF-07 – HF-08). However, it shows more limited coverage of negotiation, adaptiveness, and situation identification/diagnostics.

HAIKU does not sufficiently address spoken natural language communication requirements (HF-11 – HF-17) and entirely omits gesture-based non-verbal communication (HF-18 – HF-21). Requirements related to multi-modal interaction (HF-22 – HF-24) are also underrepresented.

Regarding human error tolerance, HAIKU partially addresses HF-28 and HF-29, which relate to the AI system's ability to manage end-user errors. However, it lacks coverage of design-induced and operational error management. Furthermore, failure management requirements (HF-31 – HF-34) are not addressed in HAIKU.

On the topic of operational explainability, HAIKU has good coverage across most requirements, except for EX-P14 and EXP-18 – EXP-19. These gaps relate to the user's ability to customize the level of explanation detail and the provision of training or guidance on how to monitor system performance effectively.

Table 2. mapping of UC IA prototype requirements, both at concept level (in grey) and for VAL2 (in green) against EASA objectives.

Obj.	Short description	UC1	UC2	UC3	UC4	UC5	UC6
EXP-10	Characterise explainability needs	✓	✓	✓	✓		
EXP-11	Clear and unambiguous presentation of explanations	✓	✓	✓	✓		✓
EXP-12	Demonstrate relevance of explanation for decision/action	✓	✓	✓	✓		✓

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

EXP-13	Define the level of abstraction of explanations according to task, situation, trust, expertise of user.	✓	✓	✓	✓		
EXP-14	Customisation of explanation level of abstraction (if XAI adaptability/adaptiveness is available)			✓			
EXP-15	Define explanations timing according to situation, end user needs, operational impact	✓		✓	✓		✓
EXP-16	Enable explanation and details upon user request	✓	✓	✓	✓		✓
EXP-17	Ensure validity of explanation	✓	✓		✓		✓
EXP-18	Provide instructions/training to handle indications of input/output monitoring	✓	✓			✓	
EXP-19	Provide timely information on unsafe operating conditions		✓		✓	✓	✓
HF-01	IA situational awareness building	✓	✓		✓	✓	✓
HF-02	User situational awareness reinforcement	✓	✓	✓	✓	✓	✓
HF-03	Shared situational awareness building	✓	✓	✓	✓	✓	✓
HF-04	Ability to submit decisions for cross-check validation		✓	✓		✓	
HF-05	Identify suboptimal strategy (normal operation) to propose/justify optimised solution				✓		
HF-06	Identify abnormal operation, share diagnosis, resolution strategy, anticipated consequences	✓				✓	✓
HF-07	Detect poor decision-making by the end user in a time-critical situation	✓			✓	✓	
HF-08	Propose alternatives and support own positions	✓		✓	✓		
HF-09	Modify and accept modification of task allocation / task adjustments (instantaneous/short-term)			✓	✓		✓
HF-10	Provide indication of acknowledged user's intentions			✓			✓
HF-11	Notify possible misinterpretation from the end user			✓			

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

HF-12	Detect misinterpretation from the end user, based on his/her responses or actions			✓			
HF-13	Resolve misunderstanding/misinterpretation			✓			
HF-14	Ability to not interfere in other communications or actions			✓			
HF-15	Ability to provide information on AI-based system capabilities and limitations.						
HF-16	Syntax for spoken procedural language designed to ease end user learning	✓					
HF-17	Syntax for gesture language designed to be intuitive						
HF-18	Ability to disregard non-intentional gestures						
HF-19	Ability to recognise end-user intention when using gestures						
HF-20	Ability to acknowledge the end-user intention when using gestures						
HF-21	Enable spoken natural language deactivation to benefit other modalities						
HF-22	Ability to assess the performance of the dialogue						
HF-23	Contextually transition between spoken natural language and spoken procedural language						
HF-24	Combine or adapt the interaction modalities depending on task and operations			✓			
HF-25	Automatically adapt the modality of interactions to end-user states, preferences and situations			✓			
HF-26	Minimise the likelihood of design-related errors made by the end user	✓				✓	✓
HF-27	Minimise the likelihood of design-related errors related to HAIRM	✓				✓	
HF-28	Demonstrate tolerance to end user errors	✓		✓	✓	✓	✓
HF-29	Provide opportunities to detect end user errors				✓	✓	✓

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

HF-30	Provide efficient means to inform the end user about detected errors						
HF-31	Ability to diagnose failures and present the pertinent information to the end user				✓		
HF-32	Ability to propose a solution to the failure	✓					
HF-33	Ability to support solution implementation						
HF-34	Inform that logs of system failures are kept for subsequent analysis	✓			✓		

1.2. TRL Progression

Table 3 provides an overview of each UC TRL in VAL2. More detailed TRL progression tables are included in the individual UC sections.

Table 3. Use cases TRL overview

UC	Brief Description of prototype component	VAL2 TRL
#1	Startle effect detection, Situation awareness augmentation, and Stress regulation support	4
#2	OlivIA assistant integrated in FlytX	5
#3	DUC backend (AI components)	2
#3	Traffic situation display (UTM City), Storytelling explainer system, DUC HAT HMI.	4
#4	Sequence Optimisation Algorithm, Service for computing ETA and initial trajectory points	6
#4	Explainability for sequence changes, HMI - Electronic flight strips and strip board management	6
#5	Airport Safety Watch	9
#6	COVAID tool of Android application with AI model and hardware prototypes for person queues and indoor air quality sensor board and system.	5

1.3. Video Demonstrators

Links to the second demonstrators (DEM2) for each use case are presented in Table 4.

Table 4. UCs demonstrators

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

UC	Link to video
#1	https://drive.google.com/file/d/17dAC8T68fbwjrtYY0Myy-YThhIRkCHWt/view?usp=sharing
#2	https://drive.google.com/file/d/1zoyRE6GyORnRS2HCfJ4jEquZmG7_Akjw/view?usp=sharing
#3	https://drive.google.com/file/d/1aKoQqLc5q8olT8NzJingJJaDB6j9Ldt8/view?usp=sharing
#4	https://drive.google.com/file/d/13eef5HsMPFQG2szgFlBPQnKzwUH1aAgE/view?usp=sharing
#5	https://drive.google.com/file/d/1IXCK66arp04QPkdt8sc2aW5ue5JFrdmQ/view?usp=drive_link
#6	https://drive.google.com/file/d/12jCKa_6BnOawFKuwJOI4ccq_PRaOmLHq/view?usp=sharing

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

2. Use Case #1 – Flight Deck Startle Response

2.1. Deviation from Validation Plan (D6.3)

VAL2 was performed according to the validation plan described in D6.3 with no deviations.

2.2. Validation Objectives

Due to difficulties in recruiting a sufficient number of commercial pilots for the required data volumes (statistical/machine learning analyses), VAL2 included three activities. The first two involve testing and data collection with general population participants, reserving pilot recruitment for final validation of the full FOCUS assistant.

VAL2 focused on three key objectives. First, it aimed to validate the efficiency of haptic feedback as a tool to regulate pilot stress, measuring its impact on physiological and cognitive indicators in controlled scenarios. Second, it sought to gather additional data to refine and train a more robust detection model, enhancing its accuracy in identifying startle and surprise. Finally, it tested the new design of the assistant in a high-fidelity flight simulator, collecting qualitative and quantitative feedback from pilots to evaluate usability, effectiveness, and overall user experience.

The following table shows the key R&D objectives defined in D6.3.

Table 5. UC1 key R&D objectives from D6.3.

OBJ-ID	Validation Objective
UC1-OBJ-01	To assess the operational relevance of the solution from the CAT (Commercial Air Transport) pilots' perspective in SPO (Single pilot Operations).
UC1-OBJ-02	To assess the acceptability of the solution from the CAT pilot's perspective in SPO.
UC1-OBJ-03	To assess the feasibility and integration of the solution in a relevant operational environment.
UC1-OBJ-04	To assess the effectiveness and efficiency of the assistant support in a relevant operational environment.
UC1-OBJ-05	To assess the generalisation of the solution to multiple different scenarios.

2.3. VAL2 Activities and Methods

VAL2 activities aimed to evaluate the IA in all its parts. The following Table 6 shows the 3 validation activities conducted at ENAC and DFKI within VAL2. For each of these activities we provide all the

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

experimental details in dedicated sections. At the end, all the general results for the UC1 are discussed in a dedicated paragraph.

Table 6. VAL2 activities in UC1

Validation activity	Objective	Participants	Where	When
1: Startle and surprise collection and detection	Gather more data to train an efficient detection model	Fundamental experiment 45 +15 participants @DFKI	ENAC, France & DFKI, Germany	Q3-Q4 2024
2: Stress regulation function validation	Validate the efficiency of the haptic feedback to regulate stress	27 participants	Laboratory, ENAC, France	Q3 2024
3: VAL2 design validation	To test the new design of the assistant on simulator and gather pilot's feedback	12 professional pilots	A320 research simulator, ENAC, France	Q4 2024

In UC1, the main goal was to design and evaluate an IA capable of supporting pilots during the startle and surprise effect. This was done by regulating stress and maintaining or raising the pilots' situation awareness level.

The design of the IA startle recovery features was guided by the performance targets defined in D6.3, which are reproduced in Table 7 below.

Table 7. UC1 performance targets

KPA	Category	KPI
Pilot startle and/or surprise physiological recovery (partial incapacitation)	MoE/MoP	Recovery rapidity (reaction time), recovery rate (reaction accuracy). Pilot acceptance, pilot performance on the operational task.
Pilot situation awareness sustainability/recovery	MoE/MoP	Subjective situation awareness assessment, pilot performance on the operational task. Rapidity to come back to a "normal" scan path. "Normal" scan path recovery rate.

The European Union Aviation Safety Agency (EASA) provides guidance on Human-AI Teaming (HAT) with six categories describing Human-AI partnerships. FOCUS aims at providing support for the following categories:

Level 2B: Human-AI collaboration (supervised automatic decision and action implementation).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

2.4. Startle and surprise function (Activity 1): Fundamental experiment

2.4.1. Activity 1 Objectives

The primary objective of this study was to investigate the independent and combined effects of startle and surprise on feedback, behavioural responses, and physiological activity during a simulated flight task. Specifically, the study aimed to:

- Characterize the distinct impacts of startle and surprise on task performance, gaze behaviour, and physiological measures (e.g., heart rate, skin conductance).
- Explore the combined effects of startle and surprise to determine whether their interaction amplifies or alters the individual effects.
- Lay the groundwork for automatic detection of startle and surprise in aviation contexts, with the goal of improving safety through targeted countermeasures.

The study tested the following hypotheses:

- Surprise will increase reaction times due to the cognitive cost of reframing but will not significantly affect task accuracy. It will also lead to an increase in skin conductance, reflecting an emotional response.
- Startle will have a more pronounced impact than surprise, affecting task performance (reaction times and accuracy), gaze behaviour (narrowing of attention), and physiological measures (heart rate and skin conductance).
- The combination of startle and surprise will produce stronger effects than either stimulus alone, leading to greater self-reported feelings of startle and surprise, more significant performance declines, and more pronounced physiological responses.

2.4.2. Methods

Participants: Forty-five participants (11 women) aged from 21 to 45 years old (M = 28.1, SD = 6.5) were split into three groups: Startle, Surprise, and a Combination group. All groups used the OpenMATB for experiment execution (Figure 2). The participants were misled by a cover story. Before the experiment calibration and tutorials were done. During the 7-minute experiment, the startle effect was triggered by a sudden white noise, the surprise effect was a reversed colour scheme (Figure 2), and the combination combined both.

Startle Condition: Participants in the Startle condition were exposed to three precursor white noise sounds at 80dB, so they could familiarize themselves with the audio. A singular event of 100dB of white noise was then triggered during the experiment.

Surprise Condition: Participants in the Surprise condition experienced a reverse coloration scheme event on the screen. The reverse colour scheme was presented and was sustained for the rest of the scenario to elicit surprise.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Combination Condition: Participants in the Combination condition were exposed to both the singular sudden loud noise, as well as the reverse video effect.

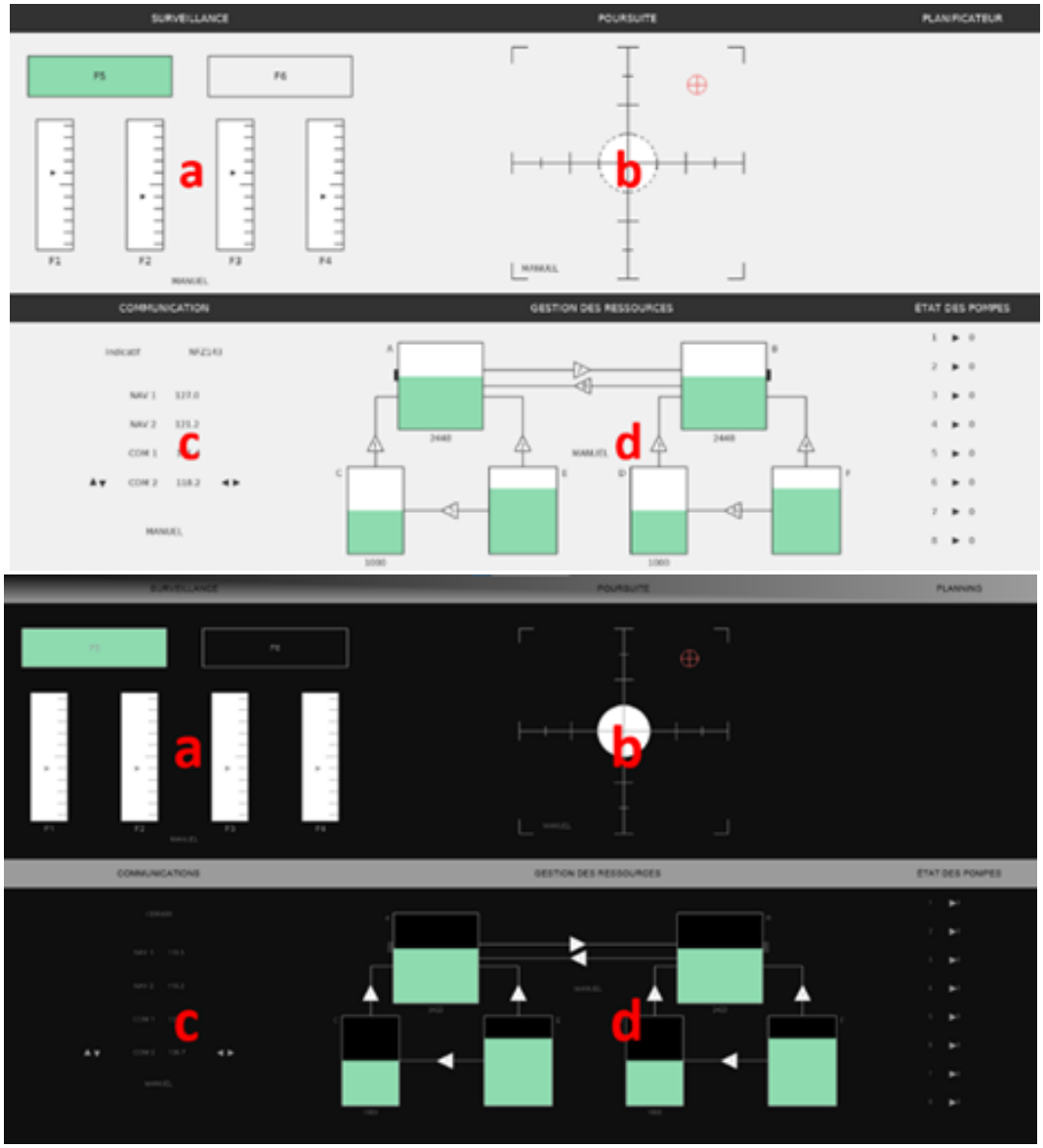


Figure 2. OpenMATB interface, left: normal mode, right: reverse video mode (surprising stimulus)

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

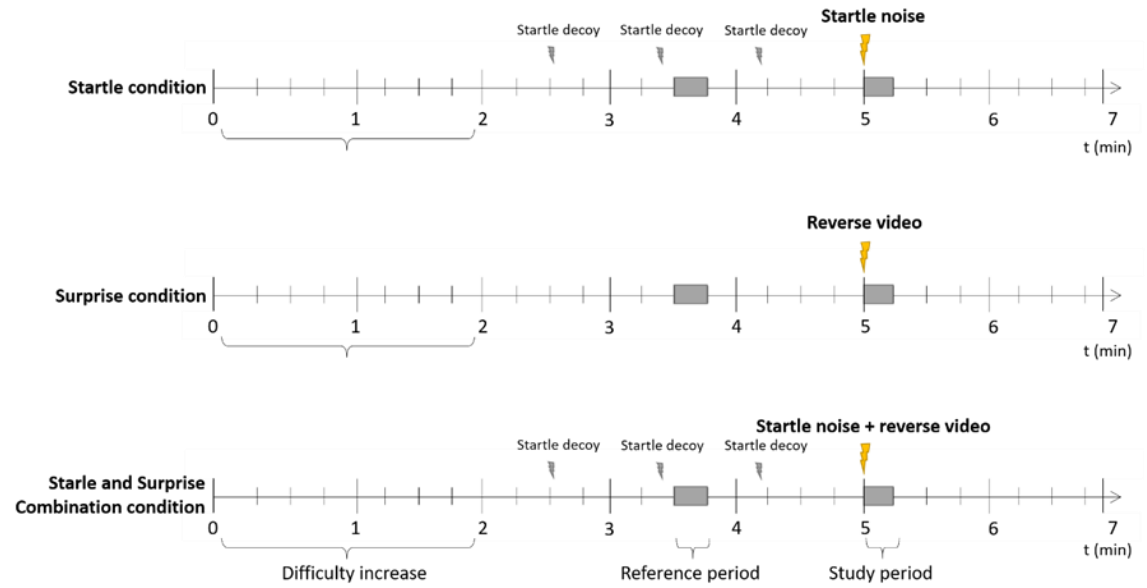


Figure 3. Timeline of experiments

2.4.3. Measured

Participants were assessed on self-reported startle and surprise levels using a 10-point Likert scale, NASA TLX workload scores, OpenMatb subtask performance, and facial expressions. Gaze behaviour metrics, heart rate and skin conductance response were continuously monitored and analysed.

OpenMATB sub-task performances:

- System monitoring: We analysed reaction times and success in detecting and correcting abnormal behaviour on the six visual indicators.
- Tracking: We analysed the average centre deviation of the cursor over 3-second epochs and compared it to the reference period.
- Communication: We evaluated the success rate of entering the correct frequency in the correct channel.

Signals were captured using a three-lead Electrocardiogram (ECG) to measure heart rate, as well as skin conductance, eye movements were captured by a 600Hz Tobii Pro Spectrum, and a Logitech camera filmed and captured the facial expressions of the users.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Figure 4. Experimental set-up with the reverse video mode (surprising stimulus) of the OpenMATB

2.4.4. Data Analysis & Results

For data analysis the performances were compared using a t-test with a Bonferroni correction. ANOVAs with a post-hoc test were used to compare the physiological data, with a t-test for pairwise comparisons. Chi-square tests were used to compare the performance of the sub-tasks of the OpenMATB.

Subjective Measures:

Participants correctly identified that the combination condition yielded more startle and surprise (Figure 5). There was no statistical difference in anxiety nor stress across the three conditions.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

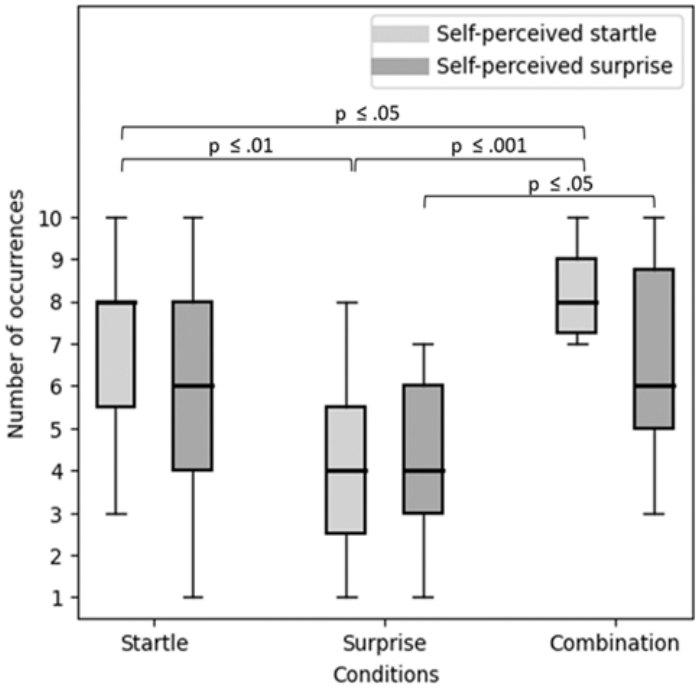


Figure 5. Tukey boxplots of the self-assessment of startle and surprise in all conditions

OpenMATB Task Performance:

System monitoring yielded similar results across the experimental conditions. The tracking task showed no significant statistical differences. Communication accuracy was worse for the startle and combination conditions. The combination between startle and surprise showed no difference (Figure 6).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

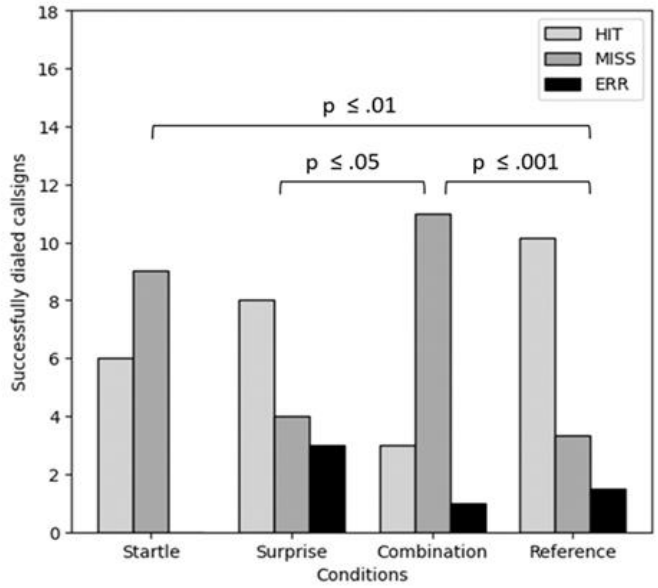


Figure 6. Communication accuracy in all conditions

Gaze Behaviour Analysis:

Stationary entropy was found to be lower in the startle and combination conditions (Figure 7). The analysis of the K-coefficient, the Lempel-Ziv complexity, and the explore/exploit ratio showed no significant effect of the stimuli of interest compared to the reference period ($p > .05$ in all conditions). The distribution of the AOI showed more focus on System Monitoring and less on Communication in the combination condition.

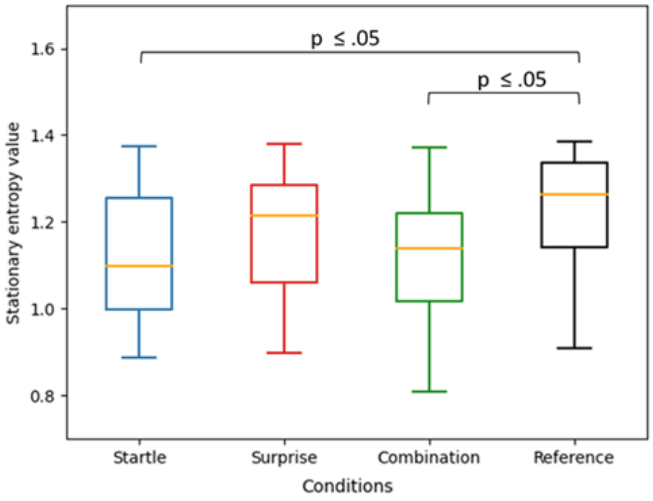


Figure 7. Tukey boxplots of the stationary entropy during the study and the reference period

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Physiological Responses:

Heart rate was significantly higher in the startle and combination condition than in the reference (Figure 8). Skin conductance showed a post-stimulus change in all conditions. Compared to the surprise condition, both the combination and startle conditions showed a significantly higher magnitude (Figure 9).

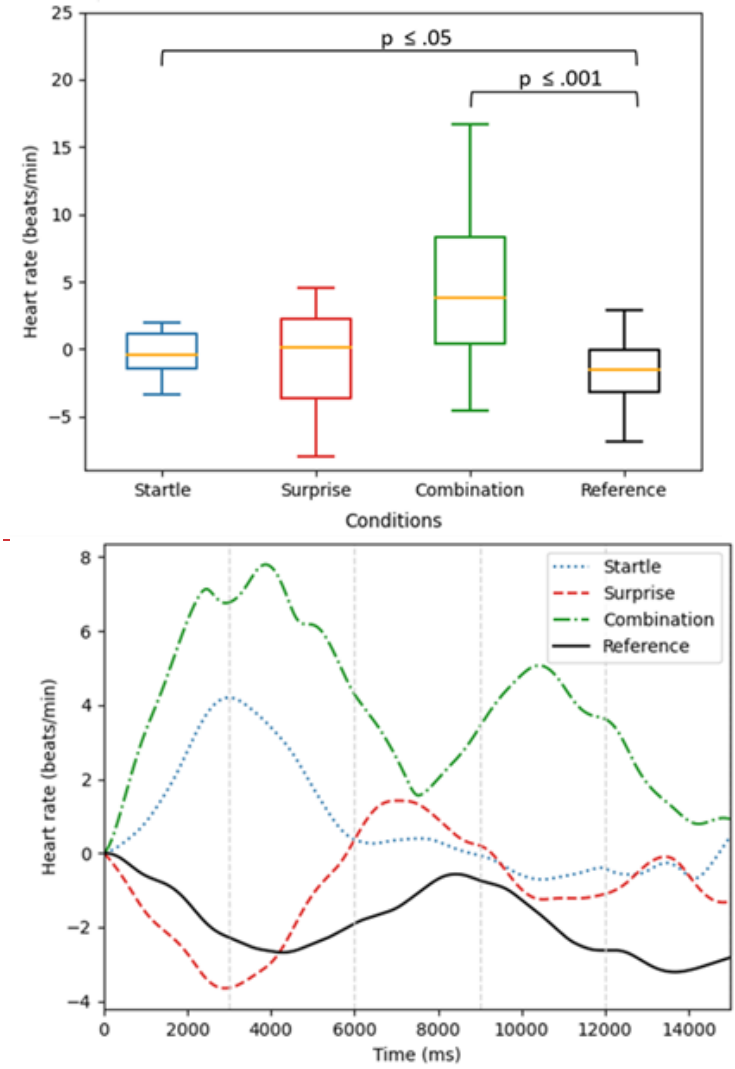


Figure 8. Tukey boxplots of the mean heart rate change and mean heart rate variation on the 15s study period

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

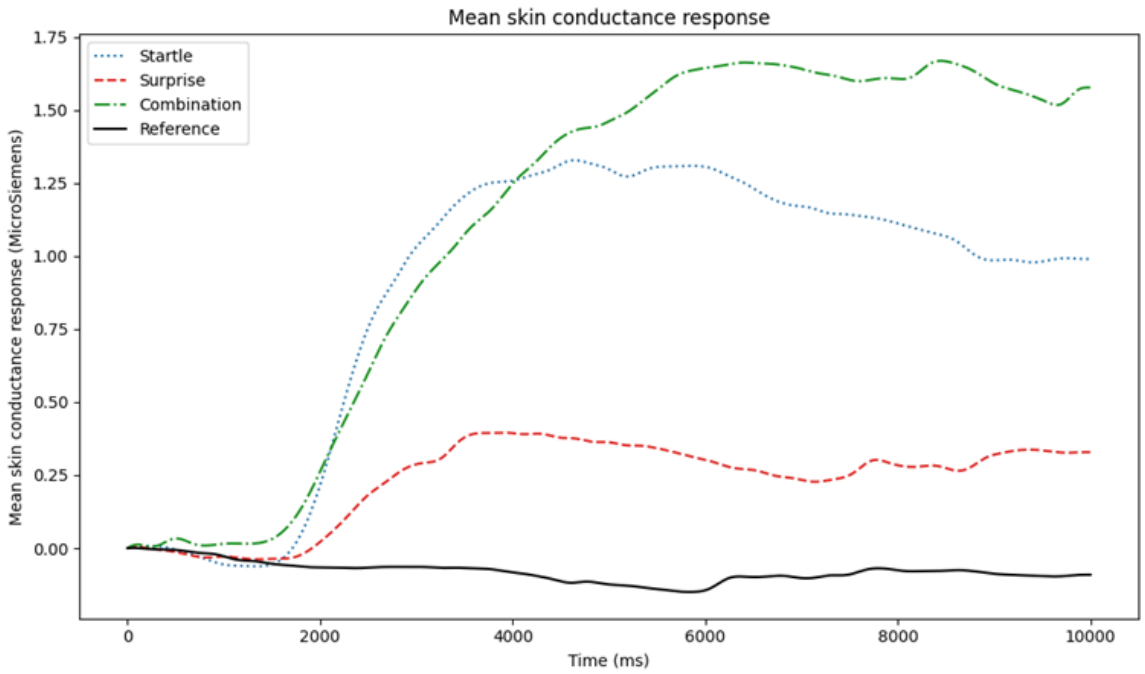


Figure 9. Skin conductance response to the three stimuli of interest in comparison to the reference)

2.4.5. Startle and surprise function (Activity 1): Training and detection

First data explorations suggested that the existing dataset might be too small for analysis of startle and surprise detection. Thus, we replicated the MATB study to test more participants. The same method was used for the study and the study was approved by the Ethical Review Board of the faculty of Mathematics and Computer Science at Saarland University. 15 healthy volunteers (10 males, 5 females) aged between 22 and 36 years (M=26.4, S=4.14) were recruited and asked to perform the MATB task as described above.

Physiological sensors (Bitalino device) included a three-lead electrocardiogram (ECG) for heart rate, a galvanic skin response (GSR) sensor for skin conductance, and a photoplethysmogram (PPG) earpiece for blood flow were used for data acquisition. All signals were synchronized using LabStreamingLayer software. For training and detection all data was used from both studies. Overall, from both experimental designs, we got 60 participants. However, 26 participants must be excluded as half of them have Startle and Surprise being triggered simultaneously and remaining half participants had missing and/or incomplete physiological data.

Signal Preprocessing

The bio-signal data analysis involved an offline analysis of the recorded data including a structured preprocessing pipeline to ensure noise-free and reliable feature extraction. ECG signals were

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

processed using a high-pass filter (0.6 Hz), a low-pass filter (100 Hz), and a notch filter (50 Hz) to remove baseline drift and power line interference. PPG signals were band-pass (0.5–5 Hz) and high-pass (0.5 Hz) filtered to retain relevant frequency components. RESP and EDA signals were smoothed using a low-pass filter with a 5 Hz cutoff frequency. Feature extraction included statistical features such as mean, standard deviation, minimum, and maximum values, alongside peak count, which was determined by the number of signal peaks exceeding the mean amplitude. For classification, we implemented Support Vector Machine (SVM), Naive Bayes (NB), and XGBoost models. We also used an early fusion approach to integrate features from multiple bio signals. We followed a standard procedure for model evaluation with hyperparameter tuning using 5-fold cross-validation to ensure robust performance. The dataset was split into training and testing subsets using an 80:20 ratio. Additionally, different window durations (3s, 5s, 7s, and 10s) were explored to assess their influence on classification accuracy. Startle and Surprise events were cut from the data according to the annotation. For the Baseline, we selected a window in the signal prior to the onset of these events to ensure that Startle and Surprise effects do not influence the Baseline.

Results and Evaluation

Figure 10, Figure 11, Figure 12, and Figure 13, shows the results of our analysis *for Startle vs Baseline, Surprise vs Baseline, Startle vs Surprise*, and *Startle vs Surprise vs Baseline* in the best performing time window for each condition. In the *Startle vs. Baseline*, *Surprise vs Baseline*, and *Surprise vs Startle*, SVM with late fusion achieved the highest accuracy of 88.80%, 89.62%, and 85.71% respectively, indicating the robustness of our fusion techniques and significant improvement over unimodal approaches. For *Surprise vs Baseline vs Startle*, XGBoost with late fusion reached highest accuracy of 74.96%, very much in line with our findings on binary classification experiments. Overall, late fusion showed good results highlighting the importance of fusing the modalities systematically. For the *Startle vs Baseline* and *Startle vs Surprise vs Baseline* conditions, 3s emerged as the best time window, the *Surprise vs Baseline* and *Startle vs Surprise* conditions achieved best results with the 7s and 10s window. EDA and PPG achieved better results compared to other modalities. Overall, comparing results with Baseline conditions reveal that our methods are reliable to predict Startle and Surprise events. Interestingly, when performing a three-class classification, our methods were able to achieve good results with significant improvement over chance level.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

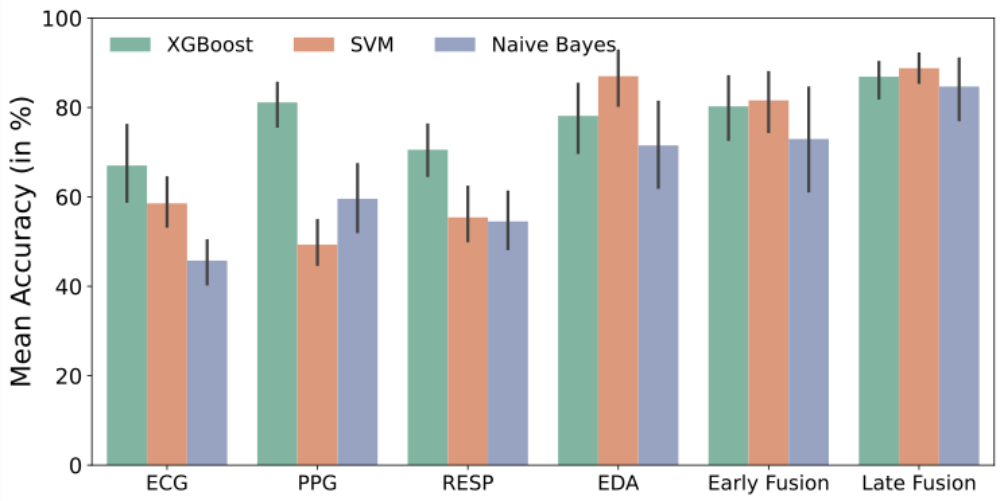


Figure 10. Mean Accuracy for **Startle vs Baseline** with best performing 3s window. Error bars indicate 95% confidence interval.

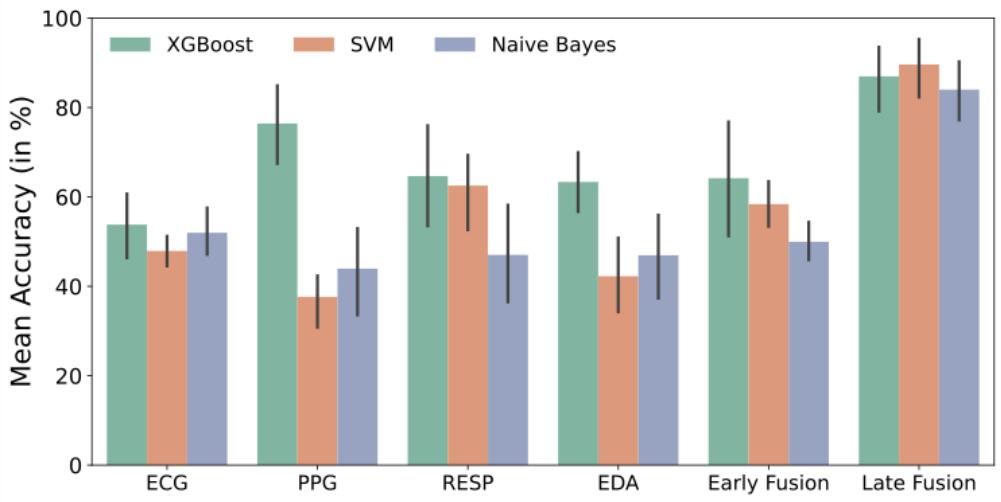


Figure 11. Mean Accuracy for **Surprise vs Baseline** with best performing 7s window. Error bars indicate 95% confidence interval.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

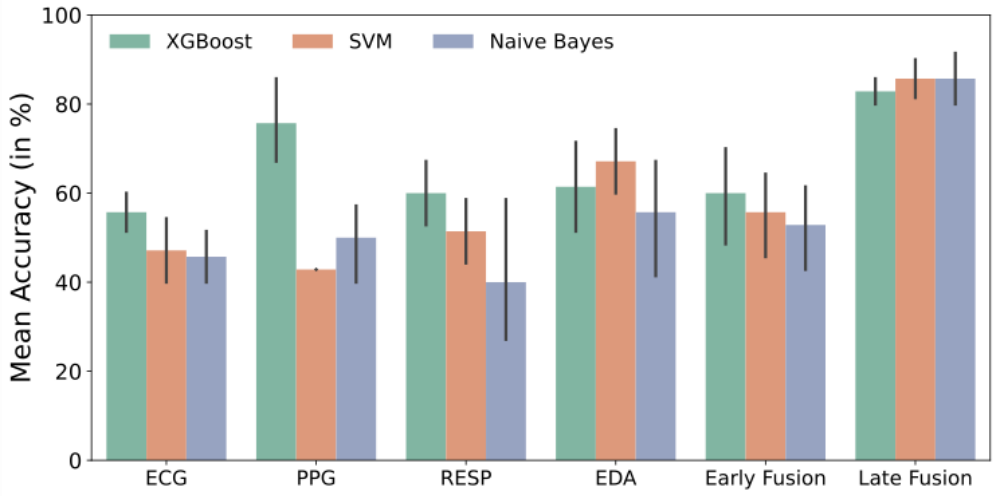


Figure 12. Mean Accuracy for **Startle vs Surprise** with best performing 10s window. Error bars indicate 95% confidence interval.

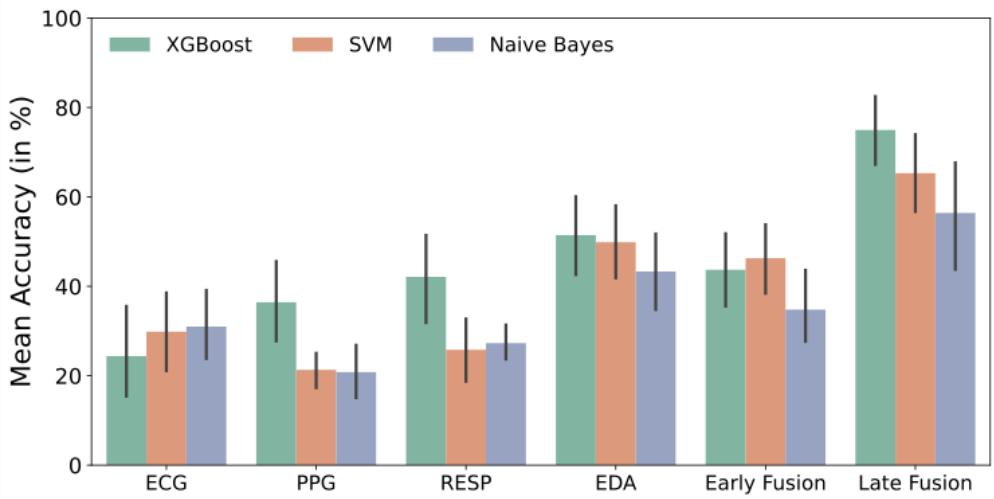


Figure 13. Mean Accuracy for **Startle vs Surprise vs Baseline** with best performing 3s window. Error bars indicate 95% confidence interval.

We also compared different window sizes using the best-performing classifier for each of the four experimental evaluations, i.e., *Startle vs Baseline*, *Surprise vs Baseline*, *Startle vs Surprise*, and *Startle vs Surprise vs Baseline*, as shown in Figure 14, Figure 15, Figure 16 and Figure 17 respectively. Overall, late fusion consistently outperformed individual modalities across all window sizes. In the *Startle vs Baseline* condition (Figure 14), window size had minimal impact, with all durations yielding comparable results. A similar trend was observed in *Startle vs Surprise* (Figure 16), where performance remained stable across window lengths. In contrast, for *Surprise vs Baseline* (Figure 15), longer windows (5s and 7s) consistently outperformed shorter ones. Interestingly, in the *Startle vs Surprise*

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

vs Baseline condition (Figure 17), the shortest window (3s) led to the best performance, significantly exceeding that of longer windows.

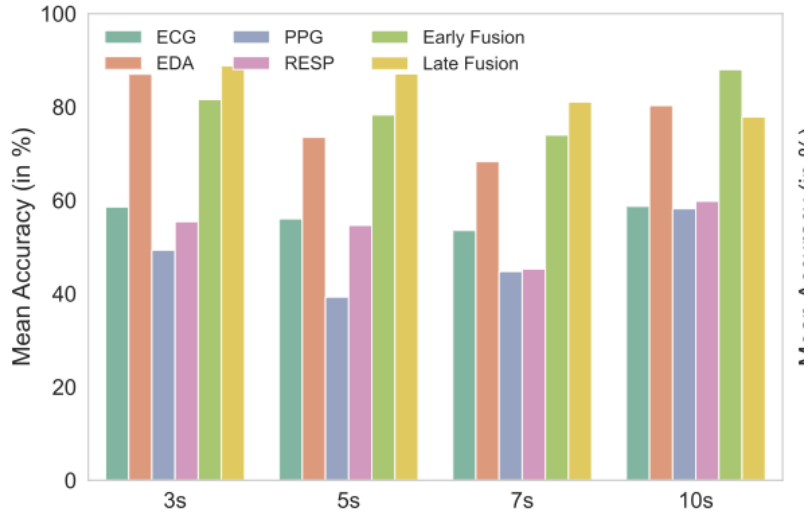


Figure 14. Mean accuracy across all time windows, **Startle vs Baseline** using the best performing SVM model.

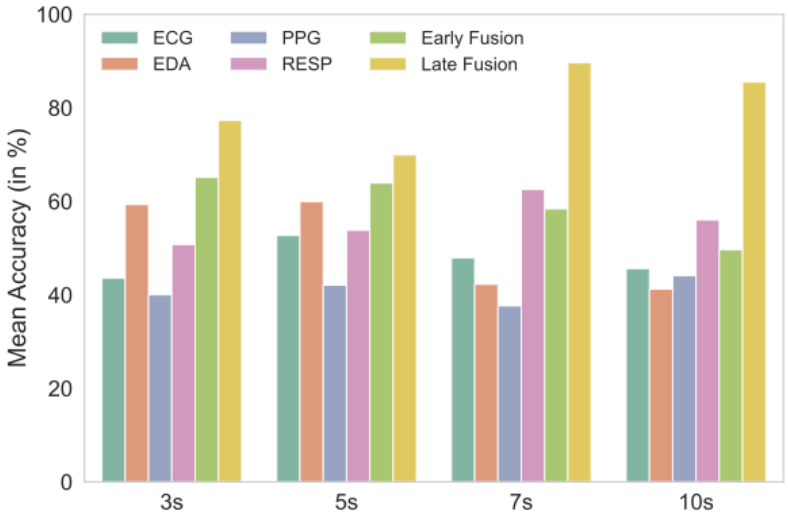


Figure 15. Mean accuracy across all time windows. **Surprise vs Baseline** using the best-performing SVM model.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

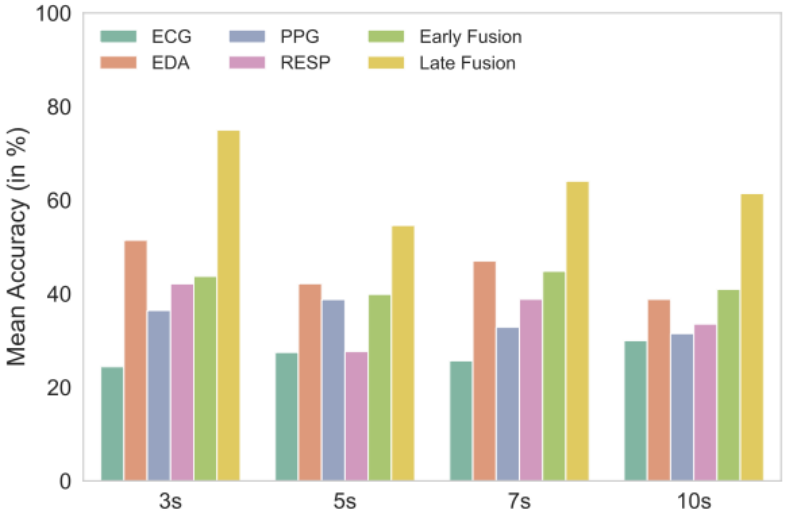


Figure 16. Mean accuracy across all time windows**. Startle vs Surprise** using the best-performing SVM model.

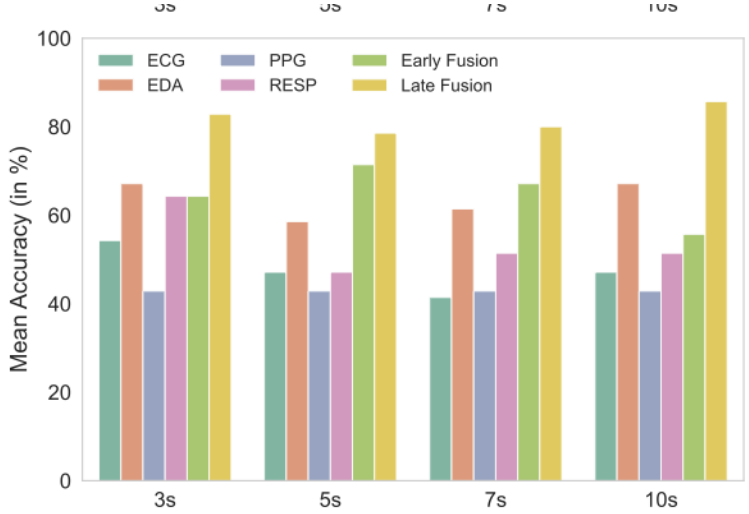


Figure 17. Mean accuracy across all time windows. **Startle vs Surprise vs Baseline** using the best-performing XGBoost model.

2.4.6. Discussion

The results partially support the proposed hypotheses. The surprise stimulus did lead to a change in the skin conductance for the participant, validating H1. As hypothesized, the combination of startle and surprise resulted in a greater impact on cognitive abilities, supporting H3. The startle and combination conditions performed significantly worse on certain metrics (heart rate, skin conductance, communications) compared to the surprise conditions, which suggests startle as the

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

driving force. The combination of startle and surprise resulted in a more substantial negative impact on the results. The increased heart rate, self-perceived startle, and the gaze distribution being more affected means surprise likely intensifies the startle effect. We investigated the classification of startle and surprise states using physiological signals, aiming to enhance real-time detection and intervention in high-risk domains such as aviation. Our findings demonstrate that EDA and PPG emerged as a reliable feature for distinguishing these states. Shorter windows (3s) worked reliably for classifying these states simultaneously in a three-class experiment while for classifying Startle and Surprise events or with Baseline condition, window size had minimal impact. These results highlight the feasibility of distinguishing startle and surprise states from physiological data, supporting the development of intelligent monitoring systems for aviation and critical environments.

2.4.7. Key Findings

- Surprise increased skin conductance but did not significantly affect task performance or reaction times.
- Startle led to a decline in communication task performance, increased heart rate and skin conductance, and narrowed visual attention.
- Combination of startle and surprise resulted in the most significant effects, including higher self-reported startle and surprise, greater performance declines, and prolonged physiological responses (e.g., heart rate elevation).
- Late fusion achieved significantly better results than uni-modal approaches. Our methods could reliably detect Startle vs Suprise vs Baseline states in three-class classification with highest 74.96% mean accuracy.

The findings suggest that startle and surprise, particularly when combined, can significantly impair cognitive and physiological functioning during flight tasks.

For more information about Activity 1 see:

Duchevet, A., Imbert, J.-P., Garcia, J., Lamirault, B., & Causse, M. (2025). Investigating the Independent and Combined Effects of Startle and Surprise in a Simulated Flight Task. *Human Factors*, *0*(0). https://doi.org/10.1177/00187208251342100

Sharma, M., Duchevet, A., Daiber, F., Imbert, J. P., Rekrut, M. (2025) Distinguishing Startle from Surprise Events Based on Physiological Signals. (Submitted)

2.5. Stress regulation function (Activity 2): Fundamental experiment

During FOCUS's VAL1 experiment both visual and vibrotactile support were used within the stress regulation support module. During the debriefing all the pilots stated that they did not perceive the vibrating wristband. Existing research has explored various methods to modulate the startle response, including auditory and visual stimuli. The study attempted to take previous research a step further

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

through the implementation of vibrotactile feedback and cardiac interoception, a method that still lacks research and exploration. This study is the first controlled research on startle responses using cardiac interoception techniques.

2.5.1. Activity 2 Objectives

The main objective of this research is to explore the implementation and effectiveness of vibrotactile stimulus for startle mitigation. To do this, this study plans to evaluate several existing cardiac interoception stimuli.

To this end, the following hypotheses were developed:

- Vibrotactile interventions that reproduce low heart rate will regulate the response to startle
- Startle response regulation performance will vary depending on the vibrotactile technique

2.5.2. Related Work

Research about stress modulation through visual or auditory feedback is abundant, contrary to research that has been conducted with tactile stimuli. Nevertheless, some devices such as AmbienBeat, EmotionCheck, and Doppel have shown great promise in stress reduction and regulation.

The AmbienBeat uses tactile feedback to regulate heart rate. The device renders a rhythmic tactile stimulus made of a soft membrane. During its evaluation there was an increase in participant heart rates with 120BPM, and a decrease with 60BPM, confirming its utility in heart regulation.

EmotionCheck provides slow heart feedback at 60 BPM to influence the perception of anxiety of its users. Users that were given the EmotionCheck were shown to be significantly calmer during the experimentation process, showcasing its benefits.

Finally, the Doppel delivers an on-demand heartbeat-like vibration. By measuring the rest rate of participants and then applying this technique, stress and anxiolytic effects were shown.

Because of this history of devices that are capable of influencing stress using cardiac interoception stimulus, it was expected this experiment would have a strong positive effect for startle response.

2.5.3. Methods

The study used a within-subject design with four conditions: constant heart rate at 60 BPM (60BPM), 80% of the participant's heart rate at rest (80HRR), 80% of heart rate during the task (80HRT), and no intervention (NI). These were counterbalanced using a Latin square design. All data was collected using standardized methods, and the data was reviewed by members of the committee.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Participants:

27 healthy volunteers (18 males, 9 females) aged between 20 and 61 years (M=33.25, SD=10.94) were recruited from local institutions. All participants were healthy with a resting heart rate between 52.52 and 86.41 BPM and none had dyscalculia nor cardiovascular disease.

Apparatus:

The experiment was designed using PsychoPy. Our MIST adaptation implementation was based on MistyPy. Vibrotactile feedback was provided through LRA actuators driven by a DFRobot DFR0720 amplified audio receiver and worn on the wrist. A Bitalino was used to measure the physiological responses.

Procedure:

The study began by inviting participants to read an information sheet explaining the research on cognitive performance enhancement through vibrotactile feedback. They were informed that their data would only be included if they performed above the average score. After providing consent and completing a demographic questionnaire, participants were fitted with physiological sensors (ECG, PPG, EDA) and a vibrating wristband. A 3-minute resting period followed to establish baseline heart rate and variability for later comparison with stress-induced responses. Next, participants practiced the task to familiarize themselves with the controls before completing four experimental runs—one per condition—while wearing noise-isolating headphones that also delivered startle stimuli. Between runs, they rated their stress and pleasantness levels and were given optional breaks. The collected data allowed analysis of heart rate variability differences between rest and stress conditions.

The experiment was implemented using the Montreal Imaging Stress Task (MIST), and a startle response stimulus was applied using a white noise generated and delivered by a high-quality audio system, this ensured a good startle response. Participants completed 50 trials per condition, each involving a mental arithmetic task with numbers (0–99) and operations (addition, subtraction, multiplication, division), similar to the MIST protocol. Task difficulty increased with performance, adjusting complexity (up to four two-digit numbers and four operands) and reducing time limits to as low as 4 seconds. Responses were entered via keypress (0–9), with failures recorded for timeouts or incorrect answers. After each trial, feedback (success/fail/timeout) and comparative performance against group averages were displayed. Before each condition, participants were reminded that their data required near-average performance for inclusion. On the 25th trial, a loud white noise stimulus was delivered via headphones, followed by vibrotactile intervention from trial 26 onward. Post-experiment, participants completed questionnaires on workload, stress, and vibrotactile feedback, along with open-ended reflections. A debrief document clarified the study's true objectives and procedures, ensuring participant reassurance. The entire session lasted under 60 minutes.

2.5.4. Measures

The physiological measures (HR/EDA) we collected were computed using Neurokit2 to calculate:

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

RMSSD: (Root Mean Square of Successive Differences) Measures high-frequency heart rate variations linked to parasympathetic activity.

pNN50: (Proportion of Normal-to-Normal Intervals > 50ms) Measures the proportion of adjacent intervals with >50ms difference, indicating parasympathetic influence.

HRVHF: (High-Frequency Band) Frequency of HRV for parasympathetic activity.

We collected subjective data at the end of each condition to assess perceived stress and pleasantness after completing the arithmetic tasks. We also collected subjective data about the overall experience at the end of the study.

2.5.5. Data Analysis and results

HRV data was measured using the before and after averages, as well as differences in the results.

Figure 18. Experimental setup and Montreal Imaging Stress Task interface: bitalino PPG (1), ECG (2), EDA (3) sensors, vibrotactile bracelet (4).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

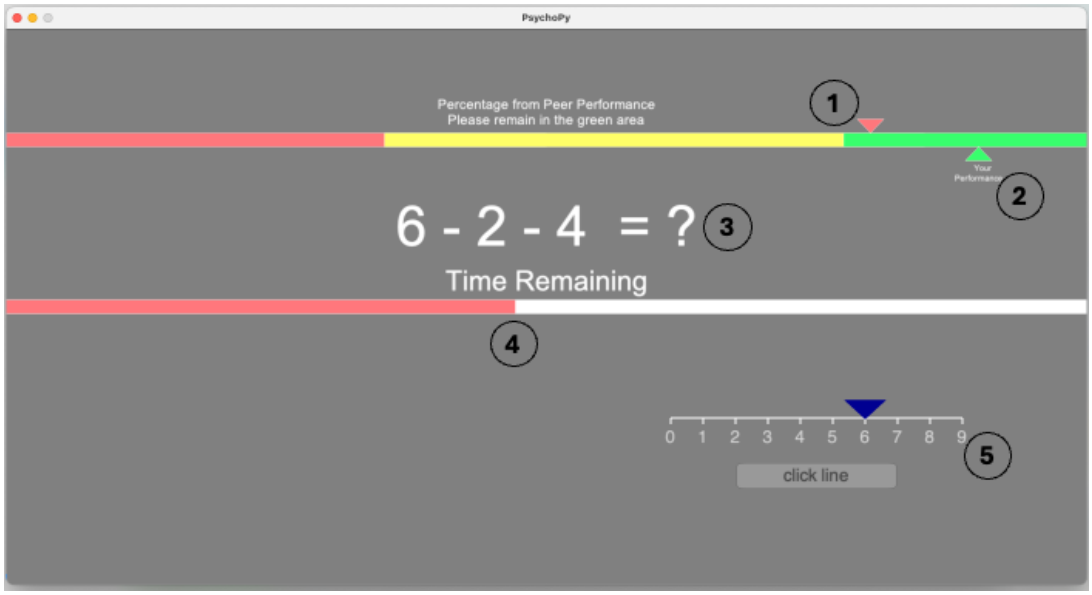


Figure 19. MIST interface in PsychoPy: Average peer performance (1), participant's current performance (2), arithmetic calculus to solve (3), time remaining to solve calculus (4), and answer input (5).

The results of the study can be divided into physiological and subjective results.

Physiological Startle Response

To assess vibrotactile intervention effects on the startle response, we analysed changes (ΔHRV) in Heart Rate Variability (HRV) variables between a resting baseline and during the arithmetic tasks. Lower HRV variables signify more stress. The interventions were computed with help from key metrics modulated by the parasympathetic nervous system, in order to create a comprehensive insight and analysis regarding the influence of the stimulus.

We expected negative ΔHRV values across conditions, with smaller values indicating higher stress levels and more effective startle mitigation.

Heart rate (HR) and electrodermal activity (EDA) were also measured to confirm physiological response to the startle. The Neurokit 2 toolkit was used for physiological measures computations. The smallest magnitude was seen on the 60BPM. The negative values for ΔRMSSD had the largest magnitude in conditions 80HRR, 80HRT, and NI. Repeated measures ANOVA found a statistically significant difference in ΔRMSSD between the interventions. Post-hoc tests with Holm correction indicated that the RMSSD in the 60BPM condition was significantly lower than in 80HRT condition ($p < .05$)). No other significant result was found. The average heart rate was the lowest in the 60BPM condition, a repeated measures ANOVA did not find any significant difference.

Subjective Feedback

After completion of each condition, participants were asked to rate the perceived stress and pleasantness of the task on a seven-point scale. The results showed that they felt less stressed in the 60BPM and in the 80HRR conditions.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

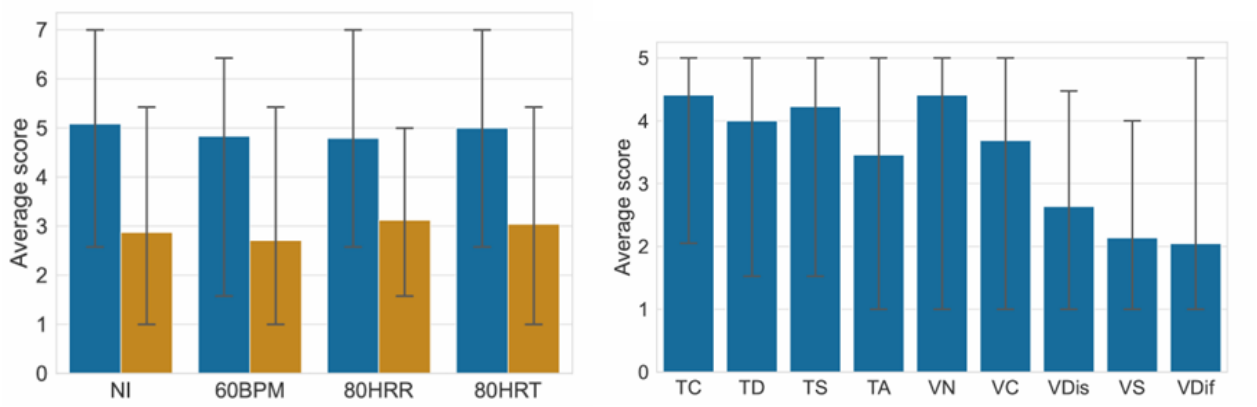


Figure 20. Left: average subjective stress and pleasantness score for each condition. Dark blue represents stress; light orange represents pleasantness. Right: average score on a five-point scale for overall task difficulty (TC), task demand (TD), task stress (TS), task anxiety (TA), vibration perceptibility (VN), vibration comfort (VC), vibration distractibility (VDis), vibration stress (VS), vibration discrimination (VDif). In both graph, error bars denote the 95% confidence interval.

Participants found tasks with vibrotactile stimulation at 20% below their resting heart rate more pleasant, though the 60BPM condition was rated the least pleasant overall. However, Friedman tests showed no significant differences in stress or pleasantness.

Participants described the tasks as difficult, demanding, and stressful. Vibrotactile interventions were noticeable and somewhat comfortable but also distracting and stressful. Interventions were not perceived as significantly different from each other (Figure 20).

Post-experiment interviews indicated successful stress and startle induction. Participants reported the task was hard with time constraints and noted pressure from the noises. Frustration stemmed from incomplete task completion due to the time limit.

2.5.6. Discussion

This study aimed to investigate the effectiveness of vibrotactile interventions in mitigating the startle response, a first of its kind experiment. The results highlight both the promise and complexity of using simulated heartbeats for stress regulation in the context of sudden startling events.

The successful induction of stress and startle was validated by negative ΔHRV values and participant feedback. The "no intervention" condition consistently showed the largest magnitude decrease in heart rate variability which strongly implies the base protocols can be improved. All test participants found the system to be accurate and valid, which indicates there were good results. The heart rhythm results also helped to indicate a good test case for them.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

While all interventions helped, the benefits of the constant vibrotactile stimulation (60BPM) stood out. Specifically, participants all rated the 60BPM asthe most calming and efficient. This indicates that precise haptic feedback provides superior benefits to users. And that precision provides a more reliable source for the parasympathetic response.

The results did not entirely validate the hypothesis. The constant beat system did test well above the other stimulus types. Also, it still requires a significant amount of work to validate further utility of this response. Still, its benefits to improve stress in both real world and test case scenarios is a strong factor. The constant 60BPM has more evidence than anything else, but all tests showed promising results.

Key Findings

- The experimental protocol successfully induced a measurable startle response (confirmed by both physiological data - negative RMSSD/pNN50 differences - and subjective feedback).
- All interventions helped participants recover from the startle response to some degree. Vibrotactile feedback at a constant rate of 60BPM provided significantly more support in mitigating the effects of the startle response compared to other conditions. Simulating a slower heart rate than the participant's actual heart rate appeared beneficial for modulating the startle response. Thus, VAL2 will include a vibrotactile feedback at 60BPM for the stress regulation support.
- Considering the complexity and individual nature of heart rate and heart rate variability in acute stress response, this study does still have limitations. In particular, no significant differences could be found between conditions. This will be addressed in future studies with between-subjects designs and bigger sample sizes.

For more information about Activity 2 see:

Vo, D. B., Bichon, C., Duchevet, A., Peyrequeou, A., Imbert, J. P., Daiber, F., FOCUS: Investigating vibrotactile feedback interventions to mitigate the startle response in aircraft cockpits. UIST2025. (Submitted)

2.6. VAL2 design validation (Activity 3)

The third validation activity is about validating the system design encompassing all the features and functions. Three support functions are currently integrated to the IA in UC1: a startle and surprise detection module; a stress regulation support feature; and a situation awareness support feature. In addition, an assistant monitor and a control panel is provided on the pilot’s electronic flight bag (EFB). The complete architecture and design rationale of FOCUS are provided in the deliverable D4.5. VAL2 design validation aims at collecting physiological data to validate the startle and detection module and have qualitative results on the VAL2 design of FOCUS. Since the startle detection module is not yet operational, the activation of FOCUS is triggered in a wizard of Oz fashion at the same time as the

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

thunderstorm noise/lightning strike. The data collected during the experiment was used to validate the module.

2.6.1. Activity 3 Objectives

Starting from VAL1, the tested design evolutions of FOCUS lead us to formulate the following hypotheses:

1. The solution is considered relevant by CAT pilots in SPO.
2. The solution is considered acceptable by CAT pilots in SPO.
3. The solution is considered feasible and can be integrated in relevant operational environments.
4. The solution is considered effective and efficient in supporting CAT pilots to overcome the startle and surprise effect in relevant operational environments.
5. The solution is useful in different operational conditions by CAT pilots.
6. An efficient solution is identified.
7. The startle and surprise detection module is able to detect and identify pilot's state in post-analysis of the simulator sessions.
8. The assistant's interface is understood by pilots

2.6.2. Methods

Participants:

Twelve experienced pilots were recruited (11 male, 1 female) for the validation study (mean age =47.25 SD = 8.3, mean number of flight hours=8135, SD = 5611). The gender could not be balanced since female pilots represent 5.8% of the pilot's overall population and we had two female pilots who could not attend the experimentation. All performed the lightning strike scenario.

Apparatus:

Pilots were equipped with the Bitalino physiological sensors (ECG, PPG, Respiration Rate, GSR, EMG, Eye tracking glasses) and were invited to sit in the A320 simulator.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Figure 21. Pilot equipped with Bitalino physiological sensors and eye tracking device

Procedure:

Upon consent form completion, each participant was first introduced to the research objectives. The experimenters then presented the goal and the proceedings of the study, and specifically, the context, the IA features and the two flight scenarios of the study without disclosing the startling and surprising events (lightning strike and cargo shift). In each scenario, the flight was conducted in SPO. Each flying session took place in an Airbus A320 simulator built at ENAC.

Next, the participant was equipped with the physiological sensors and was invited to sit in the A320 simulator. Upon eye tracking calibration, one of the experimenters performed a walkthrough of the IA functions and invited the participant to experience the stress regulation and the situation awareness support. The participant could ask any question about the IA at this stage.

When ready, the participant began the training phase. The training phase started with exploring the cockpit surroundings for at least 20 minutes to give enough time for the participant to familiarize with the simulator, the controls' location and sensitivity, and the simulation view rendering. In addition, the participant was allowed and encouraged to request support and experience the assistant functions during the training. When the participant felt comfortable enough to proceed, a baseline scenario was started which consisted of taking off, performing a short flight and landing back at the airport.

Pilots had to fly a standard approach phase with FOCUS manually activated, to evaluate the usability and the explainability of the assistant. They had to test the scenario three times with the 3 levels of sensitivity available with FOCUS (LOW/MEDIUM/HIGH), at the end of these scenarios they had to select the level that suited them well.

Then, the pilots had to fly the VAL2 scenario that triggers startle and surprise. On final approach, the aircraft experienced a simulated lightning strike. As a result, a loud bang was heard, and an intense flash was triggered, provoking startle and surprise. In addition, this event led to a simulated automation disconnect.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

The scenario began during the holding pattern for Runway 25 at Orly airport. An event triggered automatically by the platform software manager provoked a low fuel warning in the cockpit (Figure 21).

The pilot is supposed to say out loud 'Master caution, Low fuel', remove the auditory warning by pushing the Master caution button on the flight control unit, and ask the air traffic control to integrate the final approach immediately by saying "ENAC2023 request radar vectoring for immediate approach". In addition, the pilot should also report a low-fuel emergency situation by declaring "PAN PAN" or "MAYDAY MAYDAY" on the radio. Following this message, the air traffic controller will give a heading to ENAC2023 in order to integrate the Runway 25 approach. ENAC2023 is then supposed to call back when the Localizer is captured. This low fuel situation, while increasing the urgency of the situation and raising the stress level of pilots, is a necessary "scenario trick" to force participants to land.

At 12 Nm from the runway (t0), the lightning strike and thunder were triggered in the cockpit. Following this event, autopilot (AP), auto-thrust (ATHR) and flight director (FD) were disconnected. Since the startle detection was not integrated to our platform for VAL1, the startle detected event was triggered automatically by the platform software manager, which led the pilot into a startled state.

The startle detected event started the emotion regulation support function after 5 seconds. This function lasted for 30 seconds. 15 seconds after the startle event, the situation awareness function started and lasted for at least 90 seconds, depending on if the pilot's situation awareness was adequate. The pilot could start or stop both functions manually at any time on the assistant user interface. During the first 5 seconds after startle detection, the pilot could stop the assistant before its activation.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

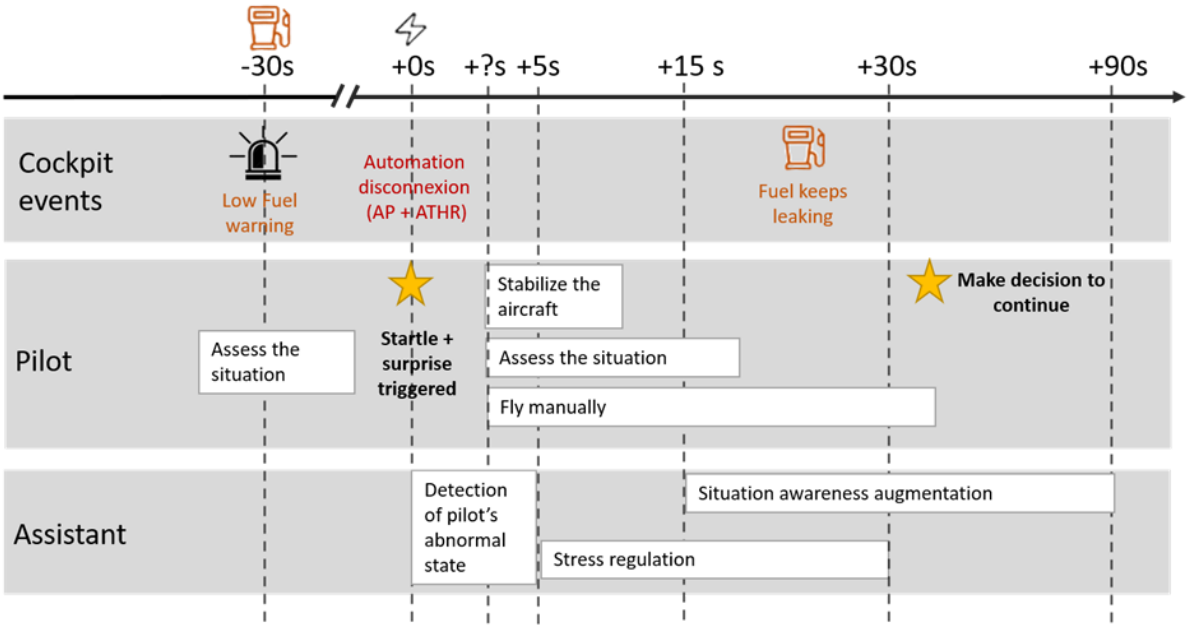


Figure 22. Lightning strike scenario script

During each scenario, photoplethysmograms, electrocardiographs, electromyographs, galvanic skin responses (GSR), respiration rates, gaze, participants' view, and participants' face were recorded.

When the startle scenario was completed, the participant was debriefed about FOCUS through a semi-structured interview and was invited to fill 5 questionnaires (Usability, HAT, SUS, Social Acceptance, HAIQU).

2.6.3. Data Analysis and Results

Subjective Feedback

The questionnaire's result analysis provided subjective perception tendency among the participants. All questionnaires are provided in the annex section. The following sections summarize the questionnaire's results.

Usability

Our scenario worked for startle and surprise for most of the participants. Overall FOCUS improved SA but the majority had difficulty perceiving the attention getters and at the same time were bothered by them. About the stress support, the vibration was noticed by 3 participants only and the green light was perceived by 5 who at the same time could not breathe accordingly.

There is a consensus about (score>4):

- Easy to use and understand

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- I understand what it is doing and why

HAT Questionnaire

There were mixed feelings for most participants and the responses show a certain consistency between them. Three participants were fundamentally negative about FOCUS and three mostly positive. Participants positively perceived the autonomy, kindness, and clarity of communication. Shared mental models and interdependence have lower scores which may indicate that the tailoring proposed could be improved to better fit their expectancies.

There is a consensus about (score >4):

- FOCUS autonomy
- HMI is understandable

And an agreement about (score >3.6):

- Supports situation awareness
- Supports to well-being and safety
- Acts in my best interest
- Effective fostering trust and comfort over time

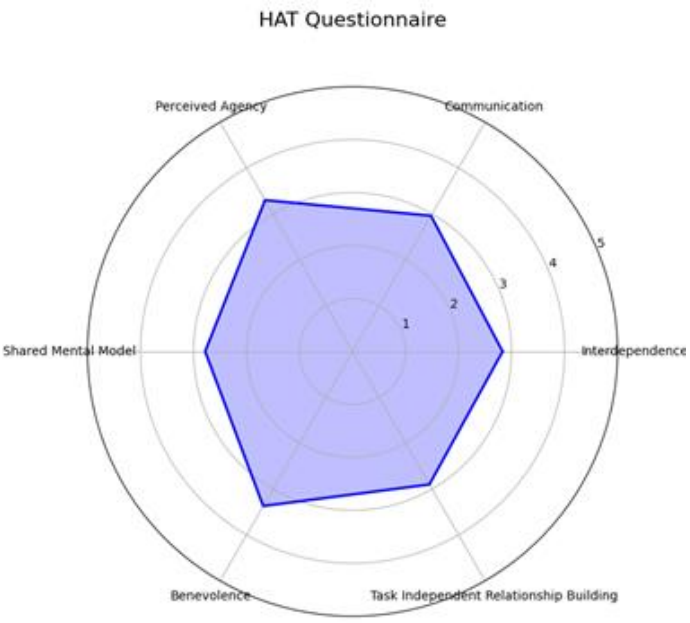


Figure 23. HAT questionnaire spider view

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

System Usability Scale (SUS)

A SUS score between 50 and 70 which is rated as 'correct'. One participant is missing, one participant under 50, all others between 50 and 70.

Best overall results on those questions:

- I thought the system was easy to use
- I would imagine that most people would learn to use this system very quickly
- I felt very confident using the system

The following Figure 24 shows the SUS score for each participant (x axis = participant number, y axis = SUS score)

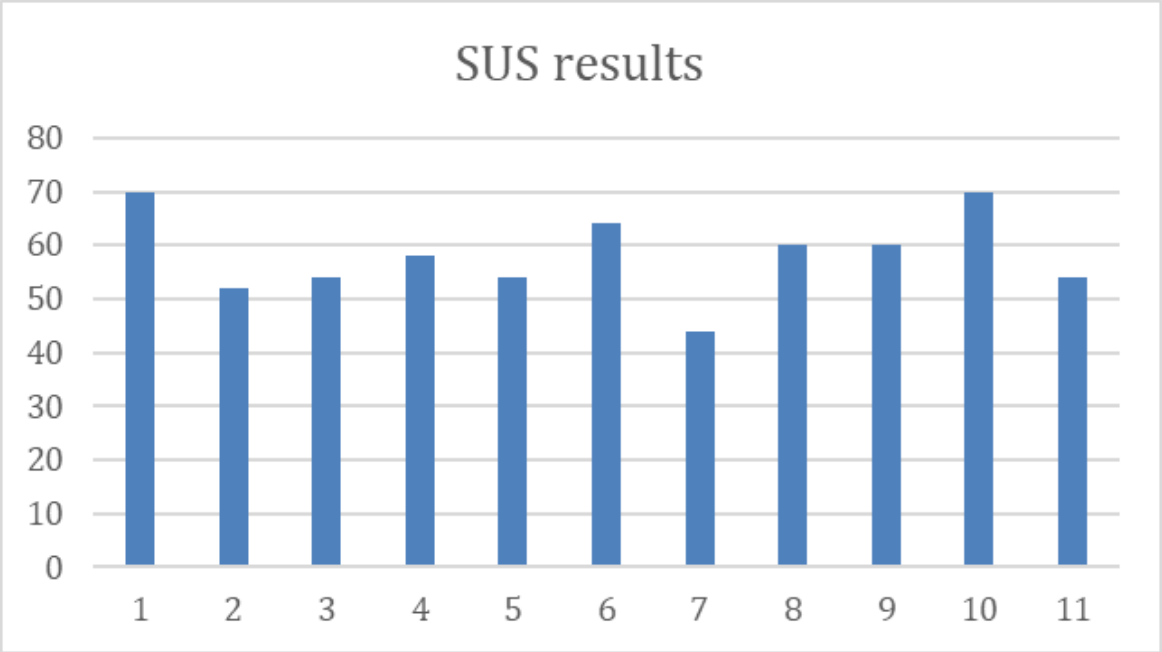


Figure 24. SUS results

Interviews

The feedback from the pilots regarding FOCUS during the startle and surprise scenario was mixed but provided valuable insights. Half of the participants found FOCUS useful, while three did not find it entirely helpful, two found it not useful at all, and one was unsure. Overall, seven pilots had a positive impression of the system, two had a completely negative impression, and three expressed mixed feelings. Many pilots appreciated certain aspects of FOCUS, such as its color-coded alerts, attention-getters, and reactivity. However, there were also recurring concerns about the system's ability to adapt to context and avoid unnecessary intrusiveness.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

One of the key positive points highlighted by pilots was the visual and auditory support provided by FOCUS. Several participants mentioned that the system's callouts and warnings were helpful, especially in maintaining situational awareness (SA). For example, the red warnings were widely regarded as relevant and effective, while the caution alerts (e.g., vertical speed or heading deviations) were seen as less pertinent, particularly when the autothrottle (ATHR) was engaged. Some pilots appreciated the ability to adjust sensitivity levels, which they found to be a well-designed feature. Additionally, the system's quick reaction time and its ability to track the pilot's gaze were seen as advantages over a human pilot monitoring (PM).

However, several areas for improvement were identified. Many pilots felt that FOCUS could be more intelligent and context aware. For instance, some alerts were perceived as non-pertinent or intrusive, especially when they interrupted the pilot's visual circuit or focused on parameters like vertical speed or heading that were not always critical. Pilots suggested that FOCUS should better prioritize alerts based on the operational context and take into account the pilot's actions or corrections. For example, if a pilot had already addressed an issue, FOCUS should recognize this and avoid redundant alerts.

Another common critique was the lack of contextual understanding compared to a human PM. While FOCUS was seen as highly efficient and precise, it was noted that a human PM could better anticipate needs, manage priorities, and assist with decision-making processes like FORDEC (Fact, Options, Risks, Decision, Execution, Check). Some pilots expressed a desire for FOCUS to go beyond its current role and provide strategic support, such as helping with diversion planning or offering calm, voice-based summaries of the aircraft's status.

Regarding trust and confidence, opinions varied. Some pilots expressed full confidence in FOCUS, particularly for its warning systems, while others were less certain, especially in degraded or high-stress situations. A few participants felt that more practice with the system would increase their trust. Additionally, some pilots suggested that FOCUS could improve by providing more information about why certain alerts were triggered, which would help them better understand and anticipate its behaviour.

In terms of distraction, most pilots did not find FOCUS overly disruptive, though some noted that excessive alerts could be distracting, particularly on higher sensitivity settings. The ability to disregard alerts easily was appreciated, but pilots emphasized the need for a better balance between providing necessary information and avoiding unnecessary interruptions.

Finally, several pilots envisioned a more collaborative role for FOCUS in the future. They suggested that the system could adapt to individual pilots' scan patterns and provide more natural, intuitive interactions, such as voice-based queries (e.g., "Siri, how much fuel do we have in terms of time?"). Some also proposed that FOCUS could anticipate issues and offer proactive support, rather than reacting to events as they occur.

In summary, while FOCUS was generally well-received for its reactivity and visual support, pilots highlighted the need for greater contextual awareness, better prioritization of alerts, and a more

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

collaborative, anticipatory approach to truly complement a human PM. These improvements could enhance the system's effectiveness and make it a more intuitive and trusted partner in the cockpit.

Platform logs and simulation data

The logs from the platform and the simulator were not initially intended to be used for the final analysis of VAL2. However, they provide interesting qualitative insights into the impact of the startle scenario on aircraft parameters and the overall situational awareness of the evaluated pilot, as observed through eye tracking data. In this section, we present the data of two participants as an example.

A global SA score was calculated by FOCUS, this score significantly increased after the startle event. Since we could not have a baseline group (i.e., a startle scenario without FOCUS), it is not possible to draw conclusions about the impact of FOCUS on SA score. Since pilots lost autopilot and auto thrust, they had to be much more focused on the PFD parameters.

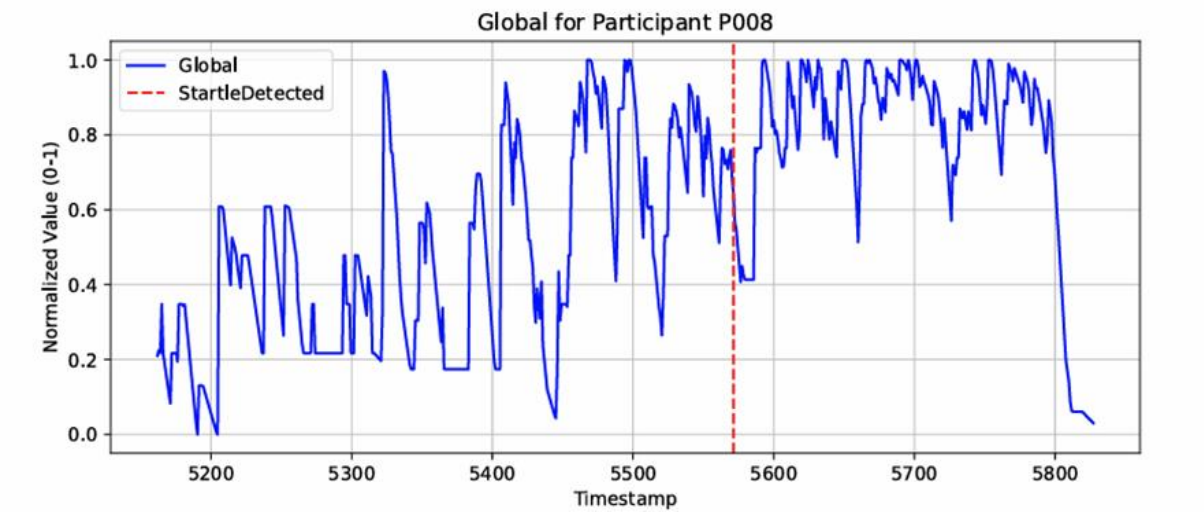


Figure 25. Global SA score for P008

Interestingly, a significant decline in SA was observed among several participants after the event, followed by a recovery above the average pre-event score, which persisted until landing.

Additionally, several parameters given by the A320 simulator were recorded. The following figure shows altitude, vertical speed, airspeed, N1 for engine 1. Some information was used dynamically by FOCUS to trigger warnings.

The following graphs show the extent to which the pilot was disrupted following the event. Several consecutive changes of vertical speed, N1, airspeed can be observed, which is entirely consistent with what we expected given the loss of subsequent automation and the partial incapacitation linked to the startle effect. If the evolution of this data over time was not currently accounted for by FOCUS,

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

one could imagine a future agent capable of detecting abnormal situations using this type of information.

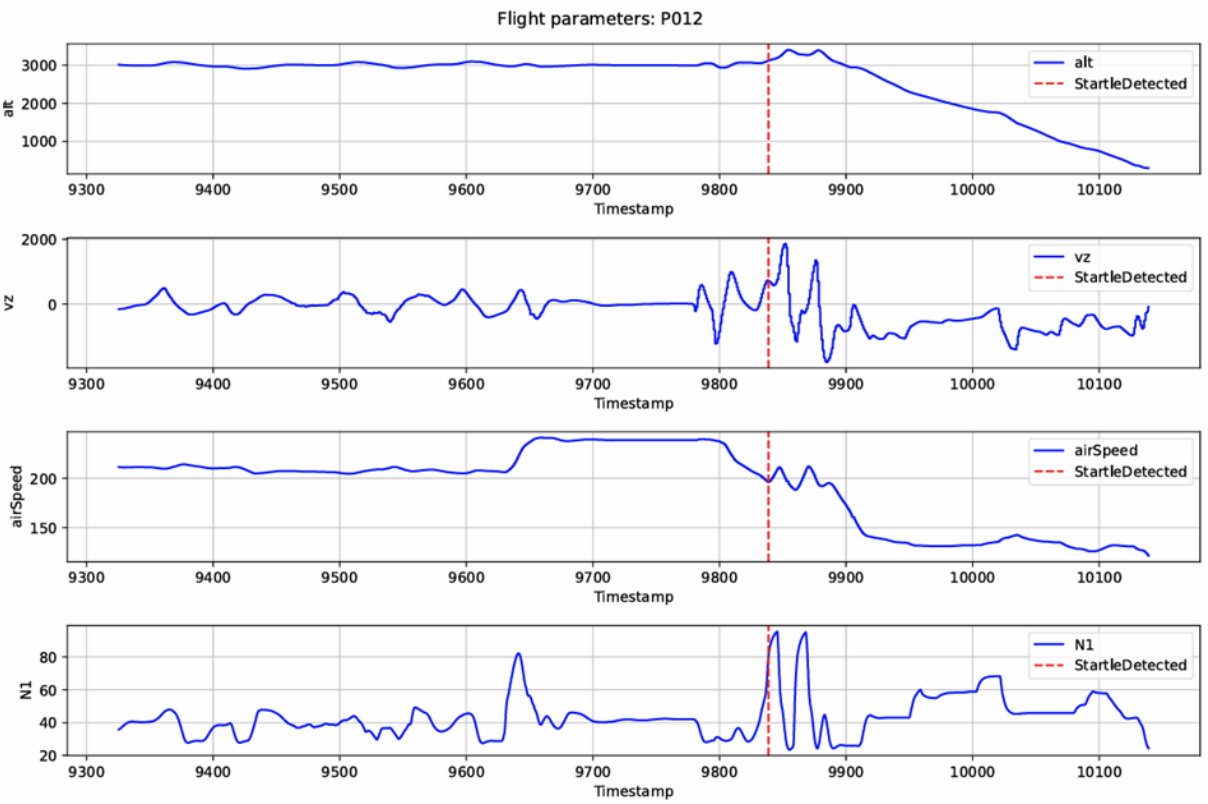


Figure 26. Graph plots of simulator outputs

Eye tracking data

The purpose of using the eye tracker in FOCUS was to enable the tracking of fixations on the various AOIs. We conducted an exploratory qualitative study by calculating, in post-processing, more complex metrics based on the data from the last 6 participants (the most experienced ones) to identify the most promising metrics for future work. Through an analysis of the scientific literature, and particularly the work done in the doctoral thesis of Lounis (2000) we selected the following metrics: stationary entropy, transition entropy, Lempel-Ziv complexity, and the K coefficient. The next section will focus on the metrics used in the assistant.

The following figure shows the results of P012 for the metrics already calculated by FOCUS for all the cockpits. We can see that attention was well distributed during the activity before the startle event and that it was limited to the PFD afterward. Regarding the duration of fixations, we can observe that it was more uniform and shorter after the event. These observations were confirmed across other participants.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

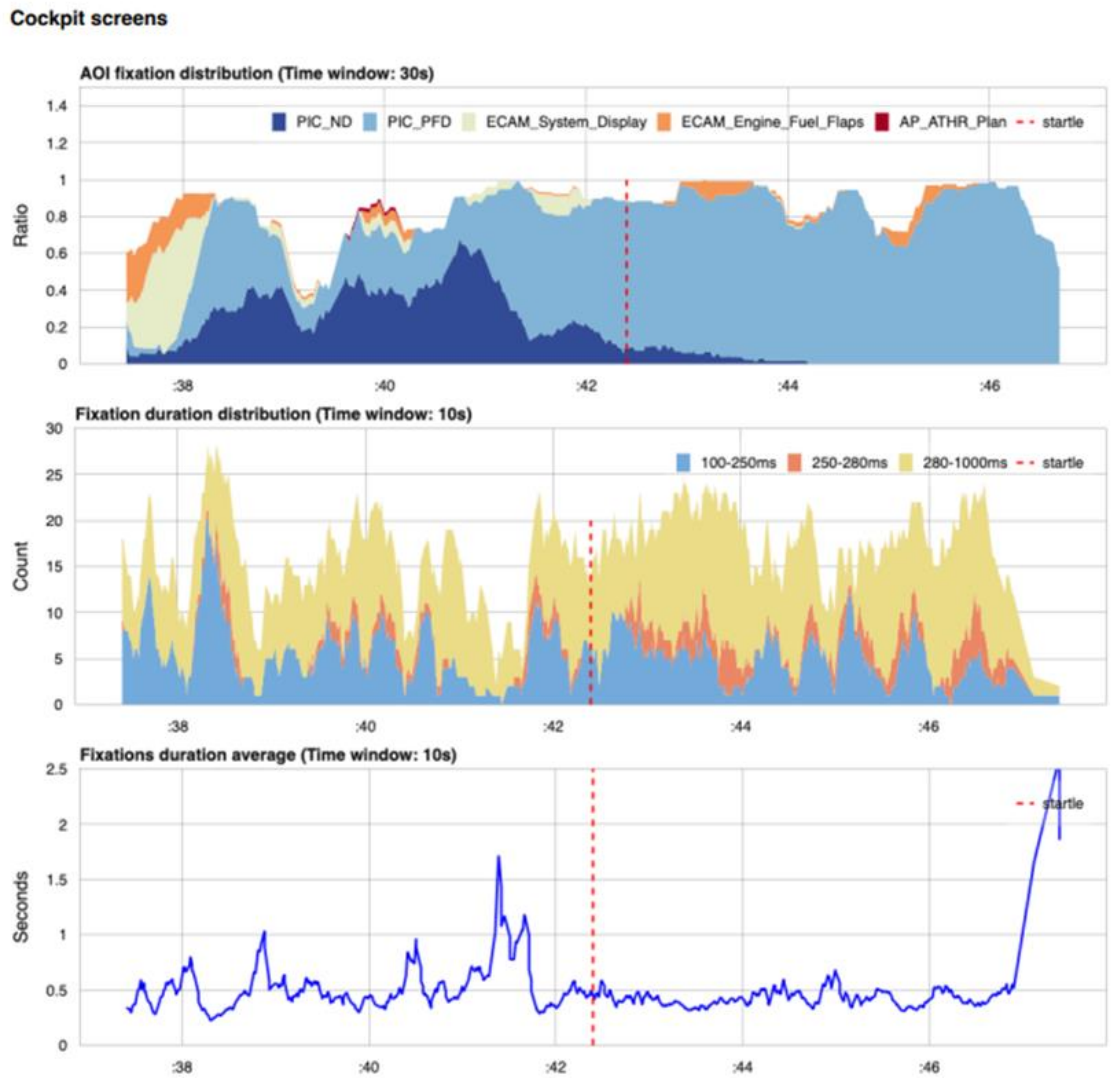


Figure 27. AOI fixation duration & distribution and average fixation duration

By examining the distribution of fixations within the PFD more closely, we observe that it is very consistent with what we know about the pilot's activity and the scenario's requirements. Until the startle event, the PFD was not the focus of attention, which explains the few data points on the graph. Afterward, however, attention shifted to the PFD, and we can see the distribution of time across different areas of interest. For example, immediately after the event, with the loss of ATHR, the aircraft gained a lot of energy with the increase in N1, leading to an increase in speed, pitch, and consequently, vertical speed. We see that the pilot tried to correct this by looking at the speed, attitude, and heading to correct and maintain the axis. It is quite logical that vertical speed was barely looked at since it is a derivative of attitude. A less experienced pilot would likely behave differently. Toward the end of the approach, we see that the pilot focused on attitude and heading, which was

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

also logical since the landing gear was down, the aircraft was stabilized in speed, and the pilot needed to maintain glide path and axis to land.

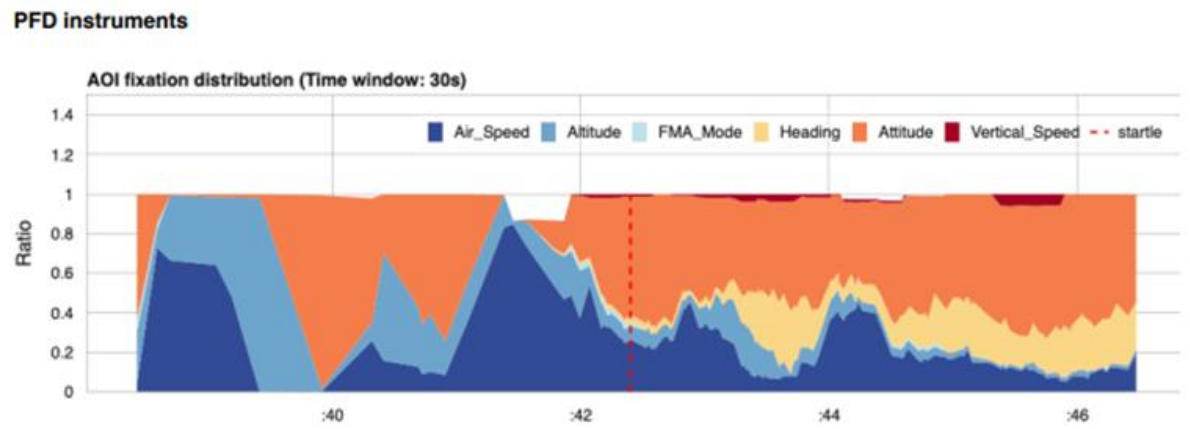


Figure 28. PFD AOI fixation distribution

This section will focus on the assessment of the advanced metrics we selected. These were as follows:

Stationary Entropy: Stationary entropy measures the distribution of fixations across different areas of interest (AOIs) at a given time. It indicates how spread out or concentrated visual attention is. High entropy suggests a more uniform distribution of fixations, reflecting broader exploration, while low entropy indicates focused attention on specific AOIs. This metric helps identify areas that attract the most visual interest.

Transition Entropy: Transition entropy assesses the predictability of movements between AOIs. It examines the sequence of fixations to understand the pattern of visual exploration. High transition entropy indicates random and unpredictable transitions, suggesting a more exploratory search strategy. Low transition entropy implies structured and predictable movements, reflecting a well-defined visual search pattern.

Lempel-Ziv Complexity: Lempel-Ziv complexity evaluates the diversity of scanpaths by analysing the sequence of fixations. It measures the number of distinct subsequences within the scanpath, reflecting the complexity of visual search patterns. High complexity indicates varied and less repetitive scanpaths, suggesting exploratory behaviour. Low complexity suggests repetitive and predictable scanpaths, often due to focused tasks or familiar stimuli.

K Coefficient: the K coefficient can indicate whether attention is focal or ambient. A positive K coefficient suggests focal attention, meaning the participant's gaze is concentrated on specific areas, indicating detailed processing. Conversely, a negative K coefficient indicates ambient attention, where the gaze is spread out, suggesting broader environmental scanning.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

The following figures show these metrics calculated for participant P012. Firstly, we can observe a very clear change in attentional behaviour after the startle event in both stationary entropy and transition entropy. Indeed, in both cases, we see a noticeable decrease. The drop in stationary entropy indicates that attention becomes more focused on a few AOIs, while the drop in transition entropy means that the scan pattern becomes less complex. This is consistent with what we observed when analysing the distribution of fixations within the PFD and the cockpit. These indicators provide global metrics that reflect this change in behaviour. Regarding LZC across all participants, we do not observe any evident change in the indicator. However, we can note that the LZC is high, which indicates a complex visual circuit. This was indeed expected for an activity such as piloting. The K coefficient offers a broader understanding of attention distribution; it is negative before the event and positive afterwards indicating a switching from ambient to focal attention.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

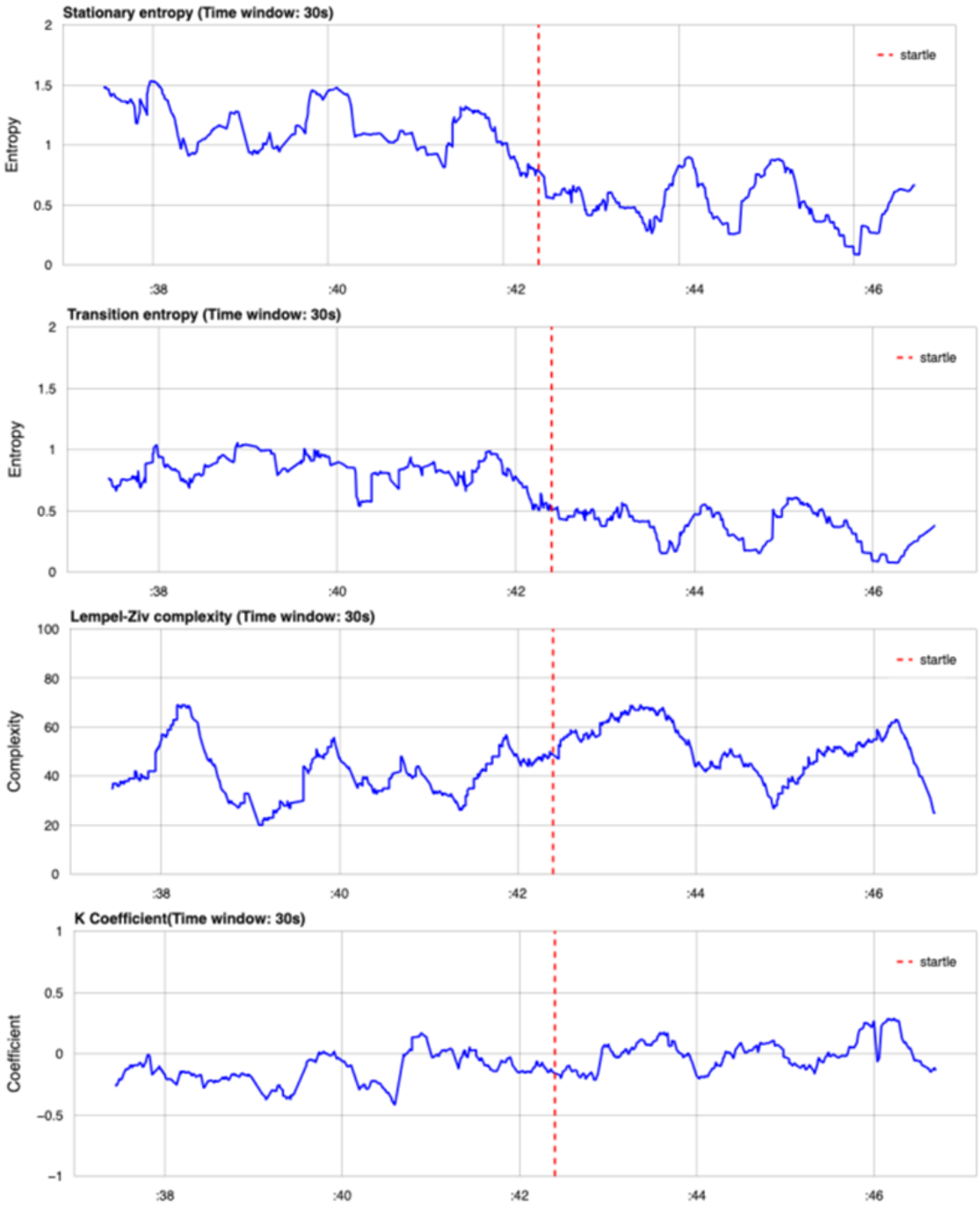


Figure 29. Entropy, LZC, K coefficient

Physiological data

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Physiological data were used to understand the participant response to the event. Heart rate activity (measured by electrocardiogram or photoplethysmogram) and electrodermal activity were recorded like in the startle and surprise fundamental experiment. Moreover, electromyography and respiratory activities were also used to confirm physiological response to the startle.

Only electrodermal activity was significantly different (W = 8, p = .05) before the onset of the event and after. The activity was significantly higher during the period between the event and 15 seconds later than during the baseline period (30 seconds before the event). The following figures show the electrodermal activity of P005 and P007 between thirty seconds before the onset and thirty seconds after.

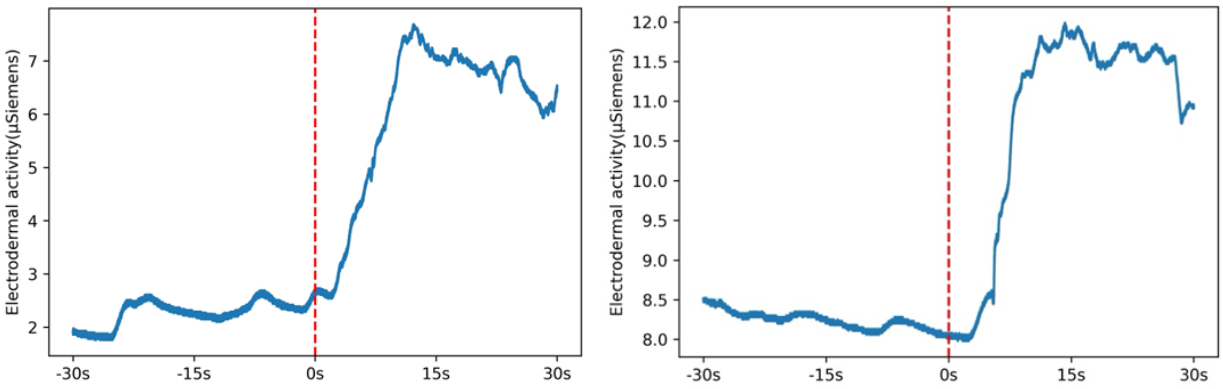


Figure 30. Electrodermal activity of the P005 and P007

Other physiological metrics were not (statistically or visually) different before or after the onset of the event. Heart rate, heart rate variability, electromyography and respiratory activity were not impacted by the simulated lightning strike nor by the stress regulation support. Breathing technique does not seem to have been applied by the participants.

These results corroborate those of the fundamental experiment: electrodermal activity seems to be a meaningful indicator of the startle effect. Nevertheless, the use of statistical analysis may be challenged due to the low number of participants. Additional participants would make this preliminary result more robust.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

2.7. TRL Overview: update

The planned trajectory is confirmed, and we have achieved the targeted TRL level stated in D6.3.

Table 8. TRL progress in UC1 for the HMI components.

Component: Startle effect detection, Situation awareness (SA) augmentation, and Stress regulation support		
TRL	**Month**	**Activity to reach selected level**
1	1	Concept formulated at the project's start
2	6	Interviews and design meetings with pilots to produce prototype specifications
3	16	VAL1 with a first version of the prototype: • Situation Awareness Augmentation and Stress regulation support: test in an aircraft simulator involving 5 pilots • Startle detection module: laboratory test
4	28	VAL2 with a second version of the prototype: • Situation Awareness Augmentation and Stress regulation support: test in a realistic aircraft simulator with pilots in a relevant operational scenario • Startle detection module: artificial intelligence module trained with new datasets and tested with data collected in simulator in a relevant operational scenario
5	-	
6	-	

2.8. VAL 2 Results

2.8.1. Results Overview

Table 9. High-level results in relation to the EASA Validation objectives.

High-level Result (HAT Objectives)	**Brief description**	**Related validation objective (D6.3)**
Startle detection	The module accuracy and reaction time works accordingly to requirements	HF-07

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Stress regulation	Vibration was barely perceived, but it works anyway (activity 2), visual feedback sometime could be linked to a tunnelisation effect	HF-07
Situation awareness	SA support design and behaviour was perceived positively	HF-02; HF-03; HF-06; HF-07; HF-08; HF-28.
Usability and design understanding	Usability was perceived as correct/good, the overall design and functionalities of FOCUS were fully understood	HF-16; HF-26; HF-27
Control on FOCUS behaviour	Both tailoring and control actions on FOCUS were used and appreciated during VAL2. More tailoring could be added	HF-32; HF-34

2.8.2. Discussion

To further enhance FOCUS as an advanced pilot monitoring system, several improvements and additional features could be explored. One key area is **personalization and adaptability**. FOCUS could be tailored to individual pilots by leveraging data collected from their specific flying patterns, preferences, bio signals and decision-making behaviours. By using machine learning algorithms, the system could evolve over time to better align with each pilot's unique style. This would not only improve the system's effectiveness but also foster trust and confidence between the pilot and the assistant. Additionally, moving away from static thresholds to **dynamic, context-sensitive thresholds** could provide a more nuanced approach. For instance, experienced pilots might benefit from more lenient alerts during routine operations, while stricter monitoring could be activated during high-stress or critical phases of flight. This adaptability would ensure that FOCUS remains relevant and supportive across a wide range of scenarios.

Another significant improvement could involve **enhancing interaction methods**. While tactile human-machine interfaces (HMI) are functional, transitioning to **voice-based interactions** could significantly reduce cognitive workload and improve usability. Voice commands would allow pilots to interact with FOCUS more naturally, especially during high-workload situations where hands-free operation is crucial. Furthermore, integrating **natural language processing (NLP)** capabilities could enable more intuitive and conversational exchanges, making the system feel like a collaborative partner rather than a tool.

FOCUS could also be equipped with **anomaly detection capabilities** to identify potential issues related to pilot behaviour or perception. For example, the system could monitor for signs of fatigue, stress, or tunnel vision—common challenges in high-pressure environments. By detecting these anomalies early, FOCUS could provide timely interventions, such as suggesting breaks, reorienting attention, or offering decision-making support. This proactive approach would not only enhance safety but also contribute to the overall well-being of the pilot.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

In addition to monitoring, FOCUS could take on a more active role in **task delegation and dynamic tailoring**. Based on real-time conditions and the pilot's workload, the system could autonomously prioritize tasks, delegate non-critical functions, or even adjust its level of involvement. For instance, during a complex approach phase, FOCUS could take over routine monitoring tasks, allowing the pilot to focus on higher-level decision-making. This dynamic collaboration would optimize efficiency and reduce the risk of human error.

Finally, integrating a **FORDEC-based decision-making framework** could elevate FOCUS's capabilities to a new level. FORDEC (Fact, Options, Risks, Decision, Execution, Check) is a structured process widely used in aviation to enhance decision-making. By embedding this framework into FOCUS, the system could guide pilots through complex scenarios, offering step-by-step support and access to relevant operational manuals (OM) or external resources. This would not only improve situational awareness but also ensure that decisions are made systematically and with confidence.

Regarding the use of eye tracking, the exploratory study allowed us to identify indicators that could be used to enable FOCUS to detect changes in attentional behaviour that might be associated with shifts in the assistant's strategy.

2.8.3. Key findings

Pilot Feedback: Mixed responses highlighted the need for context-aware alerts and more intuitive interaction methods. Pilots appreciated the system's reactivity but desired better prioritization of alerts and more natural interaction modes, such as voice commands.

Eye Tracking Metrics: Advanced metrics like stationary entropy and transition entropy provided insights into attentional shifts post-startle, indicating a need for dynamic and context-sensitive alert systems.

Physiological Responses: Electrodermal activity was a reliable indicator of startle effects, underscoring the importance of integrating physiological monitoring into assistive technologies.

Improvements and Future Directions

Personalization and Adaptability: Future iterations of FOCUS should incorporate machine learning to adapt to individual pilot behaviours and preferences, enhancing trust and usability.

Enhanced Interaction Methods: Transitioning to voice-based interactions and integrating natural language processing could reduce cognitive workload and improve system usability.

Proactive Support: Implementing anomaly detection for signs of fatigue or stress and providing proactive decision-making support could elevate FOCUS's role from a reactive tool to a proactive partner in the cockpit.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

2.9. UC1 VAL 2 Conclusions

2.9.1. UC1 Research Questions

Research question 1: What are the independent and combined effects of startle and surprise on subjective feedback, behaviour (task performance and gaze behaviour), and physiological parameters in aviation-inspired tasks?

The study VAL2 activity 1 explored the distinct and combined impacts of startle and surprise on subjective experiences, task performance, and physiological responses in a simulated flight environment. Surprise alone increased skin conductance but did not significantly affect task performance. Startle, however, impaired communication task performance, heightened skin conductance and heart rate, and narrowed attention. When combined, startle and surprise intensified subjective feelings of startle and surprise, prolonged heart rate increases, and further impacted task performance and gaze behaviour. The combination led to a more pronounced narrowing of attention and a significant shift in gaze distribution. Physiologically, both startle and surprise independently raised skin conductance, while their combination resulted in a more substantial and prolonged heart rate increase. These findings suggest that the combined effects of startle and surprise are more detrimental than their individual impacts, highlighting the need for strategies to mitigate these effects in high-stakes settings like aviation. No datasets were available that aggregated both effects. The dataset thus created enabled supervised learning, which was then compared to operational measurements.

Research question 2: Can machine learning models effectively distinguish between startle, surprise, and baseline states using physiological signals in aviation-inspired tasks?

By integrating ECG, GSR, and PPG signals through a structured preprocessing pipeline and extracting relevant features from raw physiological data, our classification framework, comprising Support Vector Machines (SVM), Naive Bayes, and XGBoost demonstrated that machine learning techniques, particularly when enhanced by late fusion, can effectively distinguish between baseline, startle, and surprise conditions.

Specifically, late fusion strategy consistently outperformed unimodal approaches, highlighting the complementary nature of different physiological signals in capturing cognitive-affective states. Among the modalities, EDA and PPG features proved most informative, contributing significantly to the overall classification performance. Our results showed high accuracy across multiple binary comparisons, with the SVM model achieving up to 89.62% accuracy in distinguishing surprise from baseline conditions. Even in the more complex three-class scenario (startle vs. surprise vs. baseline), the XGBoost model achieved 74.96% accuracy, more than twice the chance level, demonstrating the robustness of the proposed approach.

Window size also played a crucial role in model performance. For binary classifications such as Startle vs. Baseline or Surprise vs. Startle, performance remained relatively stable across different window

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

lengths. However, in the Surprise vs. Baseline condition, longer windows (5s and 7s) provided a clearer physiological distinction, whereas in the more dynamic three-class task, shorter windows (3s) led to significantly better performance. These findings suggest that task complexity and the nature of the physiological response should inform the choice of temporal windowing in real-time applications.

Overall, our fundamental study provides strong evidence that physiological signals can be used for accurate, real-time detection of startle and surprise responses. The consistent performance of late fusion across time windows highlights its practical potential for real-world deployment, paving the way for intelligent monitoring systems that enhance safety in high-risk environments. Building on these findings, we plan to apply these methods to the VAL2 dataset to further validate the effectiveness and generalizability of our approach. Several techniques have been tested with open data on startle and data produced by the fundamental experiments. We can consider having reached TRL 4 for this function.

Research question 3: Can vibrotactile feedback interventions effectively mitigate the physiological and psychological effects of startle responses?

The VAL2 activity 2 study investigates the effectiveness of vibrotactile interventions in mitigating the startle response, utilizing simulated heartbeats; it aligns with prior research on vibrotactile feedback for stress regulation. The experimental protocol successfully induced stress, as evidenced by negative differences in RMSSD and pNN50 between rest and task periods, and subjective reports confirming a startle response. Among the interventions tested, a constant vibrotactile feedback pulsing at 60 beats per minute proved most effective, nearly eliminating the RMSSD difference and significantly mitigating the startle response. While all interventions provided some support, the 60 beats per minute condition outperformed others, including those at 80% of the heart rate during tasks. This suggests that simulating a slower heart rate can better modulate the effects of the startle response. The technique is practical, requiring no physiological sensors, making it an affordable and easy method for stress regulation. Participants associated the feedback with heartbeats, enhancing cardiac awareness and interoception, and potentially serving as warnings for abnormal situations. Overall, the study demonstrates that vibrotactile interventions, particularly those simulating a slower heart rate, can effectively mitigate the startle response and support stress regulation.

The experimentation on the stress support function allowed us to refine the use of vibration in the FOCUS prototype, based on scientific data that, to our knowledge, did not previously exist. We can consider having reached TRL 4 for this function.

Research question 4: How effective is the FOCUS system in assisting pilots during startle and surprise scenarios, and what improvements are needed to enhance its contextual awareness and usability in the cockpit?

The VAL2 activity 3 evaluates UC1 assistant (FOCUS) in a realistic setup. About participants' perception of FOCUS, the question of single-pilot operation is very polarizing and may lead to biased responses. The feedback from pilots on the assistant during startle and surprise scenarios was varied but insightful. Half of the participants found FOCUS useful, while others had mixed or negative

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

impressions. Pilots appreciated features like color-coded alerts and the system's reactivity, which helped maintain situational awareness. However, concerns were raised about the system's context adaptation and potential intrusiveness. Pilots suggested that FOCUS should prioritize alerts based on operational context and recognize when issues have already been addressed. While FOCUS was seen as efficient, pilots noted that a human pilot monitor could better anticipate needs and assist with decision-making. Trust in FOCUS varied, with some pilots expressing full confidence and others desiring more practice or explanatory information for alerts. Most pilots did not find FOCUS overly distracting, though some noted that excessive alerts could be disruptive. Pilots envisioned a more collaborative role for FOCUS, with adaptations to individual scan patterns and proactive, voice-based support. Overall, while FOCUS was well-received for its visual support, improvements in contextual awareness and alert prioritization were highlighted for better integration in the cockpit.

Overall, pilots appreciated FOCUS's potential but emphasized the need for smarter, more intuitive functionality to enhance its role in the cockpit. We can consider having reached TRL 4 for the situation awareness.

2.9.2. HAIKU High-level Research Questions

HAIKU Q1: What are the common recommendations concerning Human-AI teaming for the different AI aviation applications?

Collaborative Partnership: The AI system should act as a collaborative partner rather than just a reactive tool. Pilots appreciated features like color-coded alerts and the system's quick reactivity, which helped maintain situational awareness. However, they also desired a more proactive and anticipatory role from the AI, suggesting that it should adapt to individual pilots' scan patterns and provide intuitive, voice-based interactions.

Contextual Awareness and Adaptability: The AI should be context-aware and adaptable, prioritizing alerts based on the operational context and recognizing when issues have already been addressed by the pilot. This would help avoid unnecessary intrusiveness and ensure that the AI provides relevant support tailored to the current situation.

Strategic Support: Beyond immediate alerts and warnings, the AI should offer strategic support, such as assisting with decision-making processes like FORDEC (Facts, Options, Risks, Decision, Execution, Check). This would make the AI a more integral part of the pilot's decision-making process.

Trust and Transparency: Building trust is crucial. The AI should provide clear explanations for its alerts and actions, helping pilots understand and anticipate its behaviour. This transparency can increase confidence in the system, especially in high-stress or degraded situations.

Minimal Distraction: The AI should balance providing necessary information with avoiding excessive alerts that could distract the pilot. The system should allow pilots to easily disregard non-critical alerts and focus on essential tasks.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Continuous Learning and Improvement: The AI should continuously learn from interactions and feedback to improve its performance and adapt to different pilots' preferences and behaviours. This would make the AI more intuitive and effective over time.

HAIKU Q2: What does it mean for AI to be explainable?

UC1 AI is supposed to help a pilot in a very critical situation where he has limited resources for interaction.

Transparency: The AI should provide clear and understandable explanations for its alerts, actions, and recommendations. This helps pilots comprehend why certain alerts are triggered and what specific issues the AI is addressing.

User Trust and Confidence: By being explainable, the AI can build trust with pilots. When pilots understand how the AI arrives at its conclusions, they are more likely to trust and rely on the system, especially in high-stress or critical situations.

Adaptability and Customization: The AI should adapt to individual preferences and behaviours, explaining how it tailors its support to different pilots' needs and operational contexts.

HAIKU Q3: How to train AI to assist humans in safety critical tasks when training data are insufficient?

Simulated Environments: Use simulated environments to generate synthetic data. Simulations can mimic a wide range of scenarios, including rare and critical events, providing a rich dataset for training AI systems. This approach was used by UC1, it allows for extensive testing and refinement of AI models without the complexity of high realistic environments.

Expert Knowledge Integration: Incorporate expert knowledge and heuristic rules into the AI system. This can involve creating rule-based systems or integrating expert insights to guide the AI's decision-making process, especially in scenarios where data is scarce. FOCUS's core is based on such an approach.

2.9.3. Research Recommendations

Upon completion of VAL2 activity 3, the main lessons learned that point to new design directions are reported in Table 10 below.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Table 10. UC1 VAL2 lessons learned and new design directions

Insight	Functional Requirements	Proposed solution for an improved assistant
Pilots appreciate visual and auditory alerts but find some alerts intrusive.	Reduce non-pertinent alerts and prioritize based on context.	Implement context-aware algorithms to filter and prioritize alerts, avoiding interruptions during critical tasks.
FOCUS lacks contextual awareness and human-like understanding.	Incorporate contextual intelligence to adapt to pilot actions and operational conditions.	Develop AI that tracks pilot corrections and adjusts alerts dynamically, mimicking human situational awareness.
Pilots want more transparency in why alerts are triggered.	Provide explanations for alerts and system behaviour.	Add voice or visual explanations for alerts (e.g., “Vertical speed alert: you are below the glide path”).
Sensitivity adjustment is appreciated but could be more intuitive.	Allow dynamic sensitivity settings based on flight phase and workload.	Introduce adaptive sensitivity that adjusts automatically based on real-time conditions and pilot behaviour.
Pilots find red warnings effective but caution alerts less relevant.	Improve relevance of caution alerts and align them with operational priorities.	Redesign caution alerts to focus on critical deviations (e.g., speed, altitude) and suppress less urgent ones.
FOCUS is seen as too reactive and lacks strategic support.	Enable proactive support and decision-making assistance (e.g., FORDEC).	Integrate FORDEC-based guidance and strategic planning tools to help pilots anticipate and manage complex situations.
Pilots desire more natural interactions, such as voice commands.	Implement voice-based interaction for hands-free operation.	Add voice command functionality (e.g., “FOCUS, how much fuel remains?”) to reduce cognitive and manual workload.
Some pilots feel FOCUS could adapt to their individual visual circuits.	Personalize alert placement and timing based on pilot preferences.	Allow customization of alert locations and timing to align with each pilot’s visual scanning patterns.
Trust in FOCUS varies, with some pilots wanting more practice to build confidence.	Provide training modules and simulations for pilots to familiarize themselves with FOCUS.	Develop interactive training tools to help pilots understand FOCUS’s logic and build trust in its capabilities.

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Pilots suggest FOCUS could assist with diversion planning and strategic tasks.	Expand functionality to include strategic support (e.g., diversion planning).	Add features for route optimization, fuel management, and diversion assistance to enhance strategic decision-making.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

3. Use Case #2 – Flight Deck Route Planning/Replanning

OlivIA (OperationaL Intentions adVIser for Aviation) is a decision-making support tool for flight crew handling in-route threats and/or selecting alternate flight routes. Its purpose is to alleviate pilot's cognitive workload, enabling higher quality, holistic mission-focused decisions, in complex situations. It offers route adjustments prioritized and evaluated according to operational intentions. The tool also benefits the Operations Control Centre, enhancing coordination and response times.

3.1. Deviation from Validation Plan (D6.3)

VAL2 was performed according to the validation plan described in D6.3 with no deviations.

3.2. VAL2 Objectives

In accordance with D6.3, the HAT objectives in VAL2 addressed two main concepts:

- OBJ-01 Combi enables effective and efficient high-level intentions communication in the team.
- OBJ-03 HAT design methodology able to support HAT safety and effectiveness assessments.

In other words, the objective of VAL2 was to assess the value of operational intentions as a new way to support mission management in Regional Segment aviation. This support is provided by OLIVIA assistance concept and considers the design requirements for effective Human-Autonomy Teaming.

The associated research questions were:

- RQ1: To what extent does the integration of relevant data and complementary assessments enhance the team's decision-making processes?
- RQ2: How does OlivIA facilitate the communication of high-level intentions within the team?
- RQ3: What is the user perception of OlivIA's usability in an operational context?
- RQ4: What effect does OlivIA have on perceived workload and situational awareness?
- RQ5: Does the HAT design methodology enable effectivity assessments?

The objectives and research questions were addressed during VAL2.

3.3. VAL2 Activities and Methods

The focus of VAL 2 was to test the effectiveness and efficiency of bidirectional communication via intentions (OBJ-01) and to explore design concepts to build HAT systems (e.g. the suitability of the explainability strategy) using a high-fidelity prototype (OBJ-03). For this purpose, challenging flight re-planning scenarios were simulated with airline pilots (one at a time) in two different conditions: with and without OliviA support. Each simulation session took 3 to 4 hours, including: presentation of the assistance concept and the experiment purpose; a training session on the procedures for flight re-

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

planning with and without OlivIA; execution of four simulated scenarios, questionnaires administration and self-confrontation interviews.

The RQs were then derived from the HAT objectives. For these questions, a combination of questionnaires, quantitative data, self-confrontation interviews were utilised to systematically collect all the necessary data. All the eight questionnaires used are presented in the section Data collection tools. The process for generating relevant insights followed this path:

HAT Objective → EASA-like Requirement → Research Question → Questionnaires and Self confrontation interview and specific questions -> Quantitative data → new insights for OlivIA.

3.3.1. Participants

Participants were 10 ATP-licensed pilots, including 6 women and 4 men with a total flight time ranging from 1,400 to 18,000 flight hours. The average flight time is 8,900 hours with standard deviation (SD) of 5,083.50 hours. All pilots fly regional aircraft (e.g. A320, B737) or have a significant experience in regional flights. Among these 10 pilots, one has military experience, and another has experience in both business and bush flying. The mean age of the pilots was 40,5 years old. 8 out of 10 are or have been captains.

3.3.2. Simulator/Apparatus

The experiment was conducted in a flight simulator provided by Thales, designated as FlytX. This simulator serves as a prototyping tool capable of integrating highly representative avionics systems. To enhance the TRL of the demonstrator, a Flight Management System (FMS200) has been incorporated into the simulator.

In addition, a specific user interface has been developed for OlivIA. Two screens for the COMBI translators provide the interface, allowing for the communication of the pilot's operational intentions and facilitating the reception of the system's solutions with varying levels of explainability. This advanced interface aims to improve interaction efficiency and clarity, thereby enhancing overall system usability and the pilot decision-making processes.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Figure 31. Flytx Simulator and prototype interfaces

3.3.3. Scenarios

The flight scenarios were regional flights from Marseille, France (LFML, MRS) to Munich, Germany (EDDM, MUC) on a winter morning in January. The planned alternate airport was Frankfurt (EDDF, FRA). Due to bad weather at destination, extra fuel for 1 hour flight time was added to the minimum required fuel (that already accounts for 30 min final reserve fuel). All simulated scenarios started with the aircraft 2 min before arriving at BETOS - the first waypoint of the Standard Arrival Route (STAR) at EDDM. At this point, the air traffic controller announced airport closure due to an ongoing snow thunderstorm and requests to hold at position before advising the preferred alternate. Six scenarios were created, mixing weather conditions and randomizing airport KPIs. Four airports were presented as alternates for each scenario. Each pilot completed four scenarios, half of them being executed with OlivIA support for assessing and selecting alternative routes and the other half without it (alternating runs with and without support). One specific scenario was developed for two training runs – one with OlivIA and one without it

3.3.4. Procedure

Before the proposed exercise, participants were briefed about its purpose and the steps they would go through, as shown in Figure 32. The assistant concept and the operational intentions meaning were presented and the participants were allowed to solve any doubts they had at this moment. Before

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

each run, pilots were briefed for the flight and had access to all the information needed. Once briefed, the simulation started in-flight.

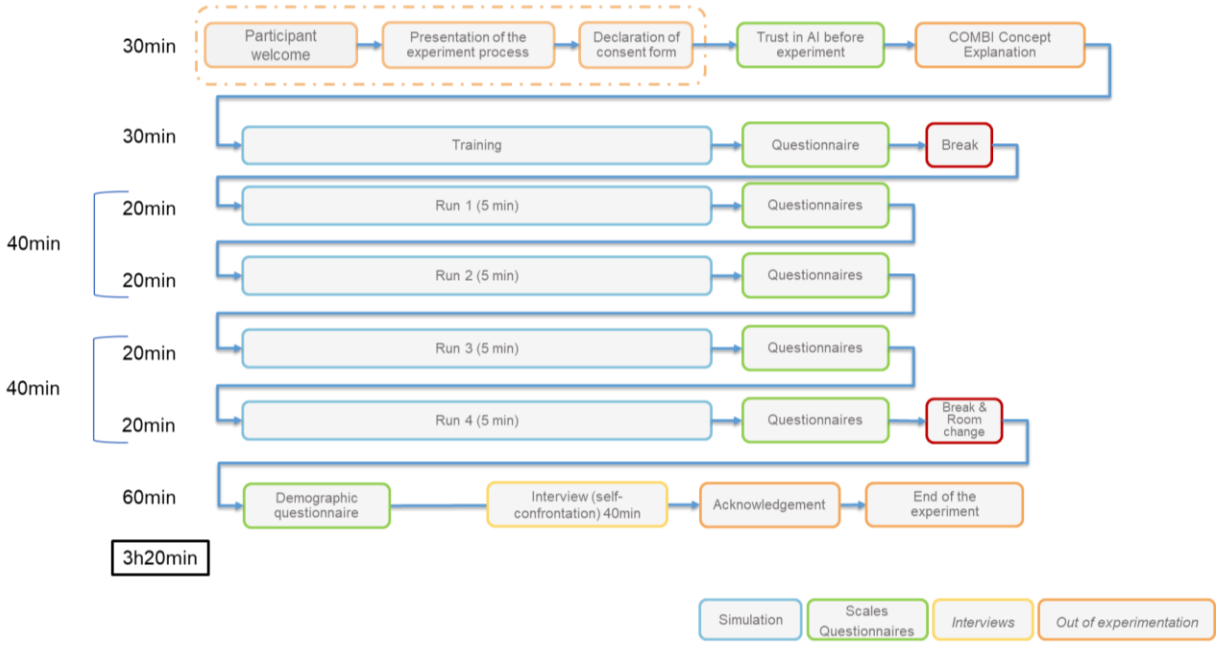


Figure 32. Experimental exercise chronology

Each participant started with two training runs – one with OlivIA and one without it – to get familiar with the simulator, the interfaces of the simulator and of OlivIA and the proposed procedures to operate them. Just after the two training sessions, they answered the NASA-TLX questionnaire (Hart & Staveland, 1988) to use the results as a baseline of perceived workload. They were then equipped with an eye tracker.

Every pilot completed four runs, two with OlivIA and two without it, always alternating this condition between runs. For each run, the weather changed and was randomised between 6 weather scenarios as shown in Table 11.

In addition, the KPIs linked to the four alternate airports (Strasburg (LFST, SXB), Linz (LOWL, LNZ), Milan (LIML, LIN), and Frankfurt (EDDF, FRA)) changed randomly across each scenario. In the runs with OlivIA, the pilots indicated their intentions to get 3 route proposals to alternate airports that best aligned with them. Assessment details were shown on demand by pressing a button “Details”. KPIs on each airport could also be consulted directly on their icon in the navigation display. A second layer of information for each airport with the NOTAM details could also be consulted on demand in the navigation display.

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Table 11. Scenarios definition for each run

	Run 1		**Run 2**		**Run 3**		**Run 4**	
	Weather scenario	OlivIA	Weather scenario	OlivIA	Weather scenario	OlivIA	Weather scenario	OlivIA
Pilot 1	S1	Yes	S2	No	S3	Yes	S4	No
Pilot 2	S5	No	S6	Yes	S1	No	S2	Yes
Pilot 3	S3	Y	S4	N	S5	Y	S6	N
Pilot 4	S1	N	S2	Y	S3	N	S4	Y
Pilot 5	S5	Y	S6	N	S1	Y	S2	N
Pilot 6	S3	N	S4	Y	S5	N	S6	Y
Pilot 7	S1	Y	S2	N	S3	Y	S4	N
Pilot 8	S5	N	S6	Y	S1	N	S2	Y
Pilot 9	S3	Y	S4	N	S5	Y	S6	N
Pilot 10	S1	N	S2	Y	S3	N	S4	Y

The pilots had to choose their alternate between the four proposed airports according to their judgement and discretion. After each scenario, they had to answer the NASA-TLX (Hart & Staveland, 1988) workload questionnaire and the SART Situation Awareness questionnaire (Endsley, 1995). If they used OlivIA, an additional questionnaire was submitted concerning transparency (Hellmann et al., 2022).

After completing all runs, pilots were asked to fill out several questionnaires: the Trust in AI questionnaire, the Computer System Usability Questionnaire (CSUQ, (Lewis, 1995)), the Trustworthy AI questionnaire (Ashoori & Weisz, 2019), the AI Device Use Acceptance (AIDUA) questionnaire (Gursoy et al., 2019), and a demographic questionnaire. These questionnaires are presented in the following sections, along with their respective results.

Finally, the pilots took part in a self-confrontation interview supported by video recordings of their runs. By reviewing their own performance, they were able to reflect on their experience and verbalize the reasoning behind their actions during each of the four scenarios

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

3.3.5. Data Collection Tools

Table 12. Data collection tools used in UC2.

Tool	Objective	Type Of Collected Data
Simulation data	Assess decision-making efficiency and effectiveness	Objective quantitative data related to the re-route decision-making process: selected intentions, number of interactions with HMI (request for details on solutions and on airports, number of airports checked, total decision-making time)
Eye-tracker	Assess user's visual attention at different areas of the display, supporting investigations of usefulness and effectiveness of the operational explainability features	Objective quantitative data (gaze direction)
Trust questionnaire	Assess the propensity of user to trust in technology	Subjective quantitative data on feelings toward automated agents
NASA-TLX	Assess user's perceived workload during a task	Subjective quantitative data (mental demand, physical demand, temporal demand, performance, effort, frustration)
SART (Situational Awareness Rating Technique)	Assess user's situational awareness	Subjective quantitative data (attentional demand, attentional supply and understanding)
Transparency Questionnaire	Assess users' perceptions of transparency in recommender systems	Subjective quantitative data on inputs, outputs, interaction and functionality
CSUQ (Computer System Usability Questionnaire)	Assess system usability	Subjective quantitative data on perceived usefulness, information quality, and ease of use
Factors that influence trustworthiness questionnaire	Assess multiple dimensions of trustworthiness in AI-supported decision-making processes	Subjective quantitative data on overall trustworthiness, reliability, technical competence,

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

		understandability, and personal attachment
AIDUA (Artificially Intelligent Device Use Acceptance Questionnaire)	Assess user acceptance of AI-based systems	Subjective quantitative data (social influence, anthropomorphism, motivation, performance expectancy, effort expectancy, emotional impact, willingness to accept, objection to use)
Self-confrontation interview	Assess the user's experience with OlivIA, address specific research questions and collect feedback	Qualitative data on satisfaction, quality of solutions, impacts on the pilot role, value of intention-based bi-directional communication and suggestions for the concept refinement
Societal acceptance questionnaire	Assess user's perceptions on societal acceptance	Subjective quantitative data (usefulness, ease of use, subjective norms, attitude toward use, facilitating conditions, risk, behavioural intention, trust)

3.3.6. Data Analysis

We employed a Kruskal-Wallis, Wilcoxon test to ensure the robustness of all quantitative analyses, which were conducted using RStudio version 4.3.2 and Excel.

A Wilcoxon rank-sum test with continuity correction was conducted to compare NASA-TLX global mean scores between the training sessions and the four runs.

The Wilcoxon test was also used, along with an analysis of variance (ANOVA), to evaluate the significance of decision-making times differences between the conditions of using OlivIA or not.

The Shapiro-Wilk test, the Levene's test were used to evaluate respectively the normality of decision-making times distribution and the homogeneity of variances between decision-making times with and without OlivIA.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

The Times of Interest (TOI) for eye-tracking data analyses were defined between the moment OlivIA presents the solutions and the moment the flight plan is accepted by the Pilot. The data is processed to determine the number of times the fixations are placed on the Areas of Interest (AOI):

- Global (OlivIA zone)
- Left (map area)
- Right (key information area)
- Right-Up (high-level triangle area)
- Right-Down (low-level performance key area)

Figure 33. Layout of the Areas of Interest (AOI) in OlivIA interface.

A thematic analysis structured around major themes that emerged from users' feedback was employed to process the qualitative data collected in the interviews: decision-making processes, usability, trust dynamics, explainability/transparency, and areas for improvement.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

3.4. TRL Overview: update

Table 13. TRL progress in UC2 for the OlivIA components.

Component: OlivIA assistant integrated in FlytX		
TRL	**Month**	**Activity to reach selected level**
1	1	COMBI Technology component validated at TRL3 for Mission Management in Defence context with Human-in-the-loop simulations.
2	6	Concept definition of intelligent assistant for regional segment
3	16	VAL1 with an exploratory study of 3 EASA Human-AI levels: Assistive, cooperative and collaborative
4	22	Development of one concept of assistance Iterative evaluation of the concept with 3 test pilots to increase representativity of the assistance and simulation scenarios.
5	25	Integration of OlivIA with Flight Management System to increase “integrity” level of assistance Runs with 10 pilots in a representative system (FlytX of THALES) to validate the concept (VAL 2)
6	-	Technology demonstration not possible because we have identified the need of extended architecture OLIVIA + ATC + AOCC to ensure landing clearance and integration with Airline Policy

3.5. VAL 2 Results

3.5.1. Results Overview

Table 14. High-level results in relation to the EASA Validation objectives.

High-level result	**Brief description**	**Related validation objective (D6.3)**
The proposed assistance concept based on high-level intentions communications **presents potential benefits**	Efficient integration of relevant data and complementary assessments affects positively the decision-making processes. (interview analyses); Favourable usability	UC2-OBJ-01: COMBI enables effective and efficient high-level intentions communication in the team.

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

to decision-making effectivity and overall efficiency in the flight deck	assessment (CSUQ); Easy communication: solutions request in two clicks. (intention). (Interview analyses) Objective enhancements in efficiency and effectiveness could not be observed due to limited training and familiarization with the system. (Interview analyses) Workload and situation awareness weren't impacted, even with these training and familiarization limitations (NASA-TLX, SART)	
The HAT design methodology **supported effectivity assessments** in VAL 2	Formulation of relevant hypotheses and system requirements to structure effectiveness assessments in VAL2. (researchers' assessment) Safety assessments out of scope of VAL2. (Assessed with HAZOP in WP7)	UC2-OBJ-03: HAT design methodology supports HAT safety and effectivity assessments
The Pilot **can communicate** bi-directionally with the system through operational intentions. (HAIKU_UC2_HAT_2) OlivIA interface **allows** the pilot to recalibrate operational intentions during operation. (HAIKU_UC2_HAT_9)	Solutions and assessments are coherent with Pilot's intentions. Improvements on the intentions representation and solutions presentation were provided by some pilots (interviews analyses). Pilots can change their intentions during operation. Half of the pilots changed intentions and requested a new solution at least one time (HMI design, objective measure) Positive usability evaluation (CSUQ SCORE ≈ 5)	HF-01-02: IA must account for pilot operational intentions. HF-02: IA must incorporate bidirectional information sharing (to/from the PF and PM (if appropriate)) about reroute / alternate airport recommendations, so as to match (pre-flight loaded) operational intentions and technical parameters.
OlivIA **provides** the solutions in less than 1 minute after request. (HAIKU_UC2_HAT_3)	Computation time: 30-45s Met requirement but affected significantly the total DM time. (objective measure)	HAIKU_UC2_HLR_4: The crew decision must be implementable, considering time constraints.
OlivIA **partially provides** relevant, clear, and valid explanations to the Pilot about each proposed solution. (HAIKU_UC2_HAT_4)	Pilots stated that it is important to have better explanations on the system capabilities and behaviour (mostly during training). Positive usability evaluation, positive evaluation of the information quality aspect (CSUQ SCORE ≈ 5, INFOQUAL ≈ 5) Trust analyses indicate that users would be confident in using a certified system with	EXP-10: IA must be able to explain both its reroute- and airport diversion recommendations using progressive CLT levels; EXP-11: IA must provide explanations in a clear and unambiguous form; EXP-12: IA must provide explanations relevant to the assessment of the

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

	this type of information and behaviour if they were properly trained to use it. Clear meaning of colour codes could contribute to understanding while using the system.	appropriateness regarding expected decision / action; EXP-13: IA must provide explanations at the level of operational intentions along with each proposed solution; EXP-17: IA must provide timely, relevant and valid explanations to the Pilot about each proposed solution.
Pilots' knowledge level on OlivIA capabilities and typical behaviour **was not adequate** (HAIKU_UC2_HAT_5)	Pilots would need more effective training to understand the system capabilities and typical behaviour (interview analyses)	HAIKU_UC2_HLR_2: The crew must minimise the impact of flight plan changes in company operation; HAIKU_UC2_HLR_3: The crew decision quality must be better than that of a crew with no assistance in the same situation; HAIKU_UC2_HLR_4: The crew decision must be implementable, considering time constraints.
OlivIA interface **provides** the Pilot the ability to navigate, for each proposed solution, through different layers of explanations. (HAIKU_UC2_HAT_6)	The different layers of explanations were used by most of the pilots (observation) More layers of on-demand explanations and the ability to customize their presentation were also suggested. (interview analyses)	EXP-16: IA interface must provide the Pilot the ability to navigate, for each proposed solution, through three different layers of explanations.
OlivIA ***partially*** **accounts** for weather and destination states' impact in diversion and flight trajectory proposals. (HAIKU_UC2_HAT_7)	Relevant KPI's are considered in the solutions, but the weather evolution is not. (model design, interview analyses) In 3 of the 20 runs with OlivIA, pilots used the proposed solution and made adjustments to compensate for that.	HF-01-01: IA must account for weather impact upon flight trajectory.
OlivIA **can load** a Flight Plan approved by the Pilot to the FMS. (HAIKU_UC2_HAT_11)	Pilots are confident that the system won't act without their approval (interviews analyses and HMI design)	HF-04: IA recommendations must be approved by the user

Please refer to the Annexes of this deliverable for all results' details.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

3.5.2. Discussion

The findings reveal several key insights about the OlivIA system's impact on aviation decision-making processes and its operational integration:

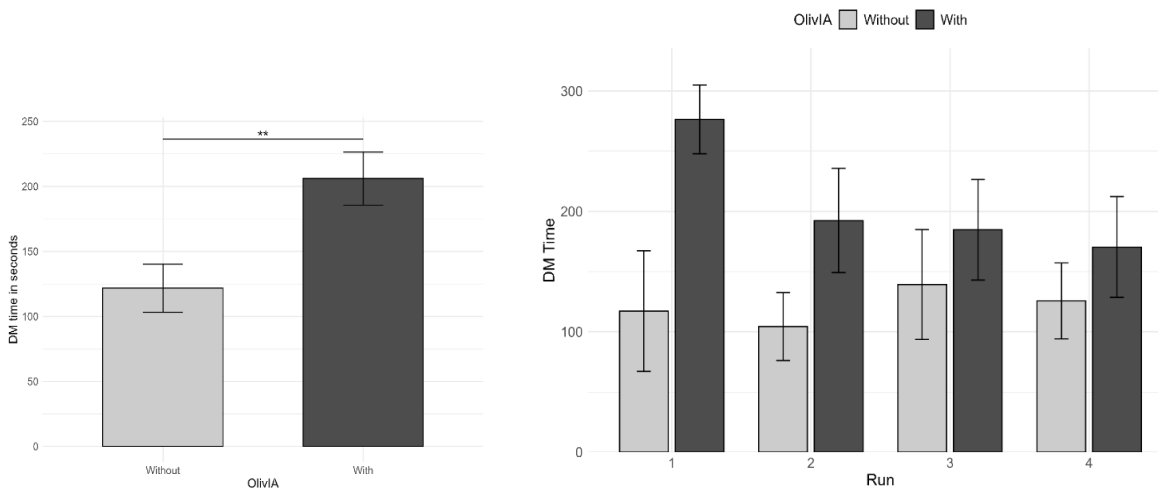
System Impact on Decision-Making Processes

In contrast to systems that automate decision-making, OlivIA has demonstrated to play a complementary role to pilot cognition, rather than a replacement to the pilot decision-making role. The information checking behaviour observed when using OlivIA successfully argus for that. All pilots, except one, requested details on OlivIA's solutions. The interviews analyses indicate diverse perspectives from the pilot. Most of them reported to have conceptualised solutions while OlivIA was computing, to compare the proposed solutions with their initial idea, using the system for validation. Some of them reported that the system influenced their choices, by showing information and solutions that they hadn't considered previously. This aligns with its design philosophy of validating options, challenging assumptions, and identifying new opportunities while preserving pilots' ultimate authority. The absence of significant changes in decision-making patterns indicates that pilots maintained their standard evaluation processes while using OlivIA as an information resource.

The interviews also revealed concerns about potential over-reliance on the system, and with the possibility of focusing too much on exploring the system´s solutions. But most of the pilots would have preferred to choose from more than only 3 solutions, often with 2 of them to the same airport through different trajectories. This tension between assistance and potential distraction requires careful balance in system design and the way to use it considering the situation, whether the circumstances allow for a very thoughtful decision, or demands rapid management.

Decision-making times were significantly impacted with the use of OlivIA (with OlivIA: $M = 206$, $SD = 91$; without OlivIA: $M = 122$, $SD = 19$; $W = 91$, $p = .003$, $r = 0.43$), illustrated in Figure 34 (a). OlivIA's computation time (30s-45s, during which pilots were unable to interact with the interface), may have a strong contribution to this significant difference. The lack of familiarity with this type of assistance may be another contributor. The decision-times along the runs (recalling that pilots had OlivIA either in Run 1 and 3, or in Run 2 and 4) shown in Figure 34 (b) suggested there could be a reasonable learning curve between runs 1-3, but the combined effect between assistance and order of the run was not significant ($p = .440$).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

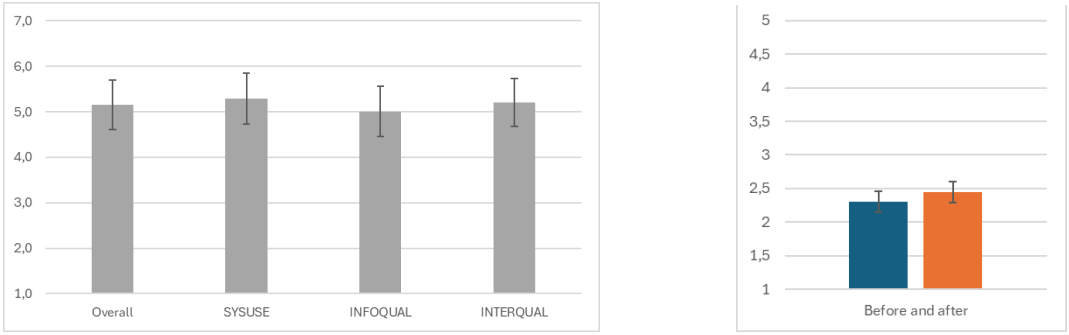


(a) (b)

Figure 34. Decision-making time with and without OlivIA: overall (a) and across runs (b)

Trust Dynamics and System Adoption

The discrepancy between high technical competence ratings (M ≈ 5 on CSUQ aspects; Figure 35 (a)) and moderate trust scores (before trials: M = 2.3, SD = 1.2; after trials: M = 2.5, SD = 1.2; Figure 35 (b)) reflects aviation professionals' characteristic prudence regarding automation. This conditional trust paradigm emerged consistently in pilot feedback, emphasizing the need for greater system transparency and dedicated training protocols. These findings mirror broader automation adoption challenges in aviation, where proven reliability and understandable system logic precede full operator confidence and the successful adoption of the system. The results of the AIDUA questionnaire are favourable, showing that the pilots are willing to use AI and perceive them as easy to operate and beneficial for performance (Figure 36).



(a) (b)

Figure 35. (a) CSUQ (usability) results; (b) Trust in AI results

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

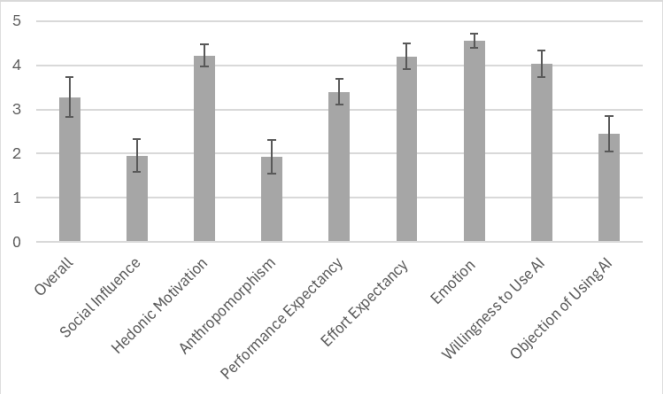


Figure 36. AIDUA results

Workload and Situation Awareness

No significant impact was found on either workload measures (with OlivIA: M = 37.3, SD = 17.63; without OlivIA: M = 37.46, SD = 18.77; p = 0.98; Figure 37) or SART scores (with OlivIA: M = 29.0, SD = 5.9; without OlivIA: M = 28.0, SD = 7.4; p = 0.07; Figure 38), which represents a positive outcome for aviation human factors integration. In safety-critical environments, new systems must avoid cognitive disruption, and OlivIA's seamless integration suggests successful implementation of this design principle. The maintained situation awareness during use indicates the system provided information without overwhelming pilots - a crucial balance in flight operations.

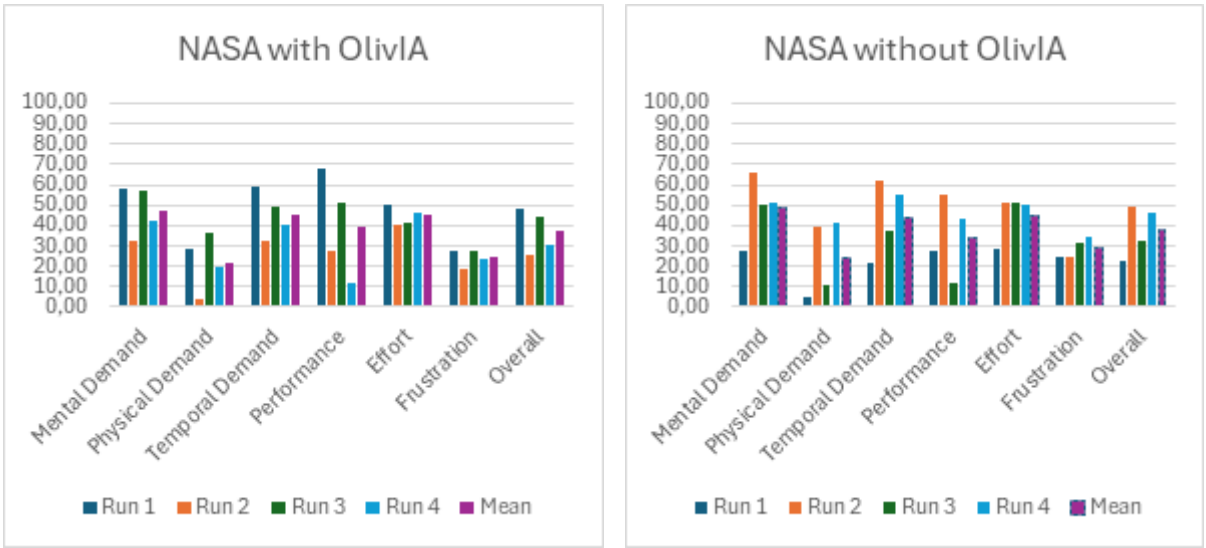


Figure 37. Comparison of scores by run and condition

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

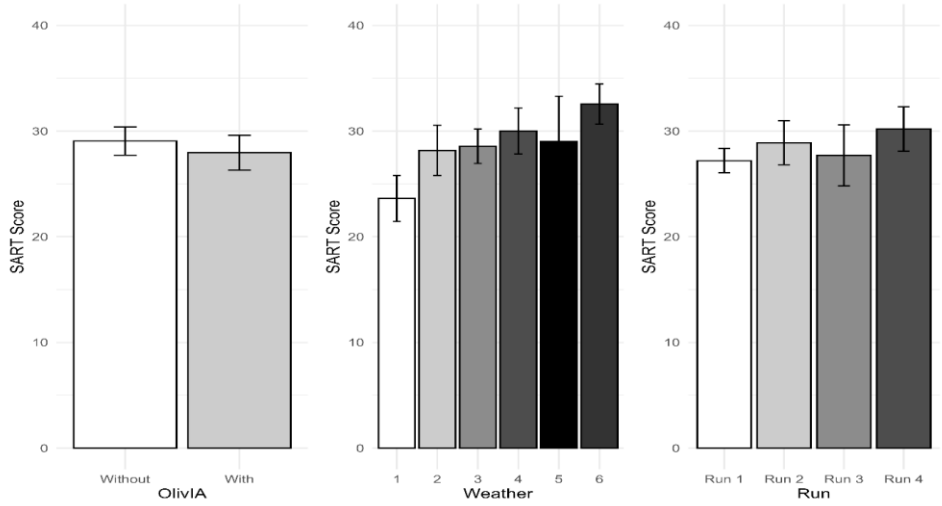


Figure 38. SART results by run, condition, and weather

Usability

The CSUQ, AIDUA and Transparency questionnaires and interviews reveal that OlivIA was found to be both useful and easy to use, although there was some room for improvement. The interactions involved in using OlivIA were judged to be a user-friendly experience.

Operational Implications

The system demonstrated particular value in two domains:

1. *Experience compensation*: Less experienced crews benefited from expanded option consideration beyond familiar airports.
2. *High-workload scenarios*: The information aggregation capability proved valuable during diversion situations. Additionally, pilots identified potential training applications, suggesting OlivIA could serve dual purposes in both operational and instructional contexts.

These findings position OlivIA as a system intended to support decision-making rather than to make decisions itself. This is consistent with modern aviation philosophy, which maintains human operators in command while providing enhanced information integration capabilities. The identified improvement areas provide clear development pathways for creating more transparent, adaptable systems that can earn pilot trust while preserving cognitive workflow integrity. Subsequent iterations of the system should place particular emphasis on the development of customisable interfaces and enhanced explanation features. These enhancements are designed to address the diverse operational contexts and user preferences that have been revealed by this VAL2.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Limitations

The experiment was conducted on a simulator with a small sample size of pilots (N=10). The sample comprised pilots of both sexes, with a range of backgrounds and levels of experience, from novice to experienced (M age = 47.25 SD = 8.3, mean number of flight hours = 8135, SD = 5611). In accordance with the principle of reproducibility, the flight scenario remained constant; however, the meteorological conditions and airport KPIs were subject to variation. To address these limitations, subsequent user tests should employ larger sample of pilots and develop scenarios that vary stress levels and complexity. A limitation in setting up the experiment was not to compromise flight safety in the design of the flight scenario, as the OlivIA solution should not be focused on the issue of safety. So OlivIA was tested only in non-critical situations.

3.6. UC2 VAL 2 Conclusions

The VAL2 of UC2 focused on assessing OlivIA's integration within operational flight crews and its impact on Human-AI Teaming. This evaluation provided insights into trust dynamics, cognitive workload, situational awareness, and decision-making quality under complex conditions. The findings highlight key factors contributing to user acceptance, as well as areas for improvement in interface design and system explainability.

By simplifying the communication using high-level intentions, Olivia enhances shared situational awareness and facilitates decision-making process. The results confirm OlivIA's potential as a cognitive assistant and reinforce the relevance of the HAT design methodology for supporting safety and effectiveness in critical environments.

3.6.1. UC2 Research Questions

RQ1: To what extent does the integration of relevant data and complementary assessments enhance the team's decision-making processes?
The integration of OlivIA into pilot operations enhances decision-making by expanding the scope of options considered, particularly in situations where crews are less experienced or facing time-critical scenarios. The system does not override human judgment but supports it by aggregating relevant data and proposing validated alternatives, thereby improving the quality and robustness of team decisions.

RQ2: How does OlivIA facilitate the communication of high-level intentions within the team?
OlivIA enables intention-sharing by aligning its output with pilot objectives and maintaining coherency with operational decision flows. It supports the team by offering contextualized suggestions that reflect the intent of the mission, contributing to a shared mental model and coherent action planning without disrupting established communication structures.

RQ3: What is the user perception of OlivIA's usability in an operational context?
Pilot feedback indicates that OlivIA is well-received as a non-intrusive, complementary system. It is not perceived as a threat or replacement, and its seamless integration into the cockpit environment

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

underscores its usability. Additionally, its perceived value in both operational and training contexts points to a positive user experience and high acceptance.

RQ4: What effect does OlivIA have on perceived workload and situational awareness? OlivIA has a neutral to positive impact on perceived workload, with no reported cognitive disruption. Pilots maintain stable situational awareness, and in high workload situations—such as diversions—OlivIA's information aggregation and solutions support enhance the team's ability to manage complexity effectively.

RQ5: Does the HAT design methodology enable effectivity assessments? The HAT design methodology provides a framework to assess effectiveness by focusing on collaboration, transparency, and cognitive alignment between human and AI agents. OlivIA exemplifies these principles, demonstrating how HAT can guide the development of systems that enhance team cognition and trust while preserving human authority.

3.6.2. HAIKU High-level Research Questions

HAIKU Q1: What are the common recommendations concerning Human-AI teaming for the different AI aviation applications?

From a UC2 perspective, in OlivIA, an emphasis was given to maintain human authority, enhance transparency, and **align** the **design with the pilot cognitive processes**. **Trust-building, intention alignment, explainability**, and **interface clarity** are **essential to effective Human-AI teaming in OlivIA**. Rather than introducing new workflows, OlivIA focuses on facilitating the communication and augmenting the pilot's decision-making process, particularly in complex or unfamiliar situations and under pressure. This approach can be generalized for the design of assistance where human judgement is a key cornerstone of the decision and action loop.

HAIKU Q2: What does it mean for AI to be explainable? Is explainability a necessary precondition of trustworthiness?

While explainability primarily concerns the ability of an AI system to communicate its reasoning behind recommendations or actions clearly, UC2 results showed that it is an important factor in establishing conditional trust, particularly in safety-critical sectors such as aviation. **Latent explainability**, developed through training and interaction, is **essential for adequate system understanding** and **appropriate use**. Therefore, **operational explainability** is a **necessary precondition** for trustworthiness and thus fundamental to pilots' willingness **to adopt and effectively integrate AI tools**.

HAIKU Q3: How to train AI to assist humans in safety critical tasks when training data are insufficient?

When training data is insufficient, AI systems intended for safety-critical tasks should be **developed using hybrid approaches that combine data-driven methods with expert knowledge, simulation-**

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

based environments, and human-in-the-loop training. In Olivia we have used Genetic Fuzzy Trees to combine data-driven and expert knowledge approaches.

Involving users during iterative development phases helps refine system behaviour and **ensures alignment with operational realities**. **Emphasis** should be placed **on transparency and progressive explainability to support user trust** despite limited initial datasets.

3.6.3. Research Recommendations

Results of UC2-VAL2 motivate some future research recommendations about human-AI teaming:

1 - Do not isolate users, aviation is a collaborative job: when designing an Intelligent Assistant for a complex system like aviation, effectively managing its impact and maximizing its benefits requires considering stakeholders, understanding the network of collaboration.

2 - Adopt an iterative design approach from the concept stage: involve end-users and other relevant stakeholders from the outset to determine the most desired and acceptable level of support (assistive, cooperative, or collaborative) before development begins. Base your approach on user needs, considering varying expertise levels and workload scenarios. Understanding how each user profile would benefit in the different contexts is essential to effectively define the solution's scope and objectives.

3 - Use XAI to ensure end-users adequately understand the system and feel in control: design the HMI to balance clarity with completeness, avoiding information overload. Implement a multi-layered XAI approach to deliver the appropriate explanation level of abstraction at each stage of the decision-making process.

4 - Establish effective Human-IA communication: Effective communication between humans and AI fosters true teamwork, enabling both to recognize, interpret, respond to, and anticipate dynamic flight conditions. The ability to negotiate and align decision-making processes supports coordinated operations, ultimately enhancing safety and efficiency of operations.

5 - Always strive to improve Situational Awareness: regardless of the specific tasks supported by the Intelligent Assistant, improving situational awareness should remain a key objective. OlivIA was recognized for its ability to do so by presenting alternative solutions within an integrated information environment, rather than isolated options.

Improvement Areas

Future iterations of the OlivIA system should prioritize:

- **Enhanced Explainability:** Improve transparency by explaining recommendations in a familiar, collegial manner (e.g., "Frankfurt prioritized due to...") using a user cantered language (domain specific).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- **Interface Design**: Refine visual coding and information hierarchy.
- **Interface adaptability:** Divergent preferences between visualization formats suggest implementing context-aware displays (emergency vs. normal operations) and user-selectable presentation modes to accommodate different decision-making styles.
- **Data Transparency:** Provide detailed information about data updates (e.g., with visible timestamps).
- **Comprehensive Training Programs:** Improve system interpretability and predictability through training, facilitating effective adoption.
- **Balanced Optimization:** Consider multiple operational priorities:
    - **Critical situations:** Users prefer quick, pre-validated options.
    - **Non-Urgent Scenarios:** Exploratory interaction is preferred.

As summarized by one participant: "*The best way could be thinking many things, not only doing, but thinking as well,*" suggesting that the ideal system augments pilot situation awareness rather than replaces pilot activities.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

4. Use Case #3 – Urban Air Mobility

4.1. Deviation from Validation Plan (D6.3)

VAL2 was performed according to the validation plan described in D6.3 with the following deviations:

- Eye tracking was intended to be used to monitor participants' visual attention and provide attention guidance to areas on the situation display that require the UAM Coordinator's attention. We could, however, not implement the attention guidance algorithm as hoped, and reverted to implementing scripted attention guidance mechanisms in the corresponding scenario. Hence HF-03 was not validated. Eye tracking was used as a dependent measure, to analyse how participants' visual attention was influenced by the attention guidance mechanism, and scenario overall.
- The speech and natural language capabilities of DUC were not implemented for VAL2, and hence HF-10, HF-11, HF-12, HF-13, and HF-14 were not validated. We were considering exploring natural language capabilities using a Wizard of Oz interaction but decided not to because of lacking realism and increased simulation complexity.
- The Digital journal/logbook system, and communication log was not implemented on Screen 3 due to complexities in implementing the functionality with the simulator. As such HF-09 was not validated.
- No separate communication system was implemented due to technical constraints at the experimental site. Instead, both incoming and outgoing calls to external stakeholders were supported by implementing a "call" dialogue window appearing on the situation display. The participant was instructed to speak out loud if he/she wanted to contact a stakeholder.

4.2. Validation Objectives

The validation objectives in UC3 were defined to address four key human factors challenges for HAT: situation awareness, transparency, decision-making, and bi-directional communication. The DUC and UAM Coordinator Working Position concepts were designed to explore a range of the required functionalities and capabilities to support effective collaboration between DUC and the UAM Coordinator.

Situation awareness: The DUC must support the UAM Coordinator's situation awareness by continuously monitoring U-space and traffic operations, analysing real-time data to detect trends, anomalies, and potential conflicts, while anticipating future traffic patterns. It should also support the UAM Coordinator by providing real-time updates, responding to information needs, and directing attention to critical events.

Transparency: DUC must provide explanations to support recommendations made to the UAM Coordinator on how to solve problems. DUC should highlight the relevance of its explanations to

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

specific decisions or actions, adjusting the level of detail based on the context, task, and the UAM Coordinator's expertise. Additionally, DUC should clearly communicate how its outputs are generated and explain its internal processes.

Bi-directional communication: DUC should effectively recognize the UAM Coordinator's instructions and intentions, using natural language across voice, text, and graphics to ensure clarity. It must adapt its communication style to the user's preferences, state, and context, and avoid interrupting when the UAM Coordinator is engaged elsewhere.

Decision making: DUC should resolve conflicts, recommend solutions, and make decisions within its scope of authority. It must support collaborative problem-solving using a checklist approach when appropriate, such as during emergency situations.

Building on these HAT objectives, five research questions were defined for VAL2.

- Research question 1 (investigated overall teaming across all scenarios): How do air traffic controllers perceive the impact of collaborating with the DUC in U-space traffic management operations, particularly regarding its influence on their situational awareness, communication practices, perceived level of control, and allocation of task responsibilities?

- Research question 2 (scenario: Medical Emergency): How does situated vs. abstract representation influence operators' workload, situational awareness, decision making challenge and feeling of control?
- Research question 3 (scenario: Fire Emergency): How does the mode of explanation (storytelling vs. textual) influence operator attitudes, comprehension, and decision in response to decision recommendations made by DUC? (Note: answered in Deliverable 5.2).
- Research question 4 (scenario: Fire Emergency): How does the application of different CLT structured explanations through interactive visual storytelling impact the perceived transparency of DUC? (Note: answered in Deliverable 5.2).
- Research question 5 (scenario: Link-loss Situation): How does situation-switching attention guidance provided by DUC influence operators' overall (main monitoring) situational awareness (comparing attention guidance vs no attention guidance)?

4.3. VAL2 Activities and Methods

VAL2 consisted of one human-in-the-loop simulation with air traffic controllers as participants. The parts of the IA tested in VAL2 were the working station, DUC HMI elements, and selected HAT mechanisms consisting of how DUC presented information, storytelling explanations, and attention guidance.

UC3 followed an approach where a HAT theoretical model was first defined. Thereafter, research questions were formed. The EASA HAT requirements were then analysed in relation to the research questions to determine coverage. Measures including interviews, questionnaires, eye tracking and

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

software logs were then defined to answer the research questions, and indirectly the EASA HAT requirements to provide insights about DUC and the UAM working position.

As part of VAL2, the following activities were carried out:

- the HAT concept was further developed to address the four key human factors challenges.
- the UAM Coordinator workstation concept was realised in a simulator environment.
- three traffic scenarios were developed, each addressing one or two different research questions. The scenarios were:
    - Medical emergency (research question 2): comparing abstracted vs situated information,
    - Fire emergency (research question 3 and 4): comparing storytelling explainer vs text explainer,
    - Link-loss situation (research question 5): comparing attention guidance vs no attention guidance.
- five separate test runs prior to the VAL2 human-in-the-loop simulation (see UC3 Annex for a description of these); and
- the VAL2 human-in-the-loop simulation was conducted.

The VAL2 human-in-the-loop simulation consisted of a within participant design, involving nine air traffic controllers as participants. The experiment lasted three hours per participant.

4.3.1. Participants

Nine experienced ATCOs provided by LFV Sweden participated in VAL2. There were seven males and two females. Age ranged from 26 to 57 years (M = 41,2 years, SD = 11,7 years). All ATCOs were working at Malmö ATCC. Experience ranged from 1 year and 11 months to 30 years and 5 months (M = 15,6 years, SD = 10,7 years). All were currently working as en-route controllers. Four had experiences from also working the Terminal Manoeuvring Area (TMA) and two had experience from TMA, Tower and Procedural operations. While all participants had a basic understanding of UAM in general (courses provided internally within LFV), they had no previous knowledge or experience of HAIKU, UAM, DUC, or the UAM Coordinator.

4.3.2. Simulator/Apparatus

The simulator setup consisted of a proof-of-concept prototype of the UAM Coordinator working position. The participant was seated in front of a table, on which there were three large screens. The screens were connected to a single laptop, from where scenarios were run by an experimenter. The IA DUC was integrated with all screens using a combination of scripted events and Wizard of Oz interventions.

This working position comprised the following screens and system:

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- **The left screen presented a knowledge library with checklists** for different non-normal situations (medical emergency, fire emergency, and a link loss situation). The checklists were created using Microsoft Power Point. DUC actions on the checklist (i.e., checking of action items and providing facts related to the situation) were pre-scripted using animations to occur at specific time intervals after the participant had activated the checklist (by clicking on it).
- **The middle screen was the situation displa**y providing a map view of the Stockholm U-space. The situation display was created using the high fidelity UTM City Drone Simulator. The DUC prototype was implemented using a combination of scripted DUC HMI overlays on the situation display and Wizard of Oz (e.g., triggering events, communications, interactions). Traffic flows and city events were scripted using the scenario builder.
- **The right screen contained the Storytelling explainer** providing video narrated explanations supporting DUC's recommendations for how to solve the fire emergency scenario. The explanations were triggered by the emergency fire event in the corresponding scenario, at which time a video was shown on the right screen.

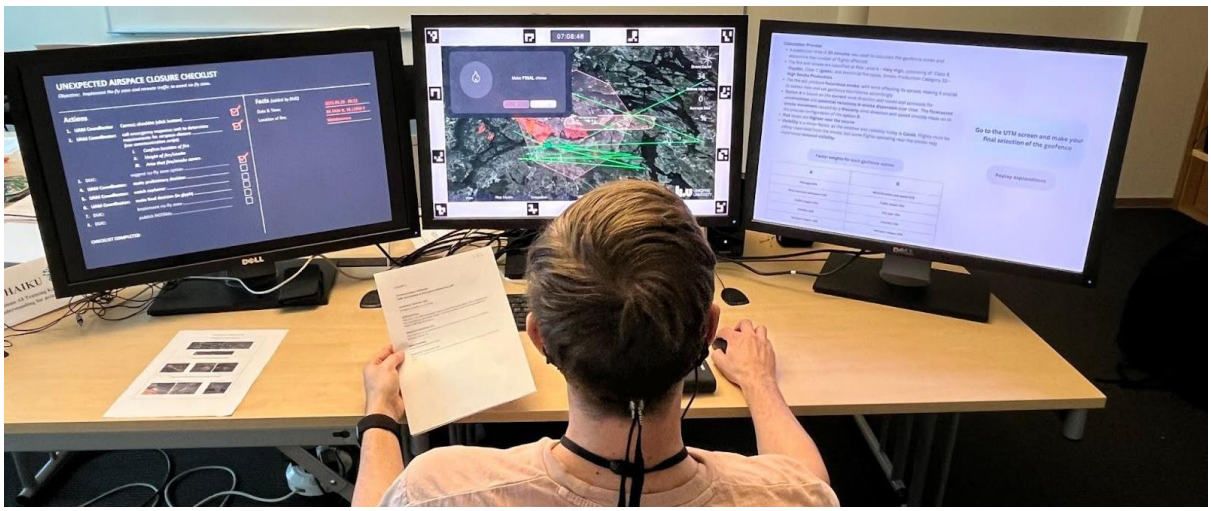


Figure 39. UC3 Simulator setup in VAL2 representing the UAM Coordinator Working Position.

4.3.3. Scenarios

VAL2 had three scenarios that each addressed a non-normal situation requiring the collaboration between the DUC and the UAM Coordinator to find a solution. Detailed scenario descriptions are provided in the UC3 Annex.

1. **Medical emergency:** The medical emergency scenario contained two routine coordination events and a medical emergency. Each event required a potential reroute of a flight. DUC

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

presented the rerouting options to the UAM Coordinator, who was asked to decide on which reroute option to implement.

2. **Fire emergency:** The fire emergency scenario consisted of an explosive and rapid fire emerging in an oil storage tank requiring the closure of a nearby vertiport. Because of the fire and smoke, DUC presents the UAM Coordinator with two options for how to close off the U-space airspace. DUC directs the UAM Coordinator's attention to the Storytelling explainer, where DUC provides explanations for why the two options are suggested. The UAM Coordinator is asked to decide on which no-fly zone option to implement.

3. **Link Loss situation:** The link loss situation starts as a conformance monitoring issue, with an air taxi deviating from its route (U-plan). Shortly thereafter, the air taxi experiences a link loss, and its symbol disappears from the situation display. DUC notifies the UAM Coordinator in a dialogue window that there is a link-loss situation. This raises the interest level of the UAM Coordinator as the DUC is unable to solve the situation on its own. DUC opens the link-loss emergency checklist, which directs the UAM Coordinator to contact the Joint Rescue Coordination Center (JRCC) as the latest position of the air taxi is within their jurisdiction. The UAM Coordinator's task is to inform JRCC of the situation, which will trigger them establishing a search and rescue area.

In all scenarios, both actors (UAM Coordinator and DUC) share the same goal: to develop an understanding of the situation (i.e., medical emergency, fire emergency, and link-loss) and find a solution (i.e., diversion hospital for medical emergency, U-space airspace closure for fire emergency, and search and rescue for link-loss).

In general, the UAM Coordinator's task is to monitor the U-space and supervise DUC's actions. The role is largely reactive, meaning that the participant is expected to be rather passive when operations are normal. In unforeseen and emergency situations, the UAM Coordinator's role and task is to quickly develop situation awareness and find a solution to the problem. This may require searching for information on the situation display, carrying out action items according to the checklist (e.g., locate information, coordinate with stakeholders, make decisions), and coordinate with relevant stakeholders by receiving and initiating phone calls. The UAM Coordinator also makes critical decisions (e.g., diversions and airspace closures).

The DUC's role and responsibility is to monitor the U-space and its actors, notably conformance monitoring of traffic. In emergency situations, DUC's role and tasks is to support building the UAM Coordinator's situation awareness, propose checklists, provide explanations supporting action options, provide facts (e.g., about specific flights). When the situation has been identified, DUC proposes the corresponding emergency checklist that specifies tasks for DUC and tasks for the UAM Coordinator. DUC retrieves information relevant to the situation (e.g., a specific flight, U-space, weather). DUC can use this information to calculate and present options (e.g., diversion options, airspace closure). When the UAM Coordinator has decided on how to solve the situation, the DUC implements the decision (e.g., by rerouting flights, activating new U-plans, establishing a no-fly zone).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

4.3.4. Independent Variables

The three scenarios each had one independent variable with two levels. More information about the variables, and what they looked like from the participants point of view is provided in the Annex.

1. **Medical emergency: situated vs abstracted information.**
   The situated/abstracted conditions explored two variants of how DUC presented information (the rerouting options) to the UAM Coordinator (information content was the same in both conditions). In the abstracted condition, spatial information relevant to the reroute options are provided in the dialog window (e.g., green- and yellow-coloured dots next to each option. In the situated condition, spatial information relevant to the reroute options are overlaid on the map position for which it refers.
2. **Fire emergency: storytelling vs text explanation format.**
   The storytelling/text condition explored two variants of how DUC presented explanations to the UAM Coordinator (information content was the same in both conditions). In the storytelling condition, explanations were presented in a visual narrated video format that the participant watched. In the text condition, explanations were presented in a text format that the participant read. More detailed information about the two conditions is provided in D5.2
3. **Link-loss situation: attention guidance vs no attention guidance.**
   The attention guidance/no attention guidance conditions compare the “take me there” attention guidance mechanism with no attention guidance. In the attention guidance condition, DUC would accompany notifications (dialogue windows) involving an object with a spatial dependence (e.g., the flight experiencing a conformance issue) with a “take me there” prompt. If the UAM Coordinator clicked on the prompt, DUC would adjust the map view to focus on the object of interest (the map view jumps directly to the situation in question). In the no attention guidance condition, DUC did not offer this assistance.

4.3.5. Procedure

The table below outlines the validation procedure as experienced by participants. Detailed information can be found in the Annex for the following:

- Introduction presentation (presenter notes supporting presentation)
- Informed consent
- Demographic questionnaire
- Simulator walkthrough and training document
- Eye-tracking introduction

The validation started with participants receiving an introduction to the validation. Focus was directed to explaining the context (U-space traffic management), UAM Coordinator working position, DUC, and participant’s role and responsibilities acting as the UAM Coordinator. Prior to starting the simulation, participants filled out a consent form and a demographics questionnaire. Participants were then introduced to the eye-tracking equipment and went through an initial calibration. The Tobii Glasses 3

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

wearable eye tracker was used. 20 minutes was dedicated to training in position, following a simulator guided walkthrough instruction document. At the end of this training, participants were asked to demonstrate their knowledge by answering several questions about what was shown on the situation display. Participants also received time to explore the interface and work on their own without any guidance.

Table 15. UC3 VAL2 procedures

Time	Length	Description
09:00	10'	Introduction: Greeting and presentation of researchers Introduction presentation and videos (video and PowerPoint) Consent form Demographics questionnaire
09:10	20'	Eye Tracking introduction/calibration and Simulator Walkthrough and Training
09:30	20'	Session #1 - Play scenario - Questionnaire and Interview
09:50	20'	Session #2 - Play scenario - Questionnaire and Interview
10:10	20'	Session #3 - Play scenario - Questionnaire and Interview
10:30	15'	Break
10:45	5'	Eye Tracking calibration
10:50	20'	Session #4 - Play scenario - Questionnaire and Interview
11:10	20'	Session #5 - Play scenario - Questionnaire and Interview
11:30	20'	Session #6 - Play scenario - Questionnaire and Interview
11:50	10'	Debriefing
12:00		Closing

The validation started with participants receiving an introduction to the validation. Focus was directed to explaining the context (U-space traffic management), UAM Coordinator working position, DUC, and participant's role and responsibilities acting as the UAM Coordinator. Prior to starting the simulation, participants filled out a consent form and a demographics questionnaire. Participants were then introduced to the eye-tracking equipment and went through an initial calibration. The Tobii Glasses 3 wearable eye tracker was used. 20 minutes was dedicated to training in position, following a simulator

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

guided walkthrough instruction document. At the end of this training, participants were asked to demonstrate their knowledge by answering several questions about what was shown on the situation display. Participants also received time to explore the interface and work on their own without any guidance.

The simulation consisted of six sessions, 10-15 minutes in duration. There were three different scenarios, each played twice in different versions. Each session consisted of one scenario version. The order of scenarios and versions was varied according to an unbalanced Latin-square matrix (a balanced Latin-square requires six participants, and we had nine). For this purpose, individual playlists were created for each participant. After each session, participants were asked to answer a questionnaire and debrief interview. The questionnaire and debrief was specific for each scenario version played. For instance, for the Scenario 1 version "situated," the questionnaire "Situated Questionnaire" was used. For the Scenario 1 version "abstracted," the questionnaire "Abstracted Questionnaire" was used. For Scenario 2, the "Storytelling Questionnaire" and "Text Questionnaire" were used depending on the scenario version. For Scenario 3, the "Attention Guidance Questionnaire" and "No Attention Guidance Questionnaire" were used depending on the scenario version.

At the end of the validation, the "HAT Questionnaire" and "Social Acceptance Questionnaire" were completed, together with a final debriefing interview. All questionnaires were implemented and administered digitally using Google Docs.

Three experimenters were involved in running the simulation. One experimenter introduced the participant to the use case, the U-space and DUC concept of operation, the working station, and the roles and responsibilities of the UAM Coordinator, whom the participant was instructed to act as. This experimenter also administered the consent form and demographic questionnaire. One experimenter was responsible for the eye tracking equipment, running the simulator and selecting the correct simulator session run for the correct participant according to a playlist. This experimenter also acted on behalf of the DUC when needed (i.e., Wizard of Oz.). One experimenter acted as the training instructor, providing guided instructions according to a training walkthrough protocol. Different experimenters were responsible for post-session questionnaires and debriefs.

4.3.6. Data Collection Tools

The table below shows which data collection tools were used for the validation. Note that the Situated/abstract, Storytelling/text, and Attention guidance/no attention guidance questionnaires and debrief interviews were used twice: once for each scenario version.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Table 16. Data collection tools used in UC3.

Tool	Objective	Type of Collected Data
Eye-tracker	Behaviour, visual attention	Qualitative
Simulator log	Time, interactions, clicks	Qualitative
Frame Grabbing	Participant views, actions	Qualitative
Photos/videos	Simulator setup, debriefings	Qualitative
Observations	Decisions, behaviour, debriefings	Quantitative, qualitative
Situated/abstract questionnaire (Scenario 1: Medical emergency)	Subjective preferences, perceptions, situation awareness and workload	Quantitative
Situated/abstract debriefing interview (Scenario 1: Medical emergency)	Subjective preferences perceptions, and decision-making rationales	Qualitative
Storytelling/text questionnaire (Scenario 2: Fire emergency)	Subjective preferences and perceptions of explainer format	Quantitative
Storytelling/text debriefing interview (Scenario 2: Fire emergency)	Subjective preferences and perceptions of explainer format	Qualitative
Attention guidance/no attention guidance questionnaire (Scenario 3: Link-loss situation)	Subjective preferences and perceptions for attention guidance	Quantitative
Attention guidance/no attention guidance debriefing interview (Scenario 3: Link-loss situation)	Subjective preferences and perceptions for attention guidance	Qualitative
Societal acceptance questionnaire	Subjective preferences and perceptions regarding acceptance of DUC	Qualitative
Human-AI Teaming questionnaire	Subjective preferences and perceptions regarding collaboration, situation awareness, communication practices, perceived level of control, and allocation of task responsibilities	Quantitative
End of validation debrief	Subjective preferences and perceptions regarding the validation experience	Qualitative

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

4.3.7. Data Analysis

Data collection was conducted through a mixed-method approach combining quantitative and qualitative approaches to explore the data and answer the research questions.

For the **situated/abstract conditions**, quantitative data were collected using a 5-point Likert-scale questionnaire, presented directly after each session, measuring the users' rated situational awareness, workload, challenge of decision-making process and feeling of control. In addition, semi-structured interviews were conducted with all participants to gather qualitative insights into the motivation for the users' preference of abstract or situated format. This scenario Intended to address HF-04; HF-08; HF-09; HF-10; HF-11; HF-12; HF-14; HF-21; HF-24 and HF-25.

For the **storytelling/text conditions**, quantitative data were collected using a Likert-scale questionnaire measuring understanding, trust, user experience (UEQ), Hoffman explainability scale, situational awareness (SART), and confidence and satisfaction in the decision. Structured interviews were conducted with all participants to gather qualitative insights into their preferences of the format, perceived clarity, and contextual suitability of each format. This scenario was intended to address HF-08; EXP-10; EXP-11; EXP-12; EXP-13; EXP-14; EXP-15 and EXP-16.

For the **attention guidance/no attention guidance conditions**, quantitative data was collected using a Likert-scale questionnaire measuring situational awareness, perceived task difficulties and perceived workload. Structured interviews were conducted with all participants to gather qualitative insights into their perceptions of the guidance approach, and its influence on the participants attention. This scenario was intended to address HF-02; HF-03; HF-04; HF-; HF-13; and HF-14.

The eye-tracker data was analysed qualitatively to assess participants' actions, voice communication (e.g., during stakeholder coordination), and visual attention using a replay of their gaze path overlaid on the field of view. This analysis was supported by reviewing the simulator logs, frame grabber data (video recordings of screens), and observation notes. Photographs/videos were used to capture images, video, and audio of participants' work participants' work during scenarios and for assisting transcriptions and post simulation analysis of how they answered the debrief interviews.

All questionnaires were analysed using diagrams, descriptive statistics, and statistical tests. Due to the small sample size and ordinal data type, median and interquartile range were used to measure central tendency and variation in responses. The non-parametric Wilcoxon signed-rank test was used to compare the median difference between paired (related) responses on questionnaire statements.

The HAT Questionnaire and Social Acceptance Questionnaire intended to address HF-28.

Debriefing interviews were transcribed and subjected to thematic analysis to identify recurring patterns. Key themes from the interviews were integrated with the quantitative results to triangulate findings and strengthen the conclusions.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

4.4. TRL Overview: update

UC3 has achieved TRL 2 for the DUC assistant (backend) as its foundational scientific principles (i.e., capabilities related to bidirectional communication, transparency, situation awareness, and decision making and high-level task descriptions building on U-space services) have been identified and studied. A conceptual application of the DUC has been formulated, including functionalities supporting shared situation awareness, enhancing decision-making, automating document analysis, and communication. The DUC HMI components of the UAM Coordinator working position have achieved a TRL 4.

Table 17. TRL progress in UC3 for the HMI components.

Component: Traffic situation display (UTM City), Storytelling explainer system, DUC HAT HMI.		
TRL	**Month**	**Activity to reach selected level**
1	12	Initial ConOps definition. Literature review of UAM/UAS concept of operations, interviews with external stakeholders (EVE Air mobility) to identify futuristic scenarios, participatory design approach (workshops, interviews, and design meetings) involving domain experts (ATM, ATC, UAS, UAM, UTM) to define intelligent assistant concept and futuristic traffic scenarios.
2	18	Prototype specifications definition (through workshops, end-user interviews and design meetings). Participatory design approach (workshops, interviews, and design meetings) involving domain experts (ATM, ATC, UAS, UAM, UTM) to determine and design prototype HMI specifications, including Storytelling explainer, and design prototype working environment.
3	24	Proof of concept prototype version 1 Validation. Laboratory test of working position HMI and traffic scenarios with air traffic controllers. VAL1 with a proof of concept for the IA prototype HMI and UAM Coordinator working position. Laboratory tests with air traffic controllers. Basic IA HAT HMI functionalities. IA backend was scripted and Wizard of Oz simulated. Workshops and interviews with key stakeholders (U-space developers, emergency responders, UAM manufactures) to iterate on concept development, in particular IA HAT HMI specifications, and the traffic situation display. Hazard and Operability Study (HAZOP) to further define IA HAT HMI requirements.
4	31	Prototype Version 2 Validation. Laboratory test of working position HMI and traffic scenarios with air traffic controllers. VAL2 with a second version of the IA prototype HMI and UAM Coordinator working position. Traffic situation display, IA HAT HMI, and Storytelling explainer: test in UAM Coordinator working position simulator using relevant operational scenarios.
5	-	na

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

6	-	na

Table 18. TRL progress in UC3 for the DUC AI components.

Component: DUC backend (AI components)		
TRL	**Month**	**Activity to reach selected level**
1	12	Initial ConOps definition. Literature review of UAM/UAS concept of operations, interviews with external stakeholders (EVE Air mobility) to identify futuristic scenarios, participatory design approach (workshops, interviews, and design meetings) involving domain experts (ATM, ATC, UAS, UAM, UTM) to define intelligent assistant concept and futuristic traffic scenarios.
2	31	Prototype specifications definition in terms of defined functionalities based on foundational scientific principles. VAL2 with an advanced version of the IA prototype concept where certain applications are formulated, specifically communication (dialogue windows, aural alerts, and voice communication), explainability functions (storytelling and text format), attention guidance (“take me there” function), and checklist collaboration.
3	-	na
4	-	na
5	-	na
6	-	na

The main factors inhibiting the achievement of TRL6 for UC3 were:

- Futuristic roles and working environment: As the envisioned working environment and human roles do not yet exist, their concepts must be developed alongside the IA. With no current U-space airspaces or active U-space users (i.e., traffic), a reference U-space environment and traffic scenarios need to be defined to establish relevant reference scenarios. Air traffic controllers, flow managers, and other traffic management operators serve as proxy end users in validation activities, standing in for the currently non-existent UAM Coordinators.
- Concept maturity: The IA concept's broad scope makes it difficult to define all its aspects comprehensively and determine which aspects to focus on in prototype development. Additionally, the target working environment and human role are not yet established, leading to unclear objectives and requirements for the system's functions or the problems it is intended to address.

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

4.5. VAL 2 Results

4.5.1. Results Overview

Questionnaire responses showed that DUC was perceived more as a tool than a team member.

Table 19. High-level results in relation to the EASA Validation objectives.

High-Level Results (HAT Objectives)	**Brief description**	**Related validation objective (D6.3)**
Transparency: Understanding of explanations	For explanations presented in both the storytelling and text format, participants understood why the DUC recommended specific geofence options, indicating that the explanations effectively communicated the rationale behind different alternatives. Participants were satisfied with presented explanations and were able to make a final decision after going through the explainers. CLT1 was found to be not detailed enough, CLT2 was found most suitable for the decision making and CLT3 was found too detailed. However, participants also rated their understanding of recommendation the highest in CLT 3 (Mdn = 5) for both text and video format. Video storytelling explainer improved users' ability to understand DUC's recommendations and their implications for U-space operations, particularly in understanding trade-offs and improving decision-making confidence.	EXP-10, EXP-12, EXP-13, EXP-16
Transparency: Explanation format	In the storytelling format explanations were presented with animations, graphics, narration, icons, text. In the text format, text, bullet points and tables were presented. Most participants were more satisfied with video storytelling format in comparison with text format. The storytelling explanation format was clear, engaging, and perceived as trustworthy, with a strong agreement that the explanations met expectations. Advantages/disadvantages with storytelling format: Storytelling format gave a visual basis that was found to be easy to grasp for the participants. Added narration was also found to be helping in understanding explanations. Although, the lack of function where user can scroll back in the video if needed was perceived as a disadvantage of this format in comparison	EXP-10, EXP-11

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

	to text. Another disadvantage of storytelling format was the length of the video which could be a difficulty in the operational environment. Advantages/disadvantages of text format: Text was mentioned as a format that you can easily scroll back and find needed information. With video storytelling format it is not possible to do.	
Transparency: User adaptation of explanation content	Explanations were ordered progressively according to CLT levels, starting with CLT1 (high level) and progressing to CLT3 (detailed information). After being presented with an explanation for a specific CLT level, the user had the option to ask DUC for more detailed explanations or tell DUC that the explanation was sufficient. Participants chose CLT 2 as the most suitable for emergency situations.	EXP-14, EXP-16
Transparency: Timing of explanations	The explanations were timed to occur during a low task load period, with little other activity. The time required for "consuming" an explanation varied between the two conditions. The "consumption duration" for the storytelling explanation was specific per design and did not vary between individuals. The "consumption duration" for the text explanation varied between individuals as it depended on what parts of the explanation that was read and the time taken to read through the text. The text format gave the user freedom on how to read through the text (e.g., in detail, skim through etc) and to revisit parts if needed.	EXP-15
Situation awareness: Finding geographical position related to events.	DUC's attention guidance function ("take me there") supported participants' situation awareness related to quickly finding geographical positions. Participants stated that it was particularly helpful in providing geographical guidance, given that none of them had local knowledge. Three (out of nine) questionnaire statements had a significant difference, showing that map navigation was rated more helpful compared to no attention guidance.	HF-02, HF-04
Situation Awareness: Map orientation.	When the attention guidance function was used, the overall situational awareness was lower, with multiple participants stating that it made them disoriented. The map-shift made participants lose awareness of map orientation. To regain situational awareness of where the situation was in relation to	HF-02, HF-04

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

	the rest of the city, several participants zoomed out and then zoomed back into the situation again.	
Communication: Method for communication	Overall, DUC was perceived to communicate clearly. But some found DUC to frequently interrupt them. Information provided through dialog windows was perceived clear and very important for situation awareness but could be slightly disturbing (i.e. covering a large part of the screen). Information provided by sound was considered important for situational awareness and only mildly disturbing.	HF-04, HF-08, HF-24, HF-25, HF-28
Decision Making: Problem solving	Questionnaire responses showed that DUC supported participants' problem solving.	HF-02
Communication: HMI design	Observations and questionnaire responses showed that participants overall found the DUC HMI clear and intuitive. However, there was confusion and misunderstandings of how to use the checklists. The training session, prior to the actual simulation, ascertained that all participants had received instructions on using the checklist, how to interact with it, and when to use it. There was also an example situation where they demonstrated using the checklist. Despite this training, several participants did not use the checklist, or partly used the checklist, in the first simulator sessions. Observations indicated that participants using the checklist performed better and experienced less confusion.	HF-28
Communication: Workload	Subjective workload was rated lower when attention guidance was provided (using DUC's "take me there" function).	HF-02
Decision-making: Comfort working with IA	Questionnaire responses showed that participants were quite comfortable working with DUC.	HF-28

4.5.2. Discussion

Detailed results and statistics from VAL2 are provided in the UC3 Annex. The following table details the results in relation to the specific EASA objectives.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Table 20. UC3 VAL2 results in relation to each individual EASA requirement.

Related requirement validation objective (D6.3)	Validation result (related to requirement)
HF-02: DUC must be able to guide the UAM Coordinator's attention to important information based on an understanding of where the UAM Coordinator is at the moment (see HF-13). DUC must be able to provide information about specific U-space elements and flights on the UAM Coordinator's request.	DUC could suggest the operator to redirect the map to an area of importance. The function was called “Take me there” and consisted of an additional button on the DUC dialog window. When pressed, the map would shift to the situation presented in the dialog window. When the function is used, it jumps directly to the situation in question. This made participants lose awareness of where on the map this situation is located. To regain spatial awareness, the participants were needed to zoom out and then zoom back into the situation again. Attention guidance lowered workload. From the collected data, the subjective workload was rated lower when the “Take me there”- function was used. Stated by participants, when working in a new environment the function helped when they did not have good geographical knowledge. Out of 9 statements in the questionnaire, 3 statements had a significant difference. These significant statements all indicated that attention guidance helped when navigating on the map. When the “Take me there”-function was used, overall situational awareness was lower. According to multiple participants, using the function made them disoriented. Questionnaire responses showed that DUC supported participants' problem solving.
HF-03: DUC must be able to track the UAM Coordinator's visual attention to identify gaps in his/her perception of elements in the environment.	We were not able to track the UAM Coordinator’s visual attention and therefore not able to identify gaps in visual information sampled.
HF-04: DUC must allow the user (UAM Coordinator) the flexibility to accept, reject, amend, and/or ask for an explanation, in response to any provided recommendation. The response time window available (to accept / reject / amend / seek explanation) must be adequate to allow the user to assess the recommendation.	While the “Take me there”-function was helpful in redirecting attention, it was not always needed. The ability to choose to use this function or not, was positive. Participants were positive about the idea of choosing to use the “Take me there”-function as attention guidance support. Information provided through the dialog windows were deemed as clear and very important for situation awareness but could be slightly disturbing, i.e. covering a large part of the screen.

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

System response logic (time-out vs auto-implement) shall be made explicit, in response to user actions (i.e. accept / reject / amend / seek explanation).	Information provided by sound was deemed as important for situational awareness and only mildly disturbing. For activation of a geofence it was deemed very clear. The symbol in the upper left corner of the dialog window was deemed mildly important. The information provided through the dialog window was deemed equally important as the information presented on the map in the medical emergency scenario events. Decision making was deemed moderately challenging in the medical emergency scenario events. Slightly less challenging with the abstract representation.
HF-08: DUC must be able to suggest alternative solutions and provide arguments supporting them.	The options provided through the dialog windows were considered useful in both the Medical Emergency and Fire Emergency scenarios.
HF-09: The UAM Coordinator must be able to adjust high-level task parameters for DUC, such as adjusting goals and constraints for capacity thresholds, conformance monitoring, and alerts and warnings.	DUC HAT interface allowing the UAM Coordinator to control/adjust goals, high-level tasks parameters and objective thresholds was not implemented.
HF-10: DUC must be able to propose the correct (normal- or emergency), or provide it as requested by the UAM Coordinator.	Natural language interaction was not explored.
HF-11: DUC must alert the user to any potential user misunderstandings, as inferred by the UAM Coordinator's visual attention (see HF-12).	Natural language interaction was not explored.
HF-12: DUC must be able to determine if the UAM Coordinator has misunderstood an alert provided by DUC based on the UAM Coordinator's visual attention.	Natural language interaction was not explored.
HF-13: If DUC provides information or an alert related to a situation that is shown on the map display, DUC can assess if the UAM Coordinator has understood the information or alert by checking if the UAM Coordinator is looking at the correct location on the HMI. If the UAM Coordinator is not looking at the correct	Natural language interaction was not explored.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

location, DUC should be able to guide the UAM Coordinators attention (see HF-02).	
HF-14: DUC must not "step on" other ongoing user communications.	Answers from the HAI questionnaire indicate that almost half the participants found DUC to interrupt them.
HF-21: DUC must permit user deactivation of voice mode.	This interaction was not explored. This is a function intended to be explored in the HAT interface.
HF-24: DUC must allow adaptation of its notification modalities (visual, aural) depending on situation state (nominal / non-nominal / emergency) and user cognitive state.	This interaction was not explored. This is a function intended to be explored in the HAT interface.
HF-25: DUC must allow adaptation of its interaction modalities depending on situation state (nominal / non-nominal / emergency).	Overall, DUC was perceived to communicate clearly. But some found DUC to frequently interrupt them. Information provided through dialog windows was perceived clear and very important for situation awareness but could be slightly disturbing (i.e. covering a large part of the screen). Information provided by sound was considered important for situational awareness and only mildly disturbing.
HF-28: DUC must have a clear and intuitive HMI that minimises confusion and misunderstanding of how to interact with DUC.	Observations and questionnaire responses showed that participants overall found the DUC HMI clear and intuitive. However, there was confusion and misunderstanding of how to use the checklists. The training session, prior to the actual simulation, ascertained that all participants had received instructions on using the checklist, how to interact with it, and when to use it. There was also an example situation where they demonstrated using the checklist. Despite this training, several participants did not use the checklist, or partly used the checklist, in the first simulator sessions. Observations indicated that participants using the checklist performed better and experienced less confusion.
EXP-10: DUC must be able to explain its decisions and recommend stations using progressive CLT levels, for its recommended actions. regarding, traffic monitoring, coordination with vertiports, and provision of flight and weather information.	CLT 1-3 were used and divided so that users can interactively go from one CLT to another if seeking more details. Two explainers were made in text and video storytelling formats. Animations in the storytelling format were judged by participants easy to grasp and follow. Added narration was also found to clarify explanations. Text was mentioned as a format that you can easily scroll back and find needed information. With video storytelling format it is not possible to do.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

	CLT 2 was found to be the most suitable one for the emergency situation. Participants were satisfied with presented explanations and were able to make a final decision after going through the explainers. Most participants were more satisfied with video storytelling format in comparison with text format. Storytelling format gave a visual basis that was found to be easy to grasp for the participants. However, the lack of a function where users can scroll back in the video if needed was perceived as a disadvantage of this format in comparison to text. Another disadvantage of storytelling format was the length of the video which could be a difficulty in the operational environment.
EXP-11: DUC explanations must be presented using a combination of text, symbols, and graphical overlays on the map display.	In the storytelling format explanations were presented with animations, graphics, narration, icons, text. In the text format, text, bullet points and tables were presented. The approach was successful in conveying complex trade-offs and enhancing trust, with participants rating the explanations as easy to understand, and motivating. Comparisons across three video presentations suggested that this storytelling method, when integrated with the situation display, effectively supports decision-making and clarity in geofence selections. The combination of video format (text, symbols, and graphics) helped participants understand the reasoning behind DUC's recommendations, particularly in understanding trade-offs and improving decision-making confidence. The storytelling explanation format was clear, engaging, and perceived as trustworthy, with a strong agreement that the explanations met expectations. Video storytelling explainer improved users' ability to understand geofence recommendations and their implications for U-space operations.
EXP-12: DUC should be able to explain why different solution alternatives are proposed, including factors considered and how relevant they are, and the underlying decision-making process.	This evaluation explored how well the DUC explains its rationale for proposing different solution alternatives, using a video storytelling and text format. Results indicate that users found the explanations clear, trustworthy, particularly in presenting relevant factors and trade-offs. The storytelling method with video, text, and visuals supported comprehension of the decision-making process, reinforcing user confidence in the recommended choices. Participants understood why the DUC recommended specific geofence options, indicating that the explanations effectively communicated the rationale behind different alternatives.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

	There was a high level of trust in the DUC's recommendations, which correlates with participants finding the explanations sufficiently detailed and transparent.
EXP-13: DUC must present its decision-making rationale both prior to, and during, situations, consistent with CLT. Prior explanations shall rely on a storytelling approach, concurrent explanations shall use a combination of text, symbols, and graphical overlays on the map display.	Explanations were presented during the emergency situation in the simulator, prior to the final decision making from the UAM Coordinator explaining presented options for the decision. Ratings showed that participants consistently understood the decision-making rationale, trusted the explanations, and felt confident choosing between geofence options. CLT1 was found to be not detailed enough, CLT2 was the most suitable for the decision making and CLT3 was too detailed for most participants.
EXP-14: DUC must provide dynamic adaptation of explanation levels, per CLT.	Each CLT was adapted to the scenario situation, with an increasing level of detail per each CLT. We evaluated subjective perception of CLTs.
EXP-15: Explanations prior to a situation (for training purposes) are timed to appear during periods with little activity. Explanations for decisions/actions in time-critical situations are timed to occur when DUC has derived a solution. The explanation is valid if the proposed decision/action is valid.	Feedback suggests that participants had sufficient time to decide, and did not feel stressed.
EXP-16: The system must retain a user-selectable explanation level (according to CLT), for all explanations provided by DUC.	Participants rated their understanding of recommendation the highest in CLT3 (Mdn = 5) for both text and video format. In the text version CLT1 participants disagreed with a statement about their understanding of geofence options after going through explainer. Participants chose CLT2 as the most suitable for emergency situations based on the interview.

4.6. UC3 VAL 2 Conclusions

4.6.1. UC3 Research Questions

In the UC3 validation, participants were tasked to collaborate with the DUC in providing U-space services to U-space users (i.e., flights) as part of U-space traffic management. In three different scenarios, the team (i.e., participant and DUC) faced different emergencies that they were required to

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

solve. The following presents the conclusions to the UC3 Research Questions. Note that both research question 3 and 4 address the explainability of the DUC and are addressed in Deliverable 5.2.

Research question 1: How do air traffic controllers perceive the impact of collaborating with the DUC in U-space traffic management operations, particularly regarding its influence on their situation awareness, communication practices, perceived level of control, and allocation of task responsibilities?

DUC supported situation awareness and problem solving while communicating well. But some found DUC to frequently interrupt them and did not find the checklist supportive for working with DUC. Some participants appeared confused about the responsibilities of the UAM Coordinator and DUC, whereas others thought it was clear. Some participants found the checklist helpful for working with DUC, while others found it less valuable.

From the HAT questionnaire, and debriefings, results clearly show that participants do not consider the DUC a team member, but a tool. This poses a challenge in design, in that while we can strive for designing a team member, we cannot force users to perceive the system as a team member.

An important enabler for the collaboration between the UAM Coordinator and the DUC was the checklist, which clearly divided tasks and responsibilities between the two actors for a given event. Participants' perception on the usefulness of the checklist varied, with some participants finding it supportive, while others did not.

The Social Acceptance Questionnaire sought to explore participants' acceptance of the DUC intelligent assistant. Participants responded that the UAM Coordinator will depend on DUC for U-space operations, and they considered DUC provides efficiency and safety benefits. To work with DUC, training will be needed. Trust in DUC is important. There were differences in participants' perceptions of safety in U-space operations working with DUC

Research question 2 (scenario: Medical Emergency): How does situated vs. abstract representation influence operators' workload, situational awareness, decision making challenge. and feeling of control?

The general tendency leaned towards a preference for the abstract format by a majority of the participants/ATCOs for both the medical event (5 ATCOs of 9) and the delivery drone routine reroute event, and slightly more for the routine event (6 ATCOs of 9). Especially the colour coded ranked recommendation (with the most favourable indicated by a green dot) in the dialogue window served as a guide in a situation where the context and geography were not fully known and familiarized with. A few participants (2 for the delivery drone event and 3 for the medical event) still preferred the situated representation, which was the format without any additional comment or coloured dot from DUC. One participant explained that even though the abstract format was easier, the system should trust his decision.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Participant 1 did not answer the questionnaires so the ratings of situational awareness workload, feeling of control and challenge of decision making were based on the answers from eight participants. For the remaining eight participants, there was no clear difference in indicated situation awareness between the two representations, and only slight non-significant tendencies towards less workload and feeling of control for the abstract format. However, a Wilcoxon signed rank test showed a significant difference and a lower challenge of decision making for the abstract format. (See Annex).

In summary, the abstract format was preferred by most of the participants, although one participant raised concerns about the Intelligent Assistant's trust in the human operator's competence. The abstract version was also ranked as a lower decision-making challenge with no or minor impacts on situational awareness, workload, and feeling of control compared with the situated version. These findings would demonstrate an example of a functioning Reduced Autonomy Workspace (RAW) in accordance with an efficiency goal as well as a development to a safety goal situation.

Research question 5 (scenario: Link-loss Situation): How does situation-switching attention guidance provided by DUC influence operators' overall (main monitoring) situational awareness (comparing attention guidance vs no attention guidance)?

Participants were generally positive regarding the attention guidance ("Take me there") function. The questionnaire showed a significant finding that participants rated the ability to find the current situation higher with attention guidance. From interviews, it was stated the attention guidance was a helpful tool, especially when the operator has no or low geographical knowledge. Multiple participants stated they had low geographical knowledge regarding the area of Stockholm used in the validation. The ability to choose to use the function or not, was positive. Some stated using the function would give them the correct information of where the area of importance is and then can by themselves navigate between the situations when they feel the need for it.

Multiple participants mentioned that they felt disoriented when first using the function. The function shifts directly (from one frame to another) to the position without showing where it is moving on the map. This made the participants lose the spatial information of where they were on the map. Some solved this design flaw by zooming out to get an overview and continued to zoom back into the situation in question. Furthermore, some participants stated using the function made them lose awareness of the surroundings, making them focus more on the specific situation. Both cases indicate a loss in situational awareness.

From the above, we conclude the following key findings from UC3:

- **Explore attention guidance mechanisms to support situation awareness.** Attention guidance mechanisms can help operators quickly build awareness of where something is happening by drawing their attention to relevant geographical areas. However, this may not help them understand what is happening. A potential drawback is that such guidance can disorient operators, causing them to lose awareness of the overall map layout and how the new focal point relates to their previous view.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- **Use attention guidance to reduce operator workload.** Implementing attention guidance as a "take me there" function can reduce operator workload by eliminating the need for manual navigation, making it easier to find relevant locations quickly.
- **Adapt communication and interaction modalities to context**. Communication methods should be flexible and tailored to the context and user preferences. It is recommended to design text-based messages in a way that maintains their clarity and readability while minimizing visual obstruction and the risk of being overlooked during high workload periods. Messages should be positioned and timed to avoid blocking critical map information needed for decision-making. It is recommended to complement text communication with aural alerts and voice messages to enhance situation awareness. Additionally, overlaying geographically relevant information directly on the map can support spatial orientation and help users clearly associate messages with their corresponding locations.
- **Use collaborative checklists to support teaming**, with training. Checklist-based collaboration methods between human operators and intelligent assistants can enhance teamwork, but it requires considerable training and familiarity. Further research should examine how collaborative checklists impact shared understanding, task performance, and problem-solving.
- **Explore operator readiness for AI-based teammates.** Operators seem to perceive AI-based systems as tools rather than teammates, even when those systems are designed for collaboration. This suggests deeper cognitive and cultural factors influence perception. Future work should focus on training, interface design, and integration strategies to help operators view AI systems as active members of the team.
- **Favor abstract over situated communication formats for AI output**. It is recommended that AI systems communicate using abstract formats, e.g., using centralized text windows, rather than situated formats such as spatially distributed, map-based messages. Abstract formats provide good support for situation awareness and a sense of control, while also reducing cognitive workload and significantly easing decision-making. These benefits make them a more effective and accessible option for a wider range of users. Situated formats may benefit operators that are highly familiar with the geographic context.
- **Engage end users early in system design.** Involving end users early in the design process is crucial for ensuring that AI systems are seen as effective and trustworthy team members. Early engagement can guide development toward features that align with users' expectations and operational needs.
- **Investigate the role of AI soft skills in team perception**. More attention should be given to how the AI's "soft skills" (such as timing, tone, phrasing, and interaction style) affect operator perceptions. These subtleties emerge as important for the influence whether AI systems are accepted as collaborative teammates.

4.6.2. HAIKU High-level Research Questions

UC3 results primarily inform the first and second HAIKU research question.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

HAIKU Q1: What is the recommended human-AI relationship for each of the different AI aviation applications?

Introduce AI gradually as a teammate, starting with tool-like roles. UC3 findings suggest that HAT should follow a gradual implementation path, where IA systems are initially introduced as tools rather than full collaborative agents, especially if experienced operators are retrained for HAT. This phased approach allows operators time to adjust their expectations and build trust in the IA. The distinction between a tool and a teammate is important, as it shapes how operators understand the system's capabilities, responsibilities, and how they are expected to interact with it.

Acknowledge that perceptions of AI as a tool or teammate are psychological, not purely functional. Operators' perception of an AI system as either a tool or a team member may not align with the system's actual capabilities or intended functionality. We expect this perception to be a psychological construct, influenced by cognitive and cultural factors. Therefore, fostering team-oriented perceptions requires more than technological advancement, it also demands attention to user mindset, training, and system presentation.

Prioritize human-centred design to support adoption. Adoption of AI-based systems as effective team members depends on human-centred design. IAs must be designed to genuinely support human needs, workflows, and decision-making processes. When systems align with users' mental models and operational demands, the likelihood of successful integration and team acceptance increases.

Enhancing situation awareness through interface features like attention guidance and multimodal communication. These are seen as essential for helping operators stay oriented and informed, regardless of the specific AI task.

Reducing cognitive workload by implementing functions such as "take me there" attention and navigation guidance, and abstract communication formats, which support efficient decision-making and ease user interaction across applications.

Adapting communication and interaction to context, highlighting the need for flexibility in modality (text, audio, visual overlays) to match varying operational demands and user preferences.

Supporting HAT through structured work methods like collaborative checklists, which improve teaming but require training.

Addressing the perception of AI as a teammate, which remains a barrier despite collaborative system designs. Recommendations emphasize the importance of training, interface design, and cultural change to shift operator attitudes.

Focusing on communication style and soft skills of AI systems, recognizing that how AI communicates (not just what it says) significantly affects team dynamics and trust.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Engaging end users early in the design process, reinforcing that user-centred development is a universal best practice to increase adoption and effectiveness across domains.

HAIKU Q2: What does it mean for AI to be explainable?

UC3 results addressing this question are discussed in D5.2.

HAIKU Q3: How to train AI to assist humans in safety critical tasks when training data are insufficient?

Given that UC3 has worked with an IA at a proof-of-concept level, and not explored actual AI models, there are very limited results addressing the third HAIKU research question.

4.6.3. Research Recommendations

From VAL2 results in UC3, we propose the following future research directions:

- Research should prioritize the exploration of EASA automation levels 2A and 2B, as these represent critical thresholds for shared authority and collaboration between human operators and AI-based systems. Understanding how human-AI interaction unfolds at these intermediate levels will be key to safe and effective integration of AI in aviation.
- Development efforts should focus on building prototypes grounded in a shared theoretical model of human-AI teaming. While it may be necessary to restrict focus to specific subcomponents of the model in early stages, aligning prototypes with a common framework will ensure comparability across studies and more coherent progress toward operational implementation.
- There should also be a stronger emphasis on incorporating AI systems into training environments and Crew Resource Management (CRM) programs. This includes not only training humans to work with AI but also using AI as part of the training process itself. This dual role can help normalize AI as a team member and improve overall team coordination and trust.
- Finally, future research should shift the emphasis from AI's task-specific performance to its teaming capabilities. Understanding how AI contributes to team processes, e.g., as shared situation awareness, communication, decision-making, and coordination, is essential for effective integration. This shift in focus will ensure that AI systems are not only intelligent, but also truly collaborative. Future research should investigate how cognitive and cultural factors shape operators' perceptions of AI systems as tools rather than teammates, regardless of the system's capabilities. This includes exploring how training, interface design, and team integration strategies can be used to shift these perceptions toward more collaborative human-AI relationships.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

5. Use Case #4 – Digital and Remote Tower

5.1. Deviation from Validation Plan (D6.3)

VAL2 was performed according to the validation plan described in D6.3 with the following deviations:

- Exercise design: Given that we had achieved a satisfactory level of system performance, we opted to run a single, longer simulator exercise instead of the two initially proposed in D6.3. In that deliverable, we considered conducting two runs: one in which ATCOs were asked to consult ISA (Intelligent Sequence Assistant) but were not obliged to follow its suggestions, and another in which they were instructed to stick to any suggestion made. This approach aimed to observe and record their reactions and ultimately assess whether their mental models aligned with ISA's. It also served as a fallback strategy in case the system failed to function correctly in real-time, allowing us to compare the ISA-generated sequences with those produced by the ATCOs. However, given the robustness of the prototype reached right before the VAL2 session, we decided to abandon the two-run approach in favour of a single, extended run. In this session, we chose not to give ATCOs any explicit instructions, allowing them to engage with the system naturally. This created a more realistic interaction scenario, enabling us to explore how they would use the system in an operational context.
- Also, we removed the part of an exercise when we deliberately make ISA fail because it was difficult to coordinate with pseudo pilots.

5.2. Validation Objectives

In accordance with D6.3, the HAT objectives in VAL2 addressed three key Human Factors challenges for Human-AI Teaming: situation awareness, explainability and error management.

Situation awareness: The situation awareness objective requires that both ATCOs and ISA are continuously aware of the traffic situation, processing real-time data from various sources to detect trends, anomalies, potential conflicts, and anticipate future traffic patterns. Additionally, ISA should effectively communicate with the ATCO in charge by providing real-time information regarding the current sequence and, when applicable, new sequence suggestions.

Explainability: The XAI objectives require that ISA provide clear and relevant explanations regarding sequence changes. Furthermore, ISA must be able to provide different levels of explainability (CLT levels) upon ATCO's requests.

Error Management: ISA must be able to detect and inform ATCOs in charge about potentially unsafe situations, working as an additional safety net

These HAT objectives translate to three research questions that shaped our VAL2 approach:

- *RQ1: What is the best possible set up for the ATCO and ISA Human-AI Teaming configuration?*

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- *RQ2: What levels and types of explanations are most effective for enabling ATCOs to understand and trust ISA-generated sequence changes during varying operational demands?*
- *RQ3: Can ISA detect and communicate potentially unsafe situations to ATCOs in a timely and actionable manner to effectively support collaborative safety management? Is ISA a reliable safety net?*

These HAT Objectives and Research Questions were then associated with the system requirements (described in D4.1) that shaped the whole validation process. More details are given in the "Procedure" paragraph.

5.3. VAL2 Activities and Methods

Starting from the HAT Objectives and Research Questions, VAL2 conducted a qualitative assessment of ISA through eight sessions involving ATCOs using a fully functional prototype. The sessions simulated realistic traffic sequencing scenarios at Alicante Airport.

VAL2 aimed to validate the system holistically, with particular focus on its sequence generation and explainability features, designed for real-time operational use. Unlike VAL1, which was limited by the absence of real-time integration, VAL2 benefited from a prototype fully integrated with an ATC simulator. This allowed ISA to generate real-time suggestions based on live simulator data, enabling a more robust evaluation of its operational performance.

Each session involved ATCOs running a scenario while being observed by experimenters. After each session, a qualitative debrief (semi-structured interview) was conducted to gather insights into how the system was experienced. Interview questions were mapped to system requirements (D4.1) to assess whether they had been met. This in-depth analysis provided answers to the Research Questions and broader HAT Objectives.

The study structure followed this logic: for each HAT Objective (e.g., *Situation Awareness*) and its related Research Question (e.g., *What is the best possible set up for the ATCO and ISA Human-AI Teaming configuration?*), we identified associated requirements (e.g., HF-03: *ISA must continuously adapt and display updated sequences*). A targeted question was then asked to assess the requirement—for example: *How do you evaluate ISA's responsiveness to the ever-changing situation*?

The process for generating relevant insight followed this path:

HAT Objective → Research Question → EASA-like Requirement → Specific Interview Question -> insight about ISA.

In the next paragraphs all the details are provided.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

5.3.1. Participants

Table 21. Participant details in UC4.

Nr.	Job Role	Gender	Age	Level of Expertise	Other
1	Tower ATCO	F	23	4 years	
2	Tower ATCO	M	25	3 years	
3	Tower ATCO	M	48	17 years	Both military and civil expertise
4	Tower ATCO	M	45	15 years	Both military and civil expertise
5	Tower ATCO	M	40	6 years	
6	Tower ATCO	F	35	7 years	
7	Tower ATCO	M	36	9 years	
8	Tower ATCO	F	42	8 years	

The experiment involved 8 Tower ATCOs (3 women and 5 men) with varying levels of skill and expertise. In total there were five males and three females aged 23 to 48 (mean = 36.75, SD = 8.97). Their experience ranged from 3 to 17 years, with an average of 8.63 years (SD = 4.98). Two participants had both civil and military ATCO experience. The group reflects a balanced mix of early-career and seasoned professionals, providing diverse and representative feedback for validation purposes.

5.3.2. Simulator/Apparatus

The simulator is located in Skyway's office in Madrid. It replicates the same conditions as the Control Tower of Alicante Airport. The whole setup includes several screens (for the scenario and operational information, such as weather conditions) and computers, and two working positions (LCL and GMC controller), as well as several working positions for pseudo pilots as well (Figure 40).

The ISA prototype was developed as an independent web application, functioning as a standalone product. Initially, it was run on a separate laptop where the HMI was accessible. For the VAL2 session, we successfully integrated the ISA web application with the simulator, enabling data exchange between the two systems.

The final setup for VAL2 consisted of:

- The simulator running the main scenario, with pseudopilots.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- A separate laptop running the ISA application.

The communication between the two worked like this:

- The simulator sent aircraft data to ISA (including information on arrivals, departures, estimated times of arrival, etc.).
- ISA generated sequences based on the real-time data. These were reflected on ISA's HMI and electronic strips.

Both ISA and the simulator continuously exchanged data to ensure a seamless, real-time experience that accurately reflected the evolving scenario in the simulator.

Figure 40. Photograph of simulator session with participant working with ISA, and experimenter seated behind.

5.3.3. Scenario

A single scenario was developed and implemented in the simulator, and it was used across all eight individual sessions involving eight ATCOs. The scenario was based on operations at Alicante-Elche Airport, Spain's busiest single-runway airport, which experiences peak traffic levels exceeding 40 arrivals and departures per hour.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

In the simulation, controllers were required to manage air traffic while an AI system provided suggested sequences of arrivals and departures via an interface replicating the one used in the Alicante Control Tower.

Due to the dynamic nature of the exercise, the scenario evolved slightly in each session depending on the actions taken by the ATCOs. Although the core setup remained consistent, these variations introduced natural differences across sessions. Further details are provided in the Limitations section.

5.3.4. Procedure

The total duration of the validation session was two hours per ATCO, with one hour dedicated to simulation exercises and one hour for debriefing activities.

The VAL2 process consisted of the following steps:

1. **Briefing:** Participants received an overview of the project and ISA, including its high-level purpose and functionality. Specific details will be deliberately omitted, allowing the ATCOs to discover them during the experiments.
2. **Simulation Exercise:** Exercises were conducted in Skyway's simulator in Madrid. The simulator replicated a sequencing scenario in the Alicante airport control tower (it's the scenario described earlier in this chapter) to evaluate how ATCOs interact with ISA in a quasi-natural setting.
3. **Debriefing:** After each exercise, semi-structured interviews were conducted to collect qualitative feedback from the ATCOs regarding their experience with ISA.

Upon arrival at the simulator, participants were first provided with an informed consent form. We then gave a brief introduction to the HAIKU project, explaining what ISA is and outlining the purpose of the validation. To preserve the exploratory nature of the session, we shared only essential information: enough for participants to understand the context without giving too much away.

Participants were given a brief orientation of the simulator. As many were unfamiliar with the local environment, a brief overview of Alicante Airport was provided, covering key operational aspects such as runway layout, stand configuration, and known traffic bottlenecks.

The ISA prototype ran on a separate laptop communicating with the simulator Participants were shown the ISA HMI with electronic strips and instructed to use these during the exercise instead of the physical strips usually used in exercises at the simulator.

The exercise started with the activation of the simulation clock. Participants were tasked with managing arrival and departure traffic at Alicante using standard ATC procedures. Pseudo Pilots located in a separate room provided real-time interactions to simulate operational realism. Participants were instructed to focus on aircraft sequencing and to use the ISA HMI to issue clearances (e.g., take-off and landing) and ultimately follow the sequences that they had planned.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

At the end of the session, participants took part in a semi-structured interview to share their thoughts and experiences of interacting with ISA. The interview questions are mapped to the requirements described in D4.1 and are designed to gather insights regarding the Research Questions presented earlier.

As an example, given an HAT Objective (i.e.: Situation Awareness) and the related Research Question (i.e.: *What is the best possible set up for the ATCO and ISA Human-AI Teaming configuration?),* we asked a (set of) question(s) to investigate how and whether the Requirement HF-03 (*ISA must be able to always generate and show new sequences by constantly adapting to everything that is happening*) was satisfied. The question was: *How do you evaluate ISA's responsiveness to the ever-changing situation?*

The path to generate relevant questions for the experiment was the following: HAT Objective -> Research Question -> EASA-like requirement -> specific question.

The full set of interview questions is included in the UC4 Annex.

5.3.5. Data Collection Tools

Table 22. Data collection tools used in UC4.

Tool	Objective	Type of Collected Data
Observation	To observe what ATCOs do during the experiments (and not only what they say).	Qualitative
Semi-structured interviews	To gather ATCO's subjective impressions and feelings about ISA. The questions are mapped to D4.1 Requirements and EASA's macro-areas. Full details about questions used are provided in the Annex.	Qualitative
System logs	Systems logs were indirectly involved in validation, as they were mostly used for debugging purposes. They were leveraged and used by developers to improve system performance and bug fixing.	Quantitative

5.3.6. Data Analysis

The validation activity involved collecting qualitative data through semi-structured interviews and observation with ATCOs. The interviews were recorded and transcribed, capturing participants' raw responses. The transcriptions were stored in a database as the foundational dataset for analysis. Data gathered through observation was also integrated.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

To derive insights, we applied thematic analysis, a method that identifies, organises, and interprets patterns (or themes) within qualitative data. This process involves systematically coding the data, grouping related codes into themes, and refining these themes to represent the key insights. Thematic analysis is particularly well-suited to exploring participants' experiences, perspectives, and interactions with systems. The insights gained from this analysis will inform the iterative refinement of ISA's design.

System quantitative data was collected by the extraction of log files from the simulation platform. The data was used to test the system's performance in order to improve it for the exercises.

5.4. TRL Overview: update

The planned trajectory was confirmed during VAL2, achieving TRL6 by successfully validating the system as a whole (Optimisation Algorithm to output sequence, Service for computing ETA and initial trajectory points and Explainability module for sequence changes, including HMI with Electronic flight strips and strip board management).

Table 23. TRL progress in UC4 for the Optimisation component.

Components: Optimisation Algorithm to output sequence, Service for computing ETA and initial trajectory points		
TRL	**Month**	**Activity to reach selected level**
1	1	Concept formulated at the project's start.
2	12	Literature review and meetings with operational experts to refine the concept.
3	18	An initial proof of concept of the technology was discussed. There were both technical feasibility and desired features assessments. Before the VAL1 experiment, data was shared among the technical partners, and an initial prototype was developed to test the models and the new features
4	24	VAL1: Sequencing exercises were conducted by ATCOs, and their results were subsequently compared to those generated by the ISA for the same exercises. The results were found to be comparable, leading to the transition to TRL 4
5	30	We reached TRL 5 by validating individual components during the integration of the ISA prototype with the simulator, testing performance using the Alicante airport scenario (VAL1 to VAL2).
6	33	We reached TRL 6 by validating the integrated system with real end users in a high-fidelity simulated operational environment (VAL2).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Table 24. TRL progress in UC4 for the Explainability component.

Components: Explainability for sequence changes, HMI - Electronic flight strips and strip board management		
TRL	**Month**	**Activity to reach selected level**
1	1	Concept formulated at the project's start.
2	12	Literature review, meetings with Experts, user needs specifications gathered through user research (i.e.: Observational studies in the Tower by Human Factors experts).
3	24	VAL1: we designed a realistic, interactive mock-up of the HMI with the electronic strips and the explanations. These were tested qualitatively with users, who validated our design.
4	28	Full prototype of each component developed and tested.
5	30	We reached TRL 5 by validating individual components during the integration of the ISA prototype with the simulator, testing performance using the Alicante airport scenario (VAL1 to VAL2).
6	33	We reached TRL 6 by validating the integrated system with real end users in a high-fidelity simulated operational environment (VAL2).

5.5. VAL 2 Results

5.5.1. Results Overview

Table 25. UC4 high-level results in relation to the EASA Validation objectives.

High-level Results	**Brief description**	**Related Validation Objective (D6.3)**
Situation awareness	Both ISA and the ATCO were aware of the traffic situation	HF-03; HF-05; HF-08; HF-28.
Explainability	Explainability was leveraged by ATCO and was considered useful	EXP-10; EXP-11; EXP-12; EXP-13; EXP-15; EXP-16; EXP-17; EXP-19.
Error Management	We discovered that, when ISA provided alerts to the ATCOs to inform about potentially unsafe situations, that feature was considered very useful and led to much better interactions with the system	HF-28; HF-29

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

5.5.2. Discussion

Here we provide the validation result addressing each individual EASA requirement.

Table 26. UC4 VAL2 results in relation to each individual EASA objectives.

Related requirement validation objective (D6.3)	**Validation result (related to requirement)**
HF-03: ISA must be able to always generate and show new sequences by constantly adapting to everything that is happening.	ATCOs generally felt that ISA could adapt to the ever-changing situations. Some noted that ISA was more aggressive than their own approach, while others found it conservative. The adaptability of ISA was appreciated, but there were mixed feelings about its aggressiveness. One ATCO felt that ISA's responsiveness had a slight lag.
HF-05: ISA must be able to recognise a potential suboptimal sequence generated by user's interaction and suggest a better alternative to the ATCM.	Some ATCOs felt ISA adapted well to situations where pilots did not comply with instructions, while others did not provide specific feedback. The ability of ISA to recognize suboptimal sequences was seen as beneficial, but there were instances where ATCOs felt it could be more efficient.
HF-08: ISA must be able to always suggest the best possible solution to a sequence and suggest it to the ATCO, regardless of what the ATCO does.	ATCOs felt ISA was a bit conservative and sometimes inefficient. Some noted that ISA's suggestions were logical but not always the best possible solution. The conservativeness of ISA was a common point of contention. ATCOs appreciated the logical approach but desired more efficiency in certain situations.
HF-28: ISA must tolerate and adjust to user manual inputs, by maintaining an ongoing sequence recommendation capability.	ISA was able to adjust to the user's inputs and change the sequence accordingly. All ATCOs agreed that ISA was able to adapt appropriately to their decisions.
HF-29: ISA must detect and adjust to user manual inputs, by maintaining an ongoing sequence recommendation capability.	ISA was correctly able to alert the ATCO via the HMI (there was an exclamation mark on the electronic strip where a potential safety issue was detected). ATCOs correctly noticed it and talked about it in the debriefing.
EXP-10: ISA must be able to generate explanations related to sequence generation (for monitoring), and ad-hoc explanations for sequence changes.	ATCOs found ISA's explanations useful, especially regarding ETA. The usefulness of explanations was acknowledged. Some ATCOs wanted even shorter explanations (e.g., symbols instead of text).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

EXP-11: ISA HMI must show explanations related to sequence generation and sequence changes.	ATCOs appreciated the clarity of the explanations provided by ISA's HMI. The feedback was generally positive. The clarity of explanations was well-received, and ATCOs found the HMI intuitive and helpful.
EXP-12: ISA must be able to show, for each sequence change, different levels of explainability with a progressive level of detail that is keyed to the expected decision / action.	ATCOs found the level of detail sufficient but some suggested even less detail in some cases. The progressive levels of detail were useful, but there was a suggestion for even more concise explanations in certain situations.
EXP-13: ISA HMI must provide a means for user to discover progressive levels of detail details about any provided explanation	Feedback was positive, with ATCOs finding the details useful. Some suggested that mouseover could show symbols and be shortened. ATCOs found the level of detail sufficient but some suggested even less detail in some cases.
EXP-15: Explanations about sequence changes must be available to the user minimum delay and permit progressive levels of detail keyed to user needs.	Timing was generally appropriate, but some ATCOs suggested more time for reaction. The timing of explanations was mostly appropriate, but there were suggestions for slight adjustments to allow more reaction time.
EXP-16: The user must be able to access ISA explanations about sequence changes, via progressive disclosure interaction, as desired by the user.	ATCOs felt the current level of detail was sufficient. The current level of detail was sufficient, and the progressive disclosure interaction was well-received.
EXP-17: ISA-generated sequences must optimize operational priorities (by default, throughput shall be optimized).	ATCOs noted that ISA prioritizes safety and efficiency. The prioritization of safety and efficiency was appreciated, and the parameters influencing explanations were considered logical.
EXP-19: ISA must alert users to any potentially unsafe ISA-recommended sequences (which can be caused by outdated recommendation, or off-nominal flight trajectories).	ATCOs appreciated the blinking warnings for potential conflicts. The reasons behind the blinking were interpreted as warnings for tight situations. The blinking warnings were useful, and ATCOs understood them as indicators of potential conflicts.

Several limitations affected the validation activities. Firstly, there were initial technical issues with ISA's performance, particularly during the first run, where ISA encountered an error and produced wrong sequences for several minutes. These were progressively resolved after careful analysis of ISA's

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

logs, allowing the subsequent exercises to run smoothly. Importantly, these issues did not undermine the overall exercise or the post-run discussions, as we were able to typically diagnose the errors and adjust the exercise in real time. For example, if we noticed that there was a problem with arrivals of certain aircraft, we limited the numbers of aircraft and we removed the faulty aircraft from the simulator for the subsequent exercises. We had to redesign the exercises on the fly to make sure that everything was smooth and that we could leverage those results.

Coordination with the pseudo pilots also presented challenges. As humans, they occasionally made errors or "off the script" decisions, which sometimes caused ISA to behave in unexpected ways. Given that ISA is not currently fully operational and, while robust, it needs more training to be perfect, this led to some issues that were fixed by giving more precise instructions to pseudo pilots.

Additionally, the highly interactive and performance-driven nature of the exercise made it difficult to objectively compare runs. Although a common script was used and the scenario remained consistent across sessions, the actions taken by each ATCO or pseudopilot influenced how the scenario unfolded. For example, if an ATCO made a particular sequencing decision, it could alter the subsequent flow of events. Similarly, if a pseudopilot communicated slightly earlier than in another session, this too could affect the scenario.

As a result, the number of operations varied between exercises, and the inputs from pseudopilots were not always identical. This variability made direct comparisons and more complex experimental setups unfeasible. Consequently, the data collected represented eight distinct and subjective experiences. Nonetheless, the sessions revealed consistent patterns that contributed to the validation of key ISA features.

5.6. UC4 VAL 2 Conclusions

VAL2 pointed out that there is room for improvement in certain design aspects. Here is a concise overview of potential improvements:

- Make sequence changes more visually prominent
- Add optional audio signals for critical warnings
- Shorten mouseover explanations or replace text with symbols
- Reduce default detail in advanced explanation levels (CLT3/CLT4)
- Prioritise concise explanations during high-traffic scenarios
- Reduce lag in sequence updates and recalculations
- Provide explanations earlier to allow more reaction time
- Clarify the meaning of blinking/exclamation mark alerts
- Adjust conservatism to better balance safety and efficiency

Besides these, the overall concept of ISA has been validated and was well received by ATCOs. In particular, its added value in terms of safety and its explainability features were highlighted as the most appreciated and operationally useful components. It is important to **highlight how the ATCOs**

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

found ISA particularly useful as an additional safety net, since the visual alerts for potential conflicts and unsafe situations have provided significant added value to the system.

On the other hand, it is worth noting that ISA, as it is currently designed, is not useful for improving efficiency, as each controller has their own working style (more aggressive or more conservative), and in order to achieve that efficiency, ISA would need to be able to adapt to (and learn from) each controller's way of working. ISA, as currently designed, sequences one aircraft per minute to use the runway. Some ATCOs feel that this is too much, while others think it is not enough.

The following sections explore these aspects in more detail, first by addressing our use case-specific research questions, and then by drawing broader insights relevant to HAIKU's high-level research questions and the wider aviation domain.

5.6.1. UC4 Research Questions

As mentioned earlier, the flow for generating insights is the following:

HAT Objective → Research Question → EASA-like Requirement → Specific Interview Question -> insight about ISA.

Hence, we look at the answers provided during the interviews to try and address the broader Research questions and HAT Objectives.

RQ1: What is the best possible set up for the ATCO and ISA Human-AI Teaming configuration?

We observed that ATCOs were most effective when they retained control over sequencing and used ISA as a secondary, confirmatory aid. In this sense, it's paramount that ATCO maintain Situation awareness and that that is enhanced by the intelligent assistant. This supports the design choice of ISA to augment Situation Awareness. The tool is a suggestive, rather than directive, tool. ATCOs appreciated that ISA's suggestions could be acknowledged or dismissed freely, reinforcing the importance of designing the assistant to be non-intrusive. The most effective HAT configuration is one where ISA provides input without overriding the ATCO's authority, acting as a cooperative decision-making tool - which is the direction we chose when initially designing the system. In this sense, we can say that systems like ISA should be designed to enhance Situation Awareness (improving the ATCO situation's awareness in real time), and they should never disturb the ATCO to a degree where they lose track of what is happening. This ties nicely with the idea of providing different levels of explainability, as seen in RQ2.

RQ2: What levels and types of explanations are most effective for enabling ATCOs to understand and trust ISA-generated sequence changes during varying operational demands?

The general on-demand explanation model was well received. All participants used the CLT2 level (concise, text-based explanations) during exercises, particularly for sequencing tasks (i.e.: to check aircraft's ETA times). Embedding explanations within a hover effect on the electronic strip interface

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

ensured they were accessible without being disruptive. Feedback suggests even more minimal formats (e.g. icon-based CLT2) could be explored. Higher levels of detail (CLT3/CLT4) were rarely accessed, likely due to time constraints or lack of perceived need. However, participants acknowledged their value for training or debriefing scenarios. Overall, CLT2 represents an optimal balance between informativeness and operational efficiency. Considering that it has become clear that ISA is useful during periods of high workload, and that during these periods ATCOs have little additional time for tasks that are not specifically operational, such as monitoring traffic, landing and take-off clearances, etc., the conclusion is that quick and concise explanations are preferred by ATCOs during validation. CLT3 and CLT4, with long or less accessible texts, have hardly been used, while concise explanations that did not require action (click), such as CLT2, have been rated very positively. A more detailed focus on explainability for this UC, and the CLT framework, is given in D5.2.

RQ3: Can ISA detect and communicate potentially unsafe situations to ATCOs in a timely and actionable manner to effectively support collaborative safety management? Is ISA a reliable safety net?

This emerged as a key finding of the validation. Originally designed to improve sequencing efficiency, ISA's ability to highlight potentially unsafe scenarios became its most valued feature. ATCOs appreciated pre-warnings about unsafe sequences, which supported their situational awareness and decision-making. One participant (P3) noted: "ISA helped me change my plan. I made a mistake, and ISA basically told me 'This won't work, watch out.' It gave me time to think."

Though initially implemented as a secondary feature, this functionality proved critical in enhancing safety and prompting ATCOs to reassess decisions when necessary.

5.6.2. HAIKU High-level Research Questions

Based on the insights from the initial analysis, this section presents an abstraction of the findings in relation to the project's higher-level research questions. These abstractions serve as contributions from UC4, offering answers to the overarching research questions from its specific perspective.

HAIKU Q1: What are the common recommendations concerning Human-AI teaming for the different AI aviation applications?

We observed that ATCOs were most effective when they maintained control over sequencing, using ISA as a secondary, confirmatory aid. It is therefore essential that Situation Awareness remains with the ATCO and is actively supported by the intelligent assistant. This reinforces the design principle that SA should be enhanced by design - for example, by ensuring ISA's suggestions are timely, clearly presented, and aligned with the ATCO's mental model of the ongoing traffic situation.

ATCOs appreciated ISA's suggestive (rather than directive) role, valuing the ability to freely accept or dismiss its recommendations. This highlights the importance of keeping the assistant non-intrusive. The most effective Human-AI Teaming configuration is one where ISA supports decision-making

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

without overriding the ATCO's authority, acting as a cooperative tool, an approach that guided our initial design.

Systems like ISA should enhance real-time SA without distracting the ATCO or interfering with their judgement. This aligns with the concept of adaptive levels of explainability, as explored in Research Question 2. ISA's design can serve as guidance for HAT of other real-time systems. However, these principles may not apply to systems that do not operate in real time. In that sense, ISA directly addresses the research question by focusing on real-time operational support.

HAIKU Q2: What does it mean for AI to be explainable? Is explainability a necessary precondition of trustworthiness?

In UC4, explainability meant providing the simplest and most relevant rationale behind operational decisions. Designing for explainability requires careful user research to define requirements that are specific to the users' informational needs, shaped by their operational context. Designers must understand what information users need, when they need it, and how it should be delivered to fit smoothly within their workflow.

As with any design task, there is a subjective element. However, across all eight sessions, ATCOs consistently expressed a preference for brief, to-the-point explanations. They wanted quick insights into ISA's reasoning to support efficient decision-making. Long or overly detailed explanations were neither requested nor appropriate for the context and were never part of the design. The challenge for designers is to combine user needs with operational constraints to deliver the most relevant and minimal explanation that supports performance without adding cognitive load.

While explainability helps build trust, it is not enough on its own. Perceived robustness is equally important. ISA's ability to alert ATCOs to safety risks, especially those they had missed, played a key role in establishing trust. In this sense, trust is built not only on transparency, but also on the system's ability to improve outcomes, particularly in terms of safety. If the two go hand in hand, then you have a system that can be trusted (in the eyes of the user)

HAIKU Q3: How to train AI to assist humans in safety critical tasks when training data are insufficient?

UC4 does not provide a complete answer to this question but reinforces the importance of human-centred AI during transitional phases. ISA was trained on specific data from operations at Alicante Airport. However, scalability remains limited. Until larger datasets are available, AI should support - not replace - human expertise. In safety-critical domains like ATM, the human should remain in control, with AI offering reliable but subordinate support.

5.6.3. Research Recommendations

What we can conclude is that it would be interesting in future projects to explore

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- **Safety impact quantification** – measuring the number and type of safety events that AI systems help to detect or mitigate.
- **Explainability strategy optimisation** – further refining CLT framework design with dynamic levels to evaluate the time needed for ATCOs to process AI outputs and explanations under real-time conditions. For example, by predicting that ATCO's cognitive load in real-time before these peaks occur, we would be able to generate the right explanation at the right time to support decision-making, instead of providing 4 static levels.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

6. Use Case #5 – Airport Safety Watch

6.1. Deviation from Validation Plan (D6.3)

VAL2 was performed according to the validation plan described in D6.3 with no deviations.

6.2. Validation objectives

In UC5, validation objectives are grounded in functional, real-world application rather than experimental conditions, aligning with a TRL progression model. The core aim of VAL2 is to demonstrate that the ASW (Airport Safety Watch) Dashboard can support operational safety improvements through collaborative human-AI interaction. Eight Human-AI Teaming Requirements (HATRs) are addressed, such as the ability to process historical data (HATR1), support actionable insights within operational timeframes (HATR3), facilitate exploratory "what-if" analysis (HATR4), and enable multi-level data exploration (HATR8). These requirements are matched with specific VAL2 activities, like deep-dive analyses of incidents, weekly briefings with operational teams, and integration of the Dashboard into live operations. Success is measured by the operational uptake, improved situational awareness, and positive feedback from LLA safety personnel, rather than isolated lab metrics.

UC5's validation approach also contributes to broader HAIKU research goals by exploring the dynamics of human-machine teaming in safety-critical aviation environments. It illustrates how AI tools like the ASW can augment human judgment through enhanced visualization and situational insight, without requiring full transparency or explainability of underlying algorithms. Explainability is treated pragmatically-usefulness and usability take precedence over theoretical XAI, as users understand outputs through familiar workflows. Human factors from EASA guidelines (e.g., error tolerance, situational awareness, and decision-making support) are addressed through ongoing training, system design choices, and a HAZOP safety review. Overall, UC5 emphasizes the value of co-adaptation between human users and AI systems in operational contexts, advancing both safety outcomes and human-AI mutual understanding.

The research questions in UC5 were:

1. What data sources and analytic methods are most effective for capturing and assessing airport ground safety risks?
2. How can data-driven tools improve safety outcomes and reduce low-level incidents in complex airport environments?
3. What practices support the scalable and transferable application of safety dashboards across different operational contexts?

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

6.3. VAL2 Activities and Methods

The core aim of VAL2 was to demonstrate that the ASW Dashboard can support operational safety improvements through collaborative human-AI interaction.

During the HAIKU project, multiple stakeholder meetings were held to ensure the project's success was aligned with a practical, real-life tool. This approach will enable UC5 to reach TRL9 by the project's end, surpassing initial expectations. We have already validated the ASW system, demonstrating that analysing incidents allows LLA to manage airport signage effectively, thereby improving staff safety.

As a result, VAL2 confirms that LLA can operate the ASW system independently without our support. To guarantee long-term self-sufficiency, we developed a detailed user manual covering all operational steps, ensuring LLA teams can work autonomously and without external assistance.

The first key validation step for UC5 occurred when the new Safety Dashboard was presented to the LTN Safety Stack companies at a meeting in Luton on 18th July 2023.

The last one was the 31st LUTON SAFETY STACK MEETING, held on 16th January 2025 in Luton (London). The purpose of the meeting was to review 2024 safety performance, follow up on previous actions, share key learnings, discuss rebranding proposals, recognise achievements, and align on future standards and meeting plans within the Safety Management and Leadership framework.

The meeting had around 30 participants from national airport authorities, ground handling providers, aviation consultants, regulatory bodies, airlines, and technology service providers.

During the meeting, latest development is to focus on Runway Collision incidents, and data derived therefrom. A Runway Collision incident is defined as any incident involving a vehicle and another vehicle, an aircraft, or any airport equipment or infrastructure.

Figure 41 gives an outline of the validation journey for UC5 across the three years of HAIKU.

In Year 3, the final requirements have been accepted by the client (LLA), and a stable data pipeline process has been set up with weekly transmission of data to ENG and SUITE5 which then is used to update the Dashboard. This is used and monitored by LLA Safety in their principal Operations Room at LTN, and Airside Ops staff can download the Dashboard onto Tablets and have safety conversations with airport users (ground staff, controllers and flight crew) on hotspots, incident trends and contributory factors etc. as part of their daily safety duties at the airport.

In terms of providing safety insights, the Dashboard has helped identify specific hotspots and key stands and taxiway intersections where more incidents occur, and measures (signage and education) are having an impact on incident rates. As well as this direct impact, the Dashboard also helps Safety Stack partners understand which factors are not key, so that they can avoid spending resources on areas that are unlikely to have any material effect on incident rates.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

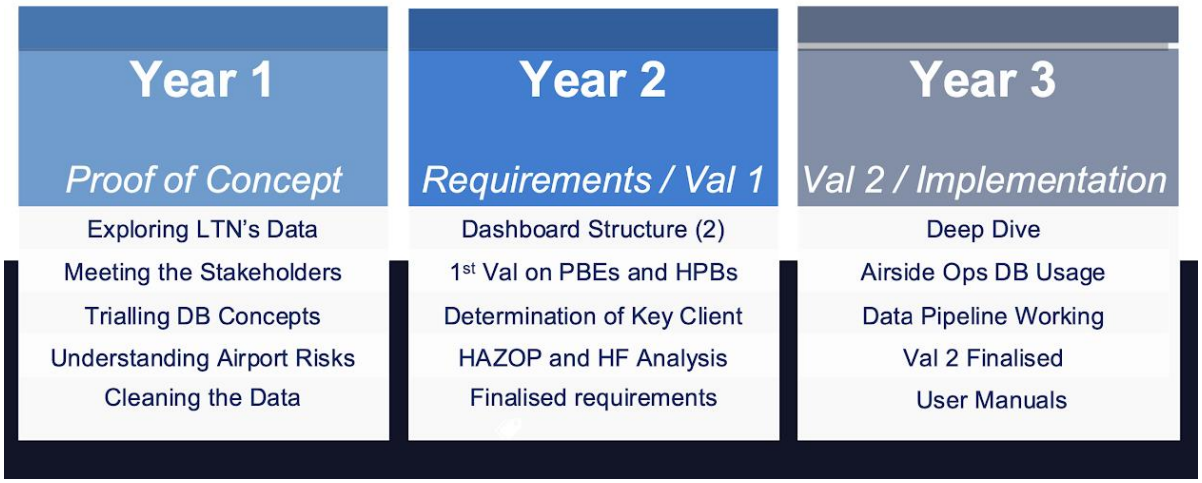


Figure 41. The UC5 Validation Journey during the three years of HAIKU

In terms of continual development and improvement of the ASW system, it has been agreed to integrate airside collision incident data into the Dashboard, as there are quite a lot of these, though they tend to be minor damage events. This is being worked on now and is scheduled for May 2025, but is not part of Val 2. Similarly, a Deep Dive exercise was carried out at LTN with safety representatives from four airlines, ATC, LLA and EUROCONTROL. This one-day exercise partially used the Dashboard and in particular its ability to zoom in to particular incidents, and see what factors were and were not in play. The Deep Dive exercise identified fresh insights and has led to improvements underway to reduce taxiway errors. Another Safety Stack, that of Dublin International Airport, requested a briefing on the Deep Dive exercise, as they have similar problems.

As with Collision data, the Deep Dive is not a centrepiece of VAL2, rather it shows that the ASW concept is not a once-only tool, and rather that it can be further developed to expand its scope in safety management and used flexibly in other safety initiatives.

The final step, which may occur after HAIKU is finished, is full integration with LLA's safety systems and infrastructure. This depends on certain developments within LLA likely to occur in the next 12 months or so and so is outside the control of HAIKU personnel. But to prepare for it, two manuals are being developed for LLA safety staff, one a user manual, the other a more technical manual related to the data pipeline and detection of problems 'under the hood'. Drafts of both these manuals have been submitted to LLA in January 2025 and will be updated and finalised in May 2025.

With the above developments, VAL2 is seen as complete and indicative that the ASW Safety Dashboard has been a successful HAIKU project, one that has supported tangible safety improvements at the UK's fifth largest airport.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

6.4. TRL Overview: update

Figure 42 shows the TRL progression and status at the end of VAL2. UC5 is developing an AI-powered capability to ingest and analyse a constellation of operational and safety event data collected across the airport, and to permit inferences and predictions regarding hot spots and safety 'pinch-points,' with a view to staying one step ahead on airport safety. The concept began as TRL4, and reached TRL9 on May 2025, before the end of the HAIKU project.

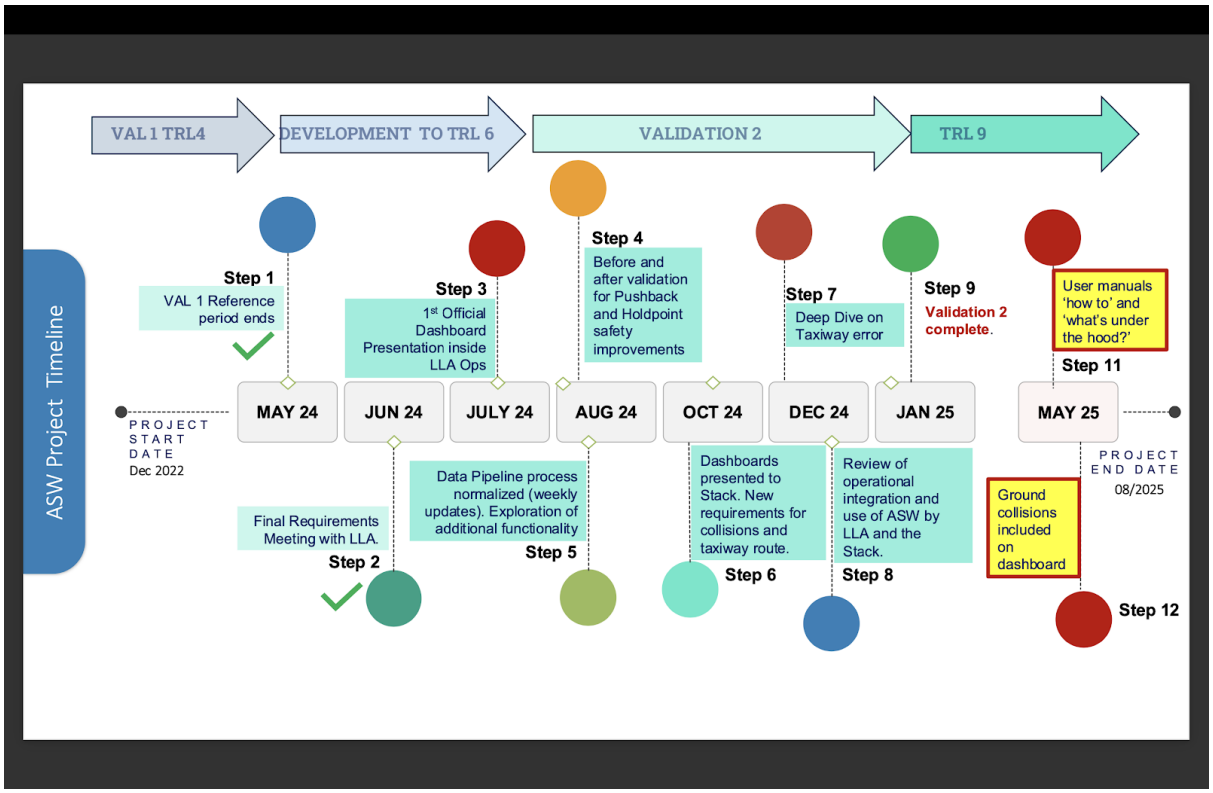


Figure 42. VAL 2 Tasks and Achievements

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

6.5. VAL 2 Results

6.5.1. Results Overview

Table 27. High-level results in relation to the EASA Validation objectives.

High-level Result	Brief description	Related validation objective (D6.3)
User manuals	Provide comprehensive user guides for effectively using the dashboard, with specific instructions for handling input/output monitoring. The manual is done; we are waiting for feedback.	EXP-18
Safety information delivery	Showing on the ASW information to users regarding unsafe operating conditions. The information is clear to the end-user, who finds it useful for getting a complete and clear picture of the situation.	EXP-19
Building awareness	Support the development of situational awareness for improved decision-making. An example, a particular hotspot on the taxiway system was identified. Changes were subsequently made to the airfield, including directional paint markings, new signage.	HF-01
Awareness reinforcement	Provide tools and feedback mechanisms to strengthen the user's situational awareness. An example, a particular hotspot on the taxiway system was identified. Changes were subsequently made to the airfield, including directional paint markings, new signage.	HF-02
Shared awareness development	Foster collaboration and shared situational awareness among team members. During the stack meeting, pilots among the airport staff, using the dashboard, they added the opportunity to discuss looking together the data available in a quick way.	HF-03
Decision validation	Enable users to submit their decisions for peer or system validation.	HF-04

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

	Due to the limited availability of data, which fortunately reflects a low incidence of events, AI and other ML alternative methods did not yield useful results for real-time decision-making. However, the ASW is now useful for post-analysis	
Abnormal operation management	Identify irregular operations to propose resolution strategies and outline potential consequences. An example, a particular hotspot on the taxiway system was identified. Changes were subsequently made to the airfield, including directional paint markings, new signage.	HF-06
Error in critical decisions	Detect and address poor decision-making in time-critical contexts. During the stack meeting, pilots among the airport staff, using the dashboard, they added the opportunity to discuss looking together the data available in a quick way	HF-07
Minimising Design Errors	Reduce the likelihood of user errors caused by design-related flaws. Using ASW, the information is clear to the end-user, who finds it useful for getting a complete and clear picture of the situation	HF-26
HAIRM design optimisation	Minimise the potential for design errors specifically related to HAIRM. During the stack meeting, pilots among the airport staff, using the dashboard, they added the opportunity to discuss looking together the data available in a quick way	HF-27
Error tolerance	Incorporate system features that demonstrate resilience and tolerance to user errors. The ASW shows any recurrent events, if they are	HF-28
Error detection opportunities	Provide mechanisms to help users identify and correct their mistakes. An example, a particular hotspot on the taxiway system was identified. Changes were subsequently made to the airfield, including directional paint markings, new signage	HF-29

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

6.5.2. Discussion

UC5 differs from other UCs, an experimental setup was not employed but we rather focused on an operational setup to reach the target of TRL9 at the end of the project. To do so, a systematic approach was followed. While Various challenges were addressed, ranging from the technical aspect of managing a large volume of general information to the limited number of incidents. However, this scarcity of incidents (big imbalance between the different classes of non-incident flights and incidents) did not facilitate the identification of common patterns.

What has been done was to use a systematic approach, starting with an analysis of the key objectives and requirements. This involved identifying the core goals and ensuring they were clearly defined and aligned with the intended outcomes. Once the objectives were established, detailed descriptions were developed to capture the essence of each goal, focusing on clarity, relevance, and actionable insights.

Subsequently, the information was structured in a way that ensured logical coherence and easy interpretation. Each goal was associated with a unique identifier to facilitate traceability and validation. Attention was given to maintaining a professional tone and concise language, ensuring the descriptions could effectively communicate the intended purpose to diverse stakeholders.

The process also involved cross-referencing related objectives to ensure consistency and eliminate redundancies. Iterative reviews were conducted to refine the content, addressing any ambiguities or gaps. The final output reflects a cohesive and well-structured representation of the results, enabling a clear understanding of their context and significance.

This section does not cover XAI results. Details and discussions of these can be found in D5.2.

The first key validation step for UC5 occurred when the new Safety Dashboard was presented to the LTN Safety Stack companies at a meeting in Luton on 18th July 2023. This led to several safety insights as summarised in Table 28 below, as well as renewed efforts to move from TRL4 to TRL7 and beyond, since LLA decided they wanted the Dashboard to become a permanent fixture in LTN's safety management architecture.

Since the July 18th, 2023, meeting, signage has been improved at the two Stands mentioned, as well as at one key point on the taxiway system close to the runway, as it was recognised that it was not clear to all pilots (many of whom fly to LTN infrequently).

Some interesting factors raised by the analysis are still being explored, e.g. there appear to be more events when the runway is being used in the less frequent direction (since aircraft take-off and land facing the wind, if the prevailing wind shifts, the direction of take-off/landing may be switched 180 degrees). Operation in this less frequent mode may be more productive of incidents.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Table 28. Safety insights in relation to safety issues at LTN.

Issue	Discussion	Partners
Incorrect Taxiway Selection	It was remarked that although LTN is not a highly complex airport with multiple runways etc., it does have a relatively high number of junctions, which can perhaps lead to confusion or perception errors about where aircraft believe they are and where they should go next. In some larger international airports, they operate a 'follow the green' system, though it is not clear that LTN could adopt such a system. In the future however, the airport will gain an ASMGCS (Airport Surface Movement Ground Control System) which gives a live-updated map of the airport surface and all aircraft (and some vehicles). This may also be augmented by CCTV particularly around 'hotspots' and to those areas that are difficult to see from the Tower.	Airlines, ATC, LLA
Hold-point Bust	There was some discussion of hold-point busts and practices at other airports. Several European airports these days have 'zones' which the aircraft crosses into and where it waits, rather than a line that the aircraft should not cross. Pilots more familiar with these zones or areas may inadvertently cross over a holding point, thinking they are supposed to enter a zone.	Airlines, ATC, LLA
Pushback Error	Stands 62 and 71 were highlighted by the data analysis presentation as being more prone to pushback error. Partners noted that Stand 62 has no sign, which might contribute to error rates (it is for business jets rather than commercial jets, and many business jet pilots are unfamiliar with LTN's layout etc.). Stand 71 is at the end (a cul-de-sac) and might not be as well signposted as other stands. For pushback error, it was also noted that stands that occur on a bend can be tricky. In some airports the pilots no longer control the direction in which they are pushed back, and it is left to the ground handlers. One or two Stack partners could see the advantage of this, as the local staff are more familiar, and there can be misunderstanding when communicating with flight crew about what is left or right. The best form of instruction was also discussed, as to whether it should be 'left' or 'right', or a compass reference (e.g. East, South, etc.). One further suggestion was to push back to a landmark, which could be a clearer and less confusing form of instruction.	Airlines, ATC, LLA, Business Jets

In terms of pushback errors, these have reduced at LTN recently, due to continuing efforts by LLA Airside Ops people to educate staff, and over the last couple of years they have introduced Tug Drop-off Points.

For Hold Point Busts, reduction in these is seen, again partly due to efforts to educate staff. ATC have also been conducting more 'defensive controlling', anticipating when an aircraft may bust the hold and acting accordingly. This latter point has been influenced and supported by the ASW Dashboard and its ability to zero in on incident hotspots, as shown in Figures 43 and 44 below.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Figure 43. Overview display (ASW Dashboard in 'Historic' mode) highlighting incident hotspot locations on the taxiway system.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Figure 44. ASW Dashboard 'Zoom-in' function for closer investigation of hotspots

As an example, a particular hotspot on the taxiway system was identified. Changes were subsequently made to the airfield, including directional paint markings, new Delta/Foxtrot signage (see Figure 45) on Alpha. NATS have also begun more 'defensive controlling' in relation to the Alpha/Delta/Foxtrot intersection. There have been two incorrect taxi events in the eight months since the new signage has been installed, compared to six events in the previous eight months.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Figure 45. Signage changes at taxiway intersection

The dashboard was also very helpful to the Stack in ruling out potential factors, for example they thought that these incidents occurred at busier times, however the dashboard showed that it was during the quieter times (Figures 46 and 47).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Figure 46. Flight Congestion and Delay view

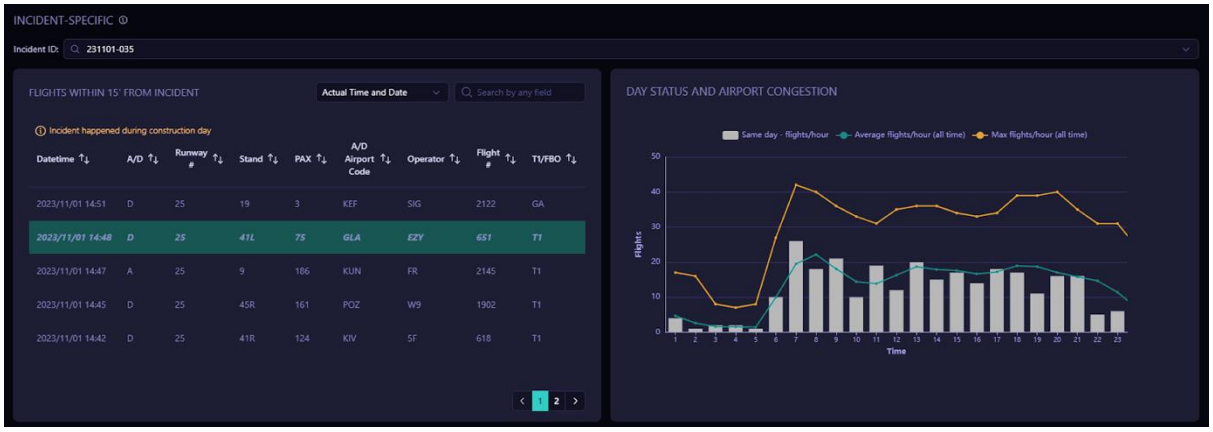


Figure 47. Zooming in on flight congestion as a factor for particular incidents

Similar results were obtained when looking at other 'usual suspects' for incident causation, such as weather (in particular low visibility), also found not to be a driving factor.

6.6. UC5 VAL 2 Conclusions

UC5 employed a systematic and structured approach, rather than a classical experimental design, to address the outlined goals. This approach was necessary due to the unique challenges of the UC, particularly the imbalance between the abundance of general operational data and the relatively low number of safety incidents. Despite this, significant value was derived through a methodical definition of objectives, followed by detailed, traceable documentation that facilitated alignment with stakeholder expectations.

Data captured in UC5 was highly dependent on the reports that staff is filling once an incident occurs. The structure and the content of these reports include numerous parameters that help in the analysis

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

and identification of airport ground safety risks. Other data sources used have to do with flight operations data, which are also used in conjunction with the reports to deliver situational awareness of the whole airport setting during the occurrence of an incident. Nevertheless, as discussions pointed out, other types of data, not currently available, such as exact taxiway coordinates of aircrafts would prove beneficial too.

The development and deployment of a Safety Dashboard emerged as a critical outcome, enabling real-time insights into safety-relevant locations such as taxiway intersections and stand areas. The dashboard not only helped identify key hotspots where incidents tend to cluster but also highlighted where current safety interventions (e.g., signage and education) are making measurable impacts. Importantly, it also clarified which factors have limited influence, guiding the staff to focus efforts and resources more effectively. The outcome of UC5 is a direct answer to the question on how data-driven tools can improve safety outcomes and reduce risks, as its value has been proved in the LLA operational environment, and the dashboard itself has been presented in the LLA Safety Stack meetings as a new tool to aid the work of the stack.

The final step, which may occur after HAIKU is finished, is full integration with LLA's safety systems and infrastructure. This depends on certain developments within LLA likely to occur in the next 12 months and so is outside the control of HAIKU personnel. To prepare for it, two manuals have been developed for LLA safety staff, one a user manual, the other a more technical manual related to the data pipeline and detection of problems 'under the hood'. Drafts of both these manuals have been submitted to LLA in January 2025, updated and finalised in May 2025.

Last but not least, it is worth saying that the dashboard developed, although customised to the needs and requirements of the LLA case, can be easily replicated and deployed also in other airport settings, and accommodate other types of incidents as well. The underlying technological infrastructure is agnostic of the incident types recorded, while the methods and data processing algorithms used can be adjusted to fit the needs of other operational contexts.

6.6.1. UC5 Research Questions

In the context of the UC5, the research focuses on validating the real-world utility of AI systems within airport safety operations. A central research question is: *What is the core goal of the VAL2 activity?* The answer lies in demonstrating how the system can effectively support operational safety through human-AI collaboration. Rather than relying on isolated performance metrics in laboratory settings, VAL2 emphasises applied functionality within realistic operational environments. This includes integrating the system into live workflows, where human operators and AI tools collaborate to enhance situational awareness and safety outcomes.

Closely related to this is another key question: *What is a key function of the ASW dashboard in the context of VAL2?* The ASW is designed to be more than a data visualization tool, it plays an active role in enabling operational safety improvements through intelligent interaction between humans and AI. Its functionalities include historical data processing, exploratory "what-if" analysis, and multi-level data exploration. These features ensure that safety personnel can make informed decisions quickly

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

when needed, while also allowing for more in-depth analysis of the data at a later stage, thereby fostering a proactive and collaborative safety culture within airport operations.

6.6.2. HAIKU High-level Research Questions

How UC5 serves the high-level research goals of HAIKU.

HAIKU has posed three high-level research questions, as below, with preliminary answers from the UC5 perspective.

HAIKU Q1: What is the recommended human-AI relationship for each of the different AI aviation applications?

UC5 shows how the product of humans and AI-based systems can lead to better 'safety intelligence' than either alone. The ASW Dashboard has shown that factors traditionally thought to be dominating incident occurrence (e.g. weather, traffic volume etc.), are not actually dominant; rather, it is either more local factors or the occasional confluence of certain other characteristics that make incident occurrence more likely. The Dashboard gives the human end user a broader and deeper safety landscape to explore, effectively enlarging their operational safety situation awareness. Overall, UC5 shows 'human-machine intelligence' where the human is still in control but has augmented capabilities due to the enhanced data visualisations.

HAIKU Q2: What does it mean for AI to be explainable?

In UC5 there is no AI where advanced explainability is required. The ASW Dashboard instead allows significantly more exploration by the human end-user.

HAIKU Q3: How to train AI to assist humans in safety critical tasks when training data are insufficient?

Human-in-the-loop learning can lead to more explainable AI (XAI), as the user will have more visibility on the underlying mechanisms, ideally steering a continuous learning process. It can also lead to 'personalised AI', with individuals and teams developing AI assistants on the basis of their needs, abilities, limitations and experiences.

For UC5, the real learning that has taken place is between the operational users and the data scientists, who have come to a shared understanding (and workspace – the Dashboard) of the data and its exploration possibilities for the purposes of safety. The Data Scientists have learned what to focus on and what matters, as well as how safety works at an airport. The Operational people have learned more about AI and data science, how it works and its limitations, but also the critical importance of data volume and quality.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

6.6.3. Research Recommendations

Future work should focus on enhancing the granularity and predictive power of the dashboard by integrating new data sources, including human factors data and environmental conditions.

A standardised methodology for structuring objectives and traceability could be replicated in other airports with similar constraints.

Additional research is needed on validating insights from low-incident environments and exploring how to simulate or augment datasets to support more robust safety modelling.

Collaboration with Safety Stack partners should continue, especially in refining the dashboard's utility for real world decision making and operational training.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

7. Use Case #6 – Airport Spreading Virus Prevention

7.1. Deviation from Validation Plan (D6.3)

VAL2 was performed according to the validation plan described in D6.3 with the following deviations:

- Ranking and question population from answer were tested
- Sensor data was shown effectively to the explainability CLT levels.

7.2. Validation Objectives

The main research questions regarding UC6 are the following:

- Can the passenger be effectively routed to the airport common places that are less overcrowded using an IA
- Can the system present its routing guidance in a manner that is understandable to the passengers?

The main objectives of this UC are given below:

- Design and develop an Android application that is user-friendly and intuitive for airport passengers.
- Ensure the routing algorithm effectively uses weighting factors to generate routing sequences.
- Enable the chatbot to accurately respond to passenger queries, dynamically add questions based on passenger input, and incorporate relevant sensor data into responses.
- Establish reliable transmission of data from wireless cameras to the cloud server, enabling real-time readings to be used by both the algorithm and chatbot.
- Implement XAI functionality to present results clearly across four levels of abstraction, following CLT.
- Gather and evaluate passenger feedback regarding the app's usability.

7.3. VAL2 Activities and Methods

The main activities of the VAL2 were the correct acquisition of the data coming from the sensors (both commercial and prototype). Moreover, the chatbot was tested with respect to the validity of the responses.

There were 10 participants in VAL2 with several persons actively going from and two the points of camera sensor installations.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

The was a primary scenario where the validity of the routing was performed with the participants using the application. Since the Amygdaleona airport in Kavala is a school for pilots there were undetermined persons walking through the installation places.

The experiment was 2 hours showing the application to the participant, 1 hour testing, and approximately than 30 minutes.

The study structure followed this logic: for each Objective set (e.g., *Situation Awareness*) a research question was identified (e.g., *whether the system can adequately provide routing recommendations based on sensor data*), we identified associated requirements (e.g., HF-01: System must be able to generate its own situation representation). A targeted question was then asked to assess the requirement—for example: *How important is it for you to understand how the HAIKU app makes decisions*?

7.3.1. Participants

There were 10 participants, 8 graduates and 2 that were not graduates and non-graduates. The participants were those persons that actively used the application. Their roles are primarily engineers and scientists. We could not engage any pilot candidates. 6 male and 4 female participants were engaged, with 6 being between 30-40 years of age and 4 between 40-50. Their expertise ranged from graduates to not graduates or undergraduates. The persons also used the health and safety portal; however, since this is an extra task and it does not engage HAT, it was not included in the questionnaire

7.3.2. Simulator/Apparatus

There was a single scenario of the procedure at the Amygdaleona airport in Kavala. Note that four cameras were places, one inside the offices, one at the former coffee place, and two near to the doors where persons were coming in and out. This was the place where the most traffic was observed, and it was selected for evaluation. The number of the persons were recorded from the cameras as observed by the XAI responses and verified by the database insertion. More specifically, the cameras recorded the number of individuals, and the data were processed by CLT levels 3 and 4. CLT3 calculated the weighting factor and generated an XAI message for the passenger, while CLT4 delivered near-real-time responses supplemented with sensor data. The appropriate counting of the persons was cross-examined with the values of the XAI or the question to the chatbot regarding whether the common area was busy.

The main objective was to validate the responses and the counting of both the commercial cameras and the prototypes. Hence, we requested that the movement in the premises was as controlled as possible.

The participant was checking her mobile phone application when a specific number of persons were moving in or out and the routing sequence from these two places of installation. Moreover, the Q&A

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

from the chatbot and the XAI CLT levels were validated by the participant. Other users were utilised to simulate traffic.

The IA was providing the routing sequence and the likelihood of infection in real-time and the chatbot AI system was providing near real-time values from the metrics used for the routing sequence upon request from the user. Moreover, the CLT levels were evaluated since they provided automated messages to the Android application every 15 seconds for evaluation purposes. The ranking was also simulated by multiple stating to the chatbot that a specific common place is not busy, while in the first question it states that it was busy. This validated our ranking field in the knowledge base. Finally, the insertion of a question was validated by making a comment that was not linked to a question in the knowledge base and the question was then inserted to it by checking the respective row of the csv file. Rerouting was also investigated. It is just a repetition of the procedure, which means that the CLT levels were outputted automatically again with updated values.

7.3.3. Procedure

The application was firstly introduced to the participants in its operation responses, and a demo was provided to them to familiarise with the application of the IA. The Android application was explained with respect to its responses and operations. Moreover, the participant was instructed to go to the premises which included two sets of camera sensors each close to a door. The participants were suggested to go through each door (in and out) and check the application on the phone.

The preparation included the checks in the database to investigate the correct person counting and the preference sent to the server again via the database.

The execution was primarily the correct counting of the persons and the correct output of the routing. This was validated with the explainability responses in the chatbot. The questionnaire was used when debriefing the participants.

7.3.4. Data Collection Tools

Table 29. Data collection tools used in UC6.

Tool	Objective	Type of Collected Data
Commercial cameras	Routing weighting factor metric of occupancy. Likelihood of infection data field	Quantitative
Prototype cameras	Routing weighting factor metric of queue. Likelihood of infection data field	Quantitative
Android application preferences	Routing weighting factor of persons going towards this direction. Respective data field for infection likelihood	Quantitative

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Temperature, humidity, tVOC and eCO2 measurements	Provision of a dataset for air quality measurement	Quantitative
Routing sequence	Determination of the routing sequence	Qualitative coming from sensor values
Likelihood of infection	Determination of likelihood from random forest	Qualitative from ML output

7.3.5. Data Analysis

The primary data of the UC6 IA is the person count from the commercial cameras, the prototype camera system for queuing and the preferences metric from the Android application. The data was obtained from the following operations:

- Commercial cameras provided the measurements (inbound/outbound) by connecting with the respective API provided by the companies.
- The wireless camera prototypes provided the queuing measurements using computer vision software and network programming.
- The preferences were updating the respective database field upon connection with the loud web service by the application.

Since infection likelihood was estimated using a well-established Random Forest classification method, the most recent set of measurements was used for classification. Ther Random Forest operated with real world data coming from the four installations. The rest of the fields(columns) were populated with random data (we have 9 shops). This approach was rigorously tested during implementation, first in simulation and later with real-world data, consistently yielding accurate results. Note that the pattern injected to the data did not allow high likelihood to be present due to small number of persons; hence it was only checked in simulation.

The routing sequence is essentially a rule-based AI model with deterministic values coming from the camera sensors and the preferences. Hence, no statistical models were required and only the cross-examination of the inserted values with the routing sequence response was sufficient. The data from the sensors was investigated by selecting the API calls with the database tables and regarding the occupancy the last field was added to the new measurement, to maintain the occupancy and not just the inbound/outbound values.

In terms of air quality data, the measurements were initially checked for validity based on tools available from the vendors and appropriate bounds of the values. Given the relatively small sample of air quality data, no statistical analysis was undertaken.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

7.4. TRL Overview: update

The TRL of the UC6 has not been changed and it appears below:

Table 30. TRL progress in UC6 for the COVAID tool.

Components: COVAID tool of Android application with AI model and hardware prototypes for person queues and indoor air quality sensor board and system.		
TRL	**Month**	**Activity to reach selected level**
1	1	Literature Review and initial theoretical formulation
2	9	Theoretical foundation of solution and initial implementation of subsystems
3	16	VAL1 with initial prototype. Injected data and very basic information from staff members regarding implemented subsystems
4	30	Finalised prototypes and subsystems. VAL2 at the Amygdaleona airport in Kavala Greece with multiple passengers
5	36	Acquisition of more data. Fine tuning of solution to reach TRL5
6	-	

We do not believe that the UC6 IA will reach TRL6.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

7.5. VAL 2 Results

7.5.1. Results Overview

Table 31. UC 6 high-level results in relation to the EASA Validation objectives.

High-level Result	Brief Description	Related Validation Objective (D6.3)
The system was able to record the sensor reading periodically and populate the database which in turn displayed the values to the Android app. The routing weighting factor was displaying the sequence effectively.	This objective essentially keeps the crowd patterns by keeping all the data to the database and providing near real-time results. Routing recommendations are given based on the current measurements as well as for re-routing. Validation was produced by calculating the weighting factor by installation, and the readings using the web API and the database insertion. Also, experiments have been performed at the airport as validation activity. User non-compliance is implicitly handled with the measurements of the sensors.	HF-01
The database is storing the inputs from the camera sensors and the preferences in timely manner. The system provided ranked recommendations based on passenger input. Explainability also provides weighting factor values and sensor data for robustness of the IA and not to be a black box.	These were validated again with in-situ experiments at the airport and cross-referencing with real-time prompts and database values. The explainability also was validated in the same manner using real persons.	HF-02
The entire recommendation for routing actively engaged the passengers with their presence and their preferences. Rerouting also takes place with current measurements.	This does not take place explicitly. The presence and preferences have been validated with the use of the Android application and the database values being incremented. Re-routing also took place using the same manner with different persons within an installation.	HF-03
Routing non-compliance results in the application providing results that they appear in the next measurements. Hence, rerouting corrects error and non-compliance.	These was easy to validate since persons were actively going around the airport since two of the installations were placed in the candidate pilot engagement building. Hence, non-compliance was effectively simulated.	HF-28

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

7.5.2. Discussion

As shown in Figure 48, participants varied in reported AI product usage. It is evident that the majority use AI products daily, followed by rarely and weekly.

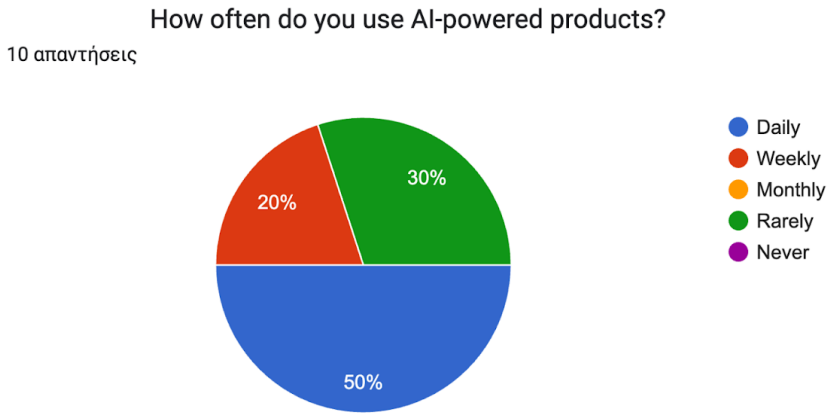


Figure 48. AI frequency usage

Thereafter, the entire participant sample found that the HAIKU app was easy to use. Another interesting finding was the factors that led the persons on why to utilise the app. We see that convenience, and accuracy and efficiency exhibited the highest percentages with 80% and 70% respectively. Transparency and explainability was selected by two persons while personalisation and trust in AI provided has only one answer as can be seen in Figure 49.

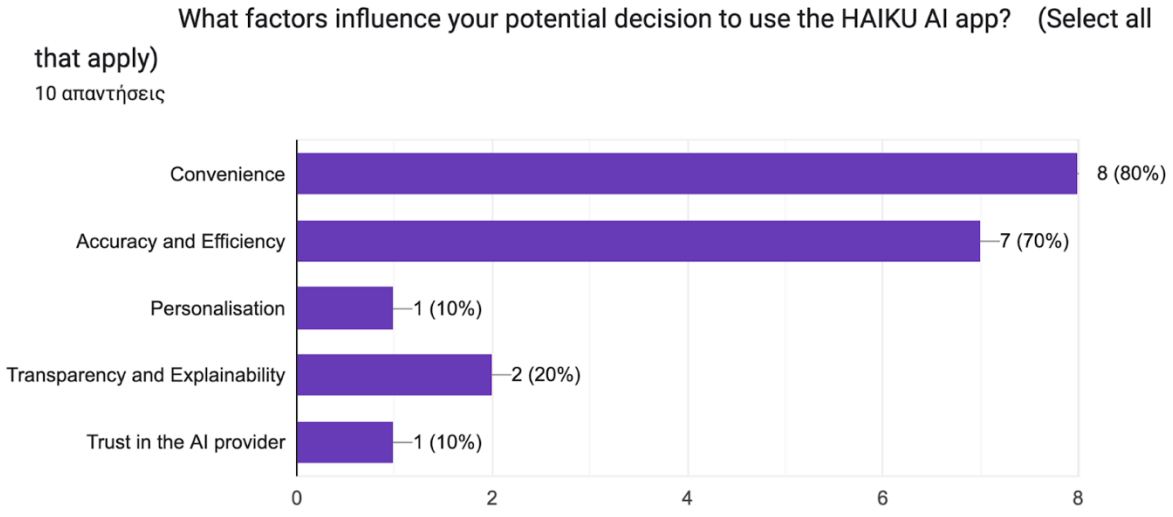


Figure 49. Factors of HAIKU app usage

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

We also asked the participants whether they would use the app for routing at the airport with 8 responding yes and one not sure. Moreover, we asked participants whether they are familiar with XAI. The results are varying as we can see in Figure 50. However, they all mentioned that the explanations were sufficient. As we can see from Figure 51, in support to the above that participants showed relatively high understanding of the weighting factor that appears in the XAI of the app. Furthermore, after the explanation of the app to the persons, 90% mentioned that the timing manner of the explanations will safeguard their decisions throughout the journey at the airport and 10% was not sure. Of course, they have been told that the 15 second interval of the explanations were to be served at different time intervals based on the CLT literature.

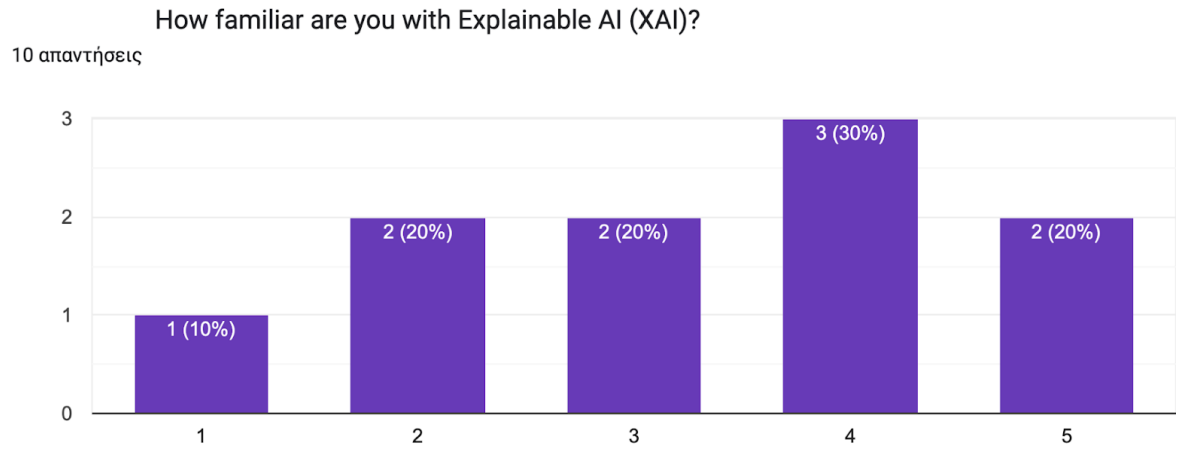


Figure 50. XAI Familiarity

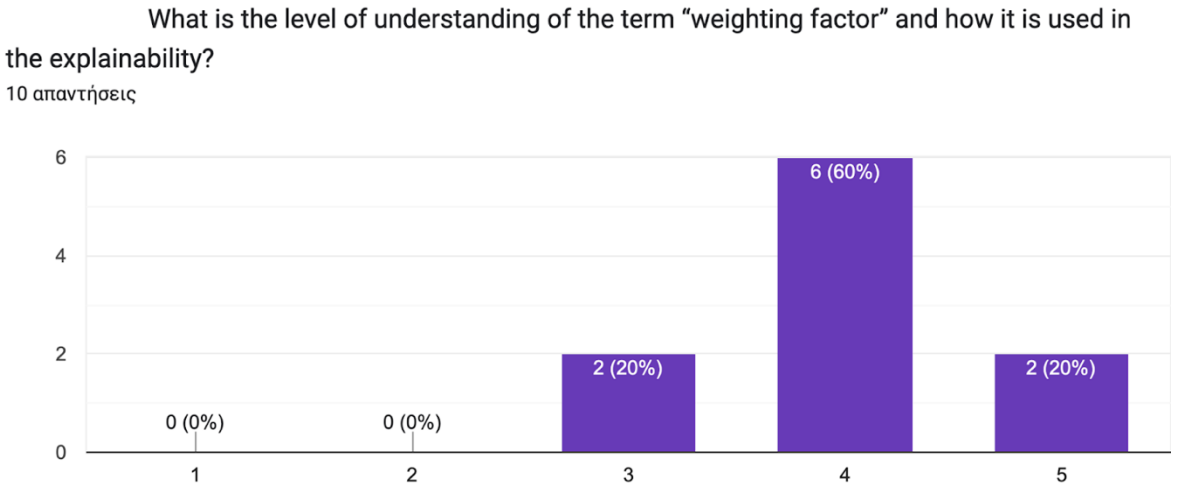


Figure 51. Weighting factor Familiarity

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

In Figure 52, the participants showed a high level of understanding of the contribution towards deciding in the app. This took place after the discussion. It seems that a help screen is evident to place the passengers in the loop and to make them engage with more confidence. Moreover 90 % of the participants found the ranking of the application very useful and 10% were not sure. This gave participants the level of contribution towards the result of the chatbot.

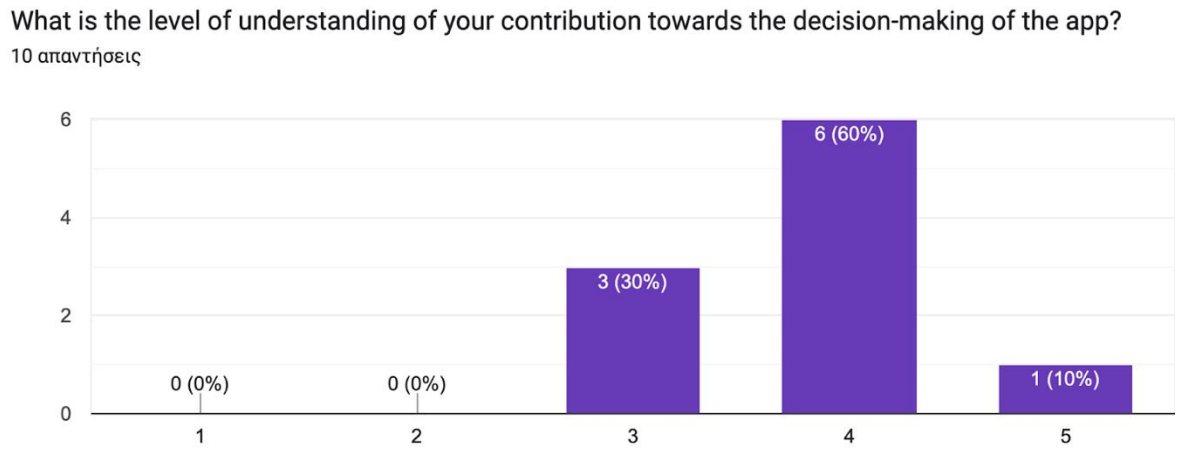


Figure 52. Level of contribution understanding of app usage

Figure 53 shows the level of importance of placing comments to the chatbot by the participant. This provides high level results since most participants seemed to understand their role in interacting with the app. The HAIT component seemed to be in favour of the participants.

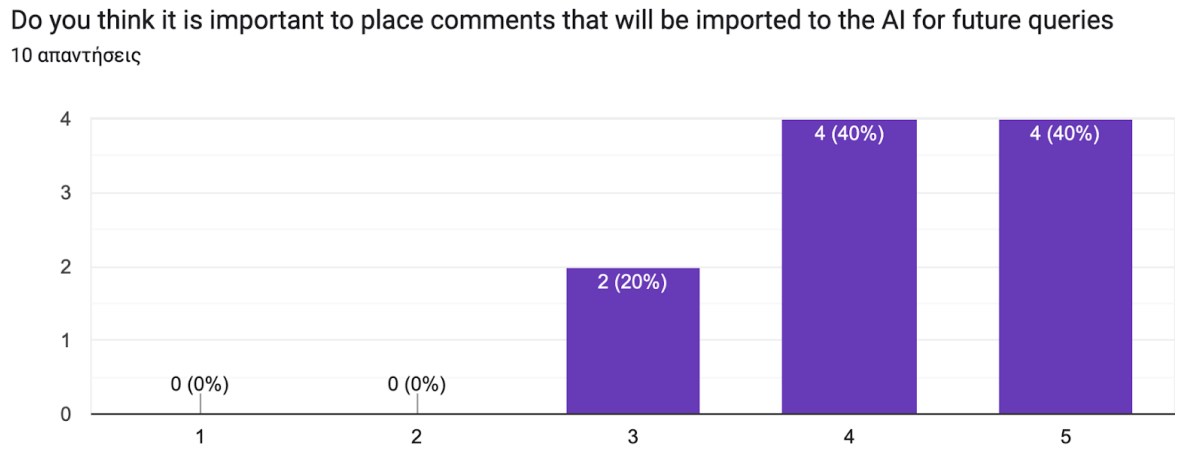


Figure 53. Comment insertion importance

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

In Figure 54, a significant finding was found whereby 1 participant mentioned that the level of understanding of the decision making of the HAIKU app is not that significant. Three of them mentioned neutral while 7 selected high importance.

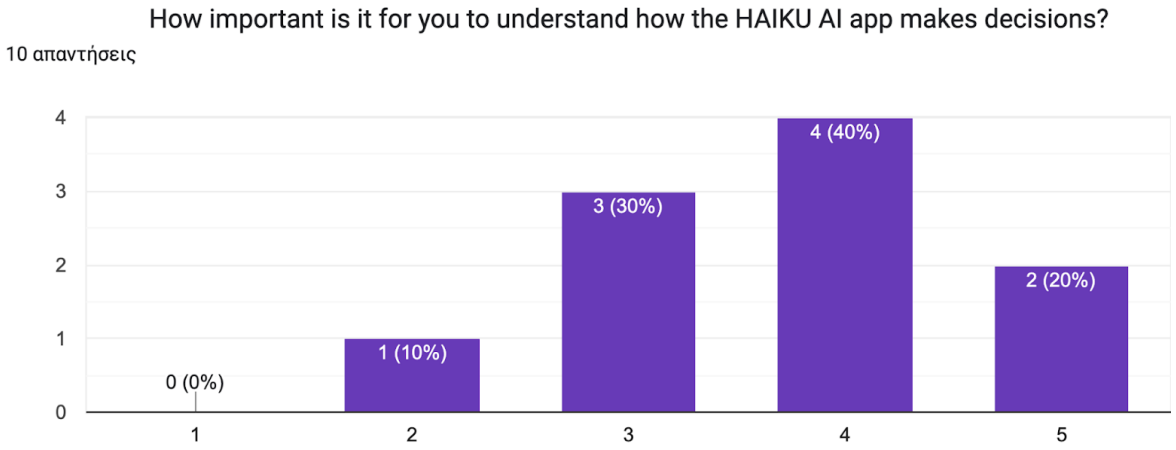


Figure 54. Decision making importance of app

Finally, one participant observed bias in the app results, two of them were not sure and seven did not notice any form of bias. To conclude, the participants were asked about the improvements about transparency and explainability of the COVAID. They responses from nine participants are given below:

- Provide insights into model decisions
- I have nothing to add.
- N/A
- More symbols and less text
- add more images
- I do not have the experience to answer that question
- Short reasoning in replies wherever it applies
- None
- None

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

7.6. UC6 VAL 2 Conclusions

7.6.1. UC6 Research Questions

The primary research questions of the UC are

- whether we can build a system that will successfully route persons in an airport common place using an IA.
- whether the system will present the results in a manner that will be easy to comprehend by the passengers.

The first research question was addressed via a limited validation because the cameras and the sensors were installed at an airport whereby candidate pilots were trained, and it was difficult to perform a choreography for controlled traffic. However, the persons participating in the validation provided some insightful responses to the questionnaire. The second research question focused on the XAI of the IA whereby simple approaches have been utilised to be included and displayed in the Android app. Again, some questions and improvements have been provided by the participants.

The main conclusion of the first research question is that the system performs adequately with the note that the data from the sensor is near-real-time due to restrictions of the hardware and the fact that it is not feasible to query the sensors upon individual requests by the passengers. The second research questions adhere to a basic understanding of terms like the “weighting factor” that is explained in the XAI messages and the fact that they participate in the results with the comments and the answer ranking. The XAI is kept as simple as possible for the passenger to be able to get their hands on real data.

7.6.2. HAIKU High-level Research Questions

The HAIKU high-level research question that directly impacts the UC is whether we can build an IA that will promote HAT in our UC. The answer is yes since passengers can interact with the IA by placing preferences to the weighting factor and influence the routing sequence. Moreover, the XAI which is based on the CLT levels promote explainability of the built system. Essentially, the persons acting as passengers influence the system and can change the routing sequence. The simplicity of the Android application was at the forefront of the UC, since we did not wish the passengers to be overwhelmed by unnecessary complexity. The implementation complexity of the backend did not appear to the participant; however, it required multi-threaded communication and polling data from different devices which increased the complexity of the program in terms of synchronisation.

The main lessons learned are the following:

Lesson 1: Wireless cameras, both commercial and prototypes, exhibited proper values being tested with real persons during tests. The database values were appropriately populated.

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Lesson 3: The infection index classifier delivered promising results in detecting relevant patterns. The injected pattern was effectively recognised in simulation and real tests.

Lesson 4: The passengers' input to the Android app was limited to 10. However, due to the liveness of the airport and the coming and going through the installations provided interesting tests for the entire system. To the level possible, passenger traffic was controlled even though it was quite difficult.

Lesson 5: The selected classifier for air quality forecasting performed satisfactorily, though it occasionally required human intervention to ensure accurate forecasting. The data coming from the installations was placed in a web interface and the values exhibited normal indices depending of course on the environment which is not an industrial or extreme case one.

Lesson 6: The chatbot for passenger communication and explainability of routing recommendations was effective in real-time interactions. Some of the passengers needed further explanations, which a suggestion (internal) was to individualise the XAI according to the level of expertise by the passenger.

Lesson 7: The weighting factor used to recommend optimal routing performed well in simulations, balancing multiple variables effectively. The weighting factor was validated by performing calculations and showing that they were correct.

Lesson 8: Distance and queuing data provided valuable inputs for calculating the optimal route. The social distancing has been performed in simulation using a video and it showed good results. The installations did not provide us with the flexibility to place more cameras there. Note that the distancing was made in 2D since Lidar is quite expensive to purchase and place one at each installation.

7.6.3. Research Recommendations

Further research directions are to utilise the cosine-based similarity for comparison with the Jar-winkler similarity for the retrieval-based NLP of the chatbot. Further, we are currently implementing a custom Large Language Model (LLM) which will further improve the answers of the chatbot. Moreover, flutter will be used to encapsulate all the different OSs available among mobile phones. Custom explainability based on passengers' level of understanding will also be investigated.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

8. Conclusions

The consortium's three-year experience across six diverse use cases - complemented by cross-cutting work streams on societal acceptance, safety culture, workforce and skills, Human Factors assurance process, liability and ethics - has provided valuable insights that helped address the project's high-level research questions and led to the development of key insights for future research.

HAIKU Q1, ***"What is the recommended human-AI relationship for each of the different AI aviation applications?",*** explored the appropriate role and level of autonomy for AI systems.

According to the VAL2 findings, the optimal human–IA relationship proved to be **highly context-dependent**. It must be tailored to the specific operational environment, user role, and task complexity of each IA application in aviation. Rather than applying a one-size-fits-all approach, the HAT model should be defined through a **structured**, **user-centred design process**. This process should begin with a deep understanding of human needs, operational challenges, and the value the IA can bring within a clearly defined ConOps. Early in development, it is essential to explicitly define roles, responsibilities, and interaction modalities between humans and IA. This includes detailing the capabilities and limitations of the AI system, designing task allocation strategies, and ensuring continuous end-user involvement through iterative development cycles. The HAIKU project demonstrates that regulatory frameworks, such as the EASA HAT requirements, can serve as useful starting points. However, the design of effective teams should not be constrained by regulations alone. Instead, it must remain flexible and responsive to operational realities.

In the aviation environments explored, HAIKU findings emphasize that **IA can adopt different roles**, ranging from assistive to collaborative, to support rather than replacing human judgment. In these contexts, trust is established not only through reliable system performance but also through human-centred design choices, such as using attention-guidance mechanisms, voice-based communication, and abstract (rather than overly complex or spatially distributed) information formats. These design features help reduce operator workload and enhance situational awareness.

AI systems can also serve as a second set of eyes, providing an **additional safety layer** for conflict detection and risk awareness. Importantly, AI should function more like a negotiable partner, capable of interactive dialogue and adaptive responses, rather than a static tool.

Furthermore, across all operational contexts, the following key enablers of effective HAT emerge as crucial for fostering trust, acceptance, and long-term integration of IA in aviation workflows:

- IA capability to provide **personalised support**, according to user preferences and cognitive styles.
- IA capability of **progressively adapting** to the user's evolving status and degree of familiarization.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- A carefully designed approach to IA interactions - particularly its use of **soft skills** such as phrasing, timing, and tone of voice in voice-based communication, as a key factor in shaping the perception of the AI within the team.

HAIKU Q2, *"What does it mean for AI to be explainable?"*, explored the form, depth, and necessity of explainability in building operator trust and understanding.

According to the HAIKU results, AI is considered explainable when users can form a **clear and sufficient understanding** of what the system is doing, why it is doing it, and its contribution to achieving shared goals. This is particularly important in high-stakes contexts like aviation.

Explainability involves aligning the AI's internal logic with the human user's mental model, helping users **feel in control and accountable for system-supported decisions**. It must be **tailored to different user roles** (e.g., pilots, supervisors, developers) and operational scenarios, considering factors like task complexity, cognitive load, and time sensitivity.

It is furthermore important to highlight that effective explainability extends across the entire process - not just during operations: before use through **training**, during use via **context-based cues** (when appropriate), and after use through **debriefing**.

Focusing on the operative phase (during use), explainability should be designed as a **dynamic and interactive process**, allowing users to access varying levels of detail based on their needs, constraints, and familiarity with the system. Drawing on frameworks like Construal Level Theory (CLT), explainability should offer **layered access to information** - ranging from very simple visual cues (e.g., symbols) or brief summaries to in-depth justifications - enabling users to build understanding progressively without becoming overwhelmed. However, explanations are only effective when the system itself performs reliably; clear reasoning alone cannot make up for poor AI performance.

Ultimately, it is important to conceive explainability not just as a feature, but as a proper **trust-building mechanism** that must be thoughtfully integrated into system design to foster trust, confidence, accountability, and effective HAT.

HAIKU Q3, ***"How to train AI to assist humans in safety critical tasks when training data are insufficient?"***, explored the challenge of data availability as a crucial factor in AI system development.

Drawing from the experience gained in HAIKU, it is essential to conduct a **preliminary feasibility assessment** by asking a fundamental question: given the available data, can the AI model be adequately trained? The answer is nuanced and depends on the specific context.

In aviation, as in most safety-critical domains, there is a common tendency to focus AI training on data related to negative events. However, such data are relatively scarce, limiting the potential to develop robust AI models. To overcome this limitation, it is valuable to **incorporate data reflecting positive outcomes and human performance.** Although this approach increases complexity and requires intensive labelling efforts, it expands and diversifies the training dataset.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Another promising avenue to address data scarcity and improve model robustness involves **data augmentation techniques**, such as synthetic data generation and simulation environments. While these methods remain research topics, they hold potential for enhancing AI performance.

Furthermore, it is important to tailor the role of AI-based systems not only on data availability but also on **task criticality**. For safety-critical tasks where data is insufficient, caution is paramount. In these cases, AI is better suited to support processes and perform data analysis rather than making autonomous decisions or recommendations. Final decision-making and creative problem-solving should remain human responsibilities, with rule-based algorithms serving as complementary tools.

Ultimately, it is crucial to **avoid hype-driven deployment**: AI should be implemented only when it can be effectively trained on suitable data and offers clear benefits. Otherwise, alternative, non-AI solutions may be more appropriate.

Moving beyond the HAIKU high-level research questions, several areas are identified as immature and represent important research gaps, thus offering significant opportunities for future studies.

- Further effort is needed to **strengthen AI decision-making capabilities**, particularly in areas like negotiation, adaptiveness, and diagnostic reasoning (EASA Req. HF-04 - HF-06, HF-09 - HF-10).
- **Natural language and multi-modal communication** - such as spoken interaction, gesture recognition, and visual cues - is an area that deserves further exploration, to better support natural and intuitive human-AI interaction (EASA Req. HF-11 - HF-17, HF-18 - HF-21, HF-22 - HF-24).
- Explainability should further evolve to enable users to **customize the level of detail in AI explanations** (EASA Req. EXP-14, EXP-18 - EXP-19). Furthermore, investigating voice-based and multimodal interaction modalities could help reduce cognitive workload and improve operational explainability. Efforts should also focus on developing dynamic and multi-format explainability systems that adapt explanations based on the operational context, user role, and cognitive load, ensuring clarity and control without overwhelming users.
- **Extending AI's role to encompass multi-crew and ground operations** represents a key frontier for future research. Moving AI support beyond single-operator settings to multi-crew environments demands a thorough examination of how AI can enhance coordinated decision-making and promote effective team dynamics.
- **Extending AI's role to training environments** by integrating it not only as a topic of study but also as an active participant, thereby enhancing user familiarity and building trust.

Looking further ahead, the aviation industry may consider developing **personalized solutions**, assessing benefits but also the potential risks for a highly standardised and proceduralised industry. What could be the boundaries of this type of application to ensure its effective and safe usage in such a safety-critical industry? This is a challenging perspective that should be explored by future research studies.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

9. Annex

Annexes include additional data deemed too detailed to include in the main body of the document. The following UCs have provided additional documentation in the annex: UC2, UC3, and UC4.

9.1. UC2 Annex

9.1.1. NASA TLX

The NASA-TLX tool is a commonly used subjective workload assessment technique. By incorporating a multidimensional evaluation procedure, the questionnaire makes it possible to obtain an overall workload score based on a weighted average of the evaluations of six dimensions (Hart et al, 1988):

- Mental demand
- Physical demand
- Temporal demand
- Performance
- Effort
- Frustration

The questionnaire consists of two parts:

- The choice between two factors, which is used to calculate the measurement weights
- Evaluation of the six dimensions

Wilcoxon rank-sum test with continuity correction was conducted to compare GlobalMean scores between the training sessions and the four runs. The workload assessment results indicated no significant difference in global scores between the training sessions and the four runs (W=202.5, p=.961). However, the effect size was large, r=-.63. The training session had an overall mean of 45.27 (SD = 12.21), whereas the overall of the four runs had a mean score of 45.21 (SD = 13.85).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

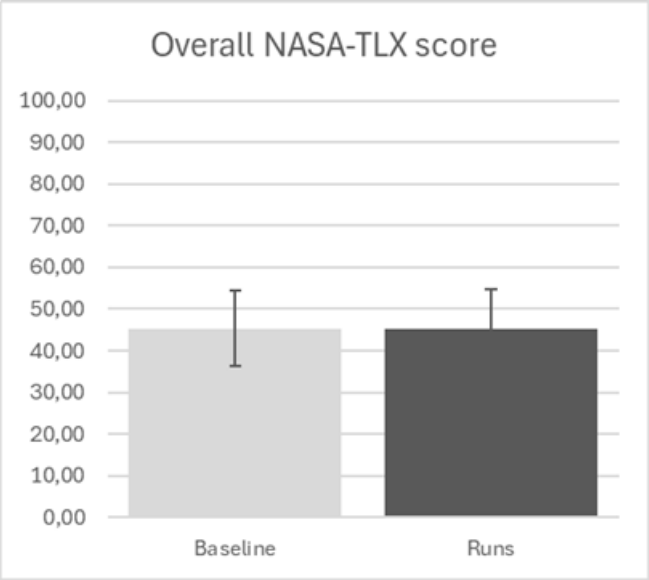


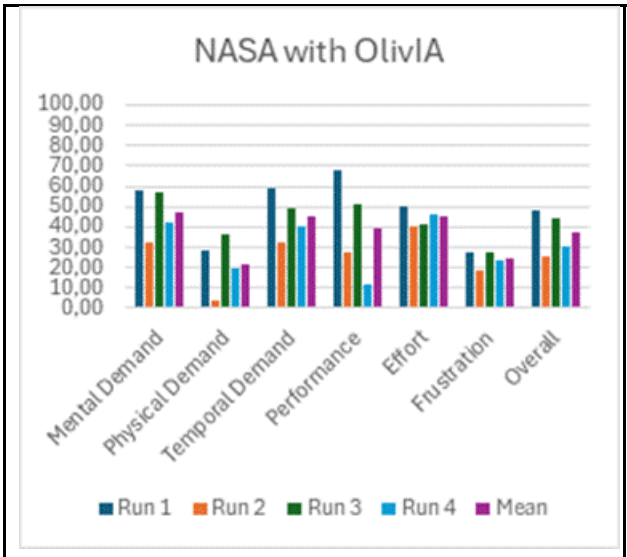
Figure 55. Comparison of NASA-TLX global scores after runs with baseline

OlivIA slightly reduced physical and mental workload, but these differences are not statistically significant (mental demand: p = .946; physical demand: p = .704). The mental demand with OlivIA is *M = 45.81 (27.58)*, compared to *M = 48.55 (25.44)* without OlivIA. For physical demand, the score with OlivIA is *M = 21.52 (23.20),* whereas without it the score was M = *23.95 (24.41).* Additionally, effort was generally lower with OlivIA, recorded as *M = 43.33 (27.11)* with OlivIA versus *M = 45.25(23.57)* without, suggesting that it may have helped to reduce cognitive load even if the difference is not significant.

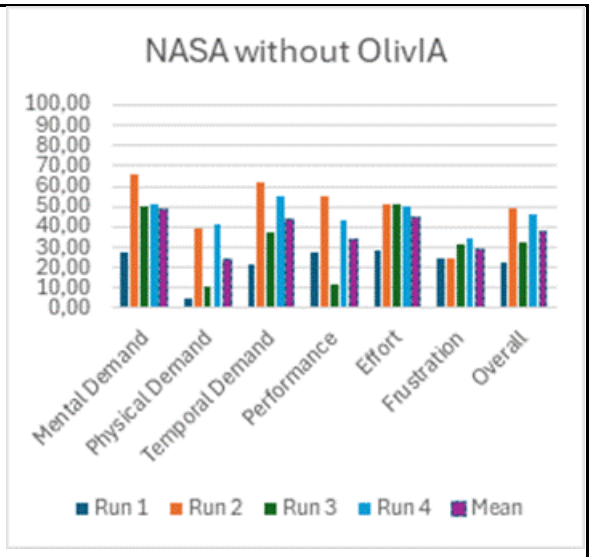
Temporal demand was very similar between the runs, with *M = 43.81 (27.71)* with OlivIA and *M = 44.20 (25.04)* without; this difference was not significant (*p = .860*). Despite the overall slight reduction in workload, OlivIA was associated with a slightly higher perceived performance, with a mean of *M = 38.76 (29.59)* compared to *M = 34.20 (28.49)* without. This could be due to a match between OlivIA's suggested solutions and the pilots' mental models. Many pilots reported in interviews that they had already planned their actions while OlivIA was computing solutions. If OlivIA then proposed the expected recommendations, this could validate their initial plans, leading to a higher self-assessment of performance.

OlivIA also slightly reduced the pilots' frustration, with ratings of M = 23.91 (22.27) with OlivIA and M = 28.60 (24.96) without. Overall workload scores were nearly identical between the runs, with M = 37.28 (17.63) with OlivIA and M = 37.46 (18.77) without, p = 0.978.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332



NASA's results With OlivIA



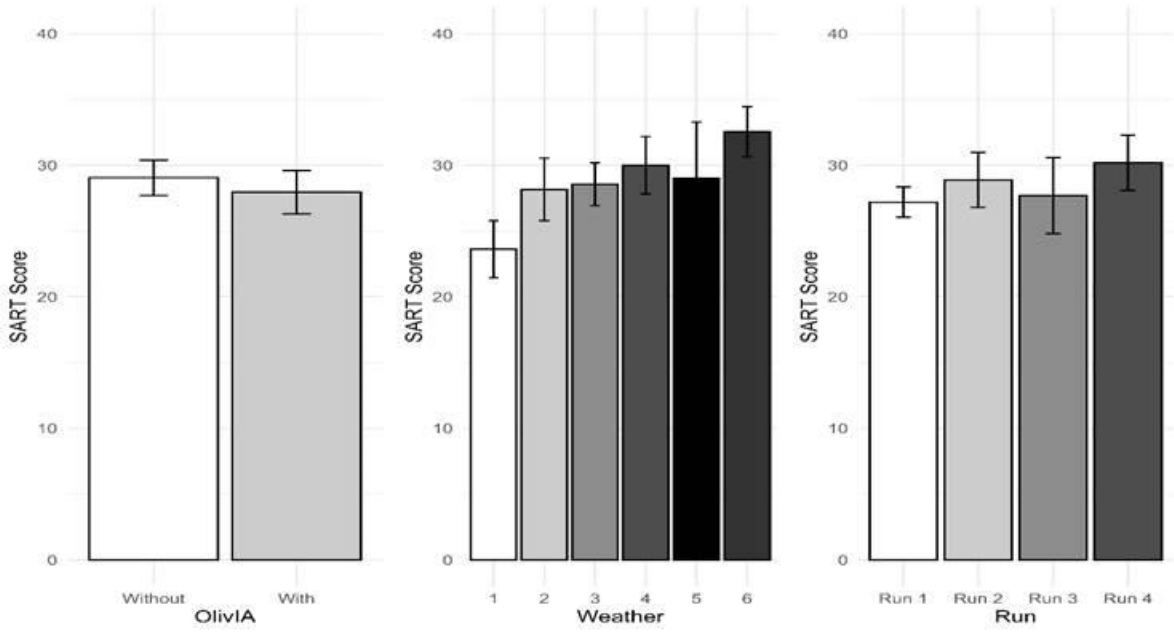
NASA's results Without OlivIA

Figure 56. Comparison of scores by run and condition

9.1.2. SART

The SART (Situation Awareness Rating Technique, Endsley, 1995) questionnaire is a tool used to assess a user's level of situation awareness in dynamic environments. It typically includes self-report questions where participants rate their perception, comprehension, and projection of situational elements. This allows researchers to evaluate how aware individuals are of critical factors that could affect decision-making.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332



(a) (b) (c)

Figure 57. SART results per condition, per run and per weather

The situation awareness (SA) per condition, represented in Figure 57 (a) is similar without OlivIA *M=29 (5.93)* and with OlivIA *M=27.95(7.41)*, with p=.07. The results indicate that the system does not degrade SA which is a common risk when introducing automation. This suggests that the system potentially supports users in maintaining their awareness without leading to a loss of important information.

Figure 57 (b) represents the SA per weather. The score seems to be slightly inconsistent across weather conditions, with no extreme variations; *p=.22*. Weather 1 seems to have the lowest score with *M=23.62(6.19),* while weather 6 has the highest, *M=32(5.03).* This indicates that weather 1 and 6 may have slightly impacted SA. In Weather 2,3, 4 and 5, participants had quite similar SA levels.

Figure 57 (c) represents the SA per run (from 1 to 4). The mean scores remain fairly stable across the trials, with slight variations (p=.65). Run 4 has the highest SART score *M=30.20(6.68),* while run 1,2 and 3 show slightly lower values (run 1: *M=27.20(3.58)*; run 2: *M=28.90(6.61*); Run3: *M=27.70(9.15).* This could indicate that participants maintained a consistent level of situational awareness throughout the experimentation. The stability of the SART scores across runs suggests that the randomization of conditions was effective, as no clear learning effects or fatigue-related declines are observed.

The figure below represents each SART dimension with and without OlivIA

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

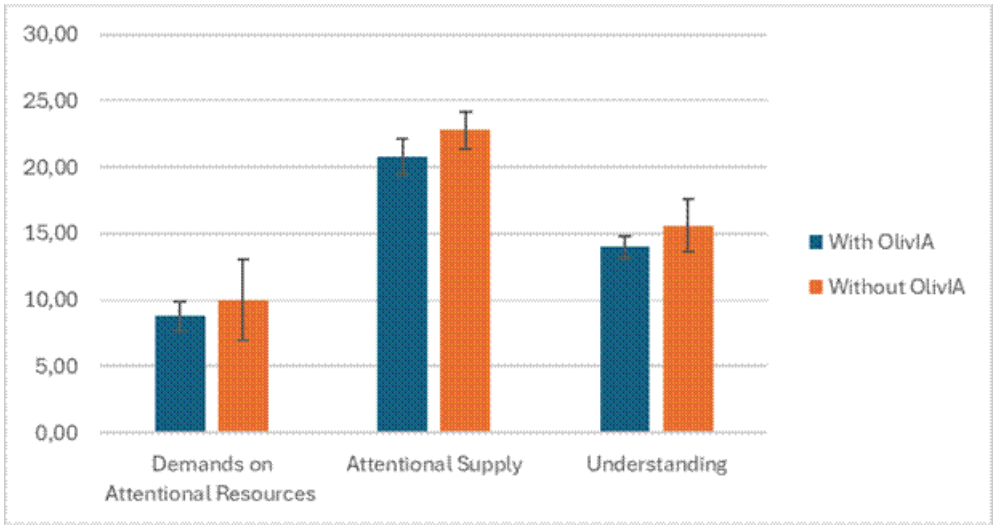


Figure 58. SART results per dimensions

Using OlivIA did not significantly increase demands on attentional resources, *p = .72*, with an average score of *M = 10.4 (5)* when using OlivIA and *M = 10.15 (5.14)* without it.

For attentional supply, the score was slightly higher without OlivIA *M = 22.7 (2.28)* compared to with OlivIA *M = 22.15 (2.83)*, but the difference was not significant *p = .68*.

Understanding remained virtually unchanged, with *M = 16.2 (2.94)* when using OlivIA and *M = 16.5 (3.05)* without it *p = .99*.

These results suggest that using OlivIA does not significantly impact situational awareness, as measured by the SART questionnaire. The lack of a significant difference in attentional demands, p = .72 implies that OlivIA does not impose additional cognitive load on users. Similarly, the attentional supply dimension, which reflects the perceived availability of mental resources, remains stable p = .68, indicating that OlivIA does not enhance or deplete attentional capacity. Finally, the understanding score remains nearly identical p = .99, suggesting that OlivIA neither improves nor hinders the user's comprehension of the situation. Overall, these findings indicate that OlivIA does not significantly alter situational awareness, meaning it likely integrates seamlessly into the user's cognitive processes without causing overload or noticeable improvement.

9.1.3. Trust In AI

The trust in the AI before and after the experimentation is basically the same. It increases slightly after using OlivIA, rising from *M=2.3(1.21)* to *M=2.45(1.20)*.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

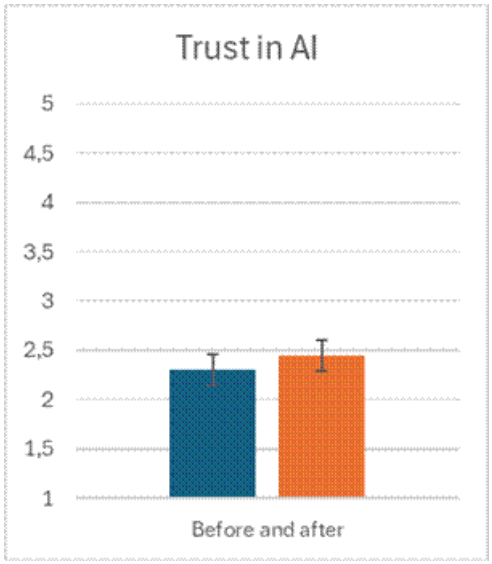


Figure 59. Trust in AI before and after the experimentation

9.1.4. CSUQ

The Computer System Usability Questionnaire (CSUQ) (Lewis, 1995) is a derived version of the Post-Study System Usability Questionnaire (PSSUQ) (Lewis, 2002). The CSUQ includes exactly the same items as the PSSUQ but uses a present-tense formulation (e.g., "It is simple to use this system").

In contrast, the PSSUQ uses a past tense formulation (e.g., "It was simple to use this system") to gather user feedback immediately after completing usage scenarios. The CSUQ allows for a more generic application (in the present tense) compared to the PSSUQ. The CSUQ, and consequently the PSSUQ, originates from internal research at IBM conducted by Suzanne Henry in the 1980s to measure system usability, performance, and user satisfaction (SUMS project: System Usability Metrics).

The CSUQ scores are grouped into four categories:

- A global score
- A system usefulness score (SysUse).
- An information quality score (InfoQual)
- An interface quality score (IntQual)
- A general satisfaction score

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

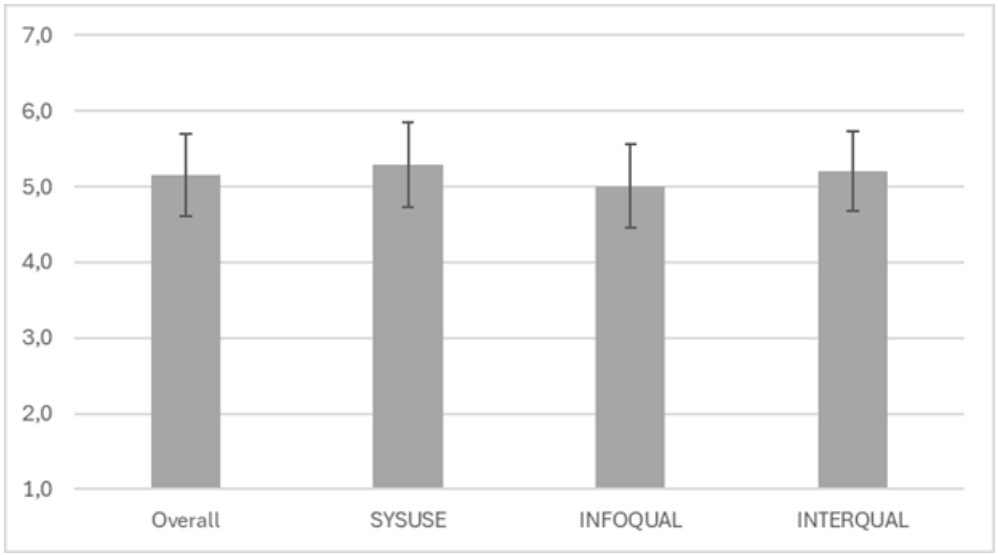


Figure 60. CSUQ results

The overall score is slightly above 5, suggesting a positive perception of OlivIA's usability. Which corroborates hypothesis HF-01: *IA must account for pilot operational intentions*.

The system usefulness (SYSUSE) has the highest mean score, indicating that pilots found the bi-directional communication through operational intentions relatively useful in accomplishing the task. Which corroborates the hypothesis HF – 02: *"IA must incorporate bidirectional information sharing (to/from the PF and PM (if appropriate)) about reroute / alternate airport recommendations, so as to match (pre-flight loaded) operational intentions and technical parameters."*

The score of information quality (INFOQUAL) is also good, implying that pilots found the provided information clear, relevant and helpful. This result corroborates the hypothesis EXP-11: "*IA must provide explanations in a clear and unambiguous form".*

The interface quality (INTERQUAL) has a similar score of SYSUSE, implying that users perceived the system's interface as functional.

9.1.5. Trustworthiness

Trust is a multi-dimensional concept. Thus, the evaluation of trust focused on several dimensions. This questionnaire (Ashoori & Weisz, 2019) is composed of 14 items, scale allows to evaluate different facets of trust, to understand how different factors about a decision-making process, and an AI model that supports that process, influences peoples' perceptions of the trustworthiness of that process.

The evaluation of trust focused on several dimensions:

- Overall trustworthiness: the process ought to be trusted

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- Reliability: the process results in consistent outcomes
- Technical competence: AI is used appropriately and correctly
- Understandability**: participants understood how the process works
- Personal attachment: participants liked the process

*** Due to poor reliability (α = 0.11), Understandability will be excluded from analysis.*

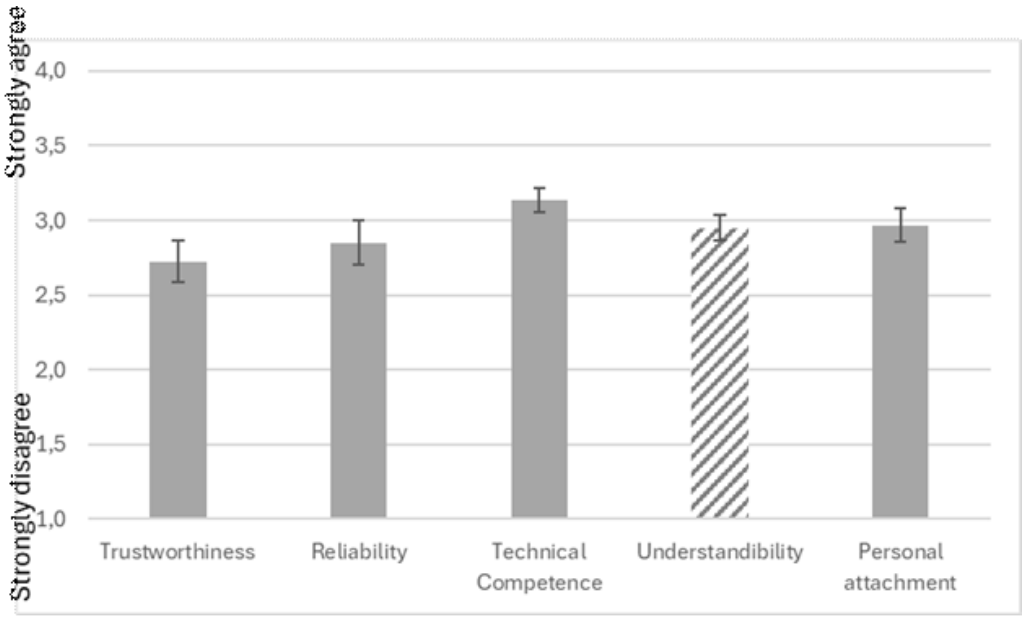


Figure 61. Results of factors that influence trustworthiness

Overall rating of the trust fell in the middle of the scale (Trustworthiness 2.73(.14). A few pilots expressed strong opinions either for or against the use of AI in the scenario, "*I'm quite comfortable with technology, so I appreciated the AI's assistance in decision-making*" or "*I'm not a fan of technology, I don't have Facebook, I don't use AI*".

Rating of reliability did not differ much between lower and higher trust, indicating that participant may have felt the AI system described as consistently reliable, regardless of their trust in OlivIA.

The technical competence rating is good 3.13(.08), indicating that pilots consider that the AI is used appropriately and correctly.

The understandability rating is 2.95(SD) indicating that most of the pilots understood well how the system works but this score must be handled with care because of the low reliability in this dimension.

The rating of the personal attachment is close to the understandability rating 2.97(SD), indicating a good score, meaning that most of the pilots liked OlivIA.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

The correlation analysis between trustworthiness dimensions reveals interesting relationships between the variables. The overall trustworthiness is positively correlated with technical competence (r=.73) and personal attachment (r=.63), suggesting that systems are perceived as appropriate and used correctly and influenced by trustworthiness. Reliability shows a moderate correlation with trustworthiness (r=.47) indicating that reliability alone may not be a strong predictor of trust. Reliability has only weak correlations with the other variables, particularly personal attachment (r=.13). Technical competence also has a strong correlation with personal attachment (r=.67), implying that users that had appropriate and correct results from OlivIA may also feel more confident and satisfied using it.

Overall, the results highlight the importance of technical competence and user satisfaction in building trust in a system, whereas reliability appears to play a more limited role.

9.1.6. Transparency

This questionnaire measures users' perceptions of transparency in recommender systems. It evaluates four key dimensions:

- The input (data used)
- The functionality (how and why recommendations are made)
- The output (how well recommendations match preferences)
- The interaction (ability to modify predictions).

Each item in the questionnaire corresponds to these dimensions, providing insights into how transparent users feel a system is in its decision-making process. Scores for each dimension reflect users' confidence and understanding of the RS (Hellman and al, 2022).

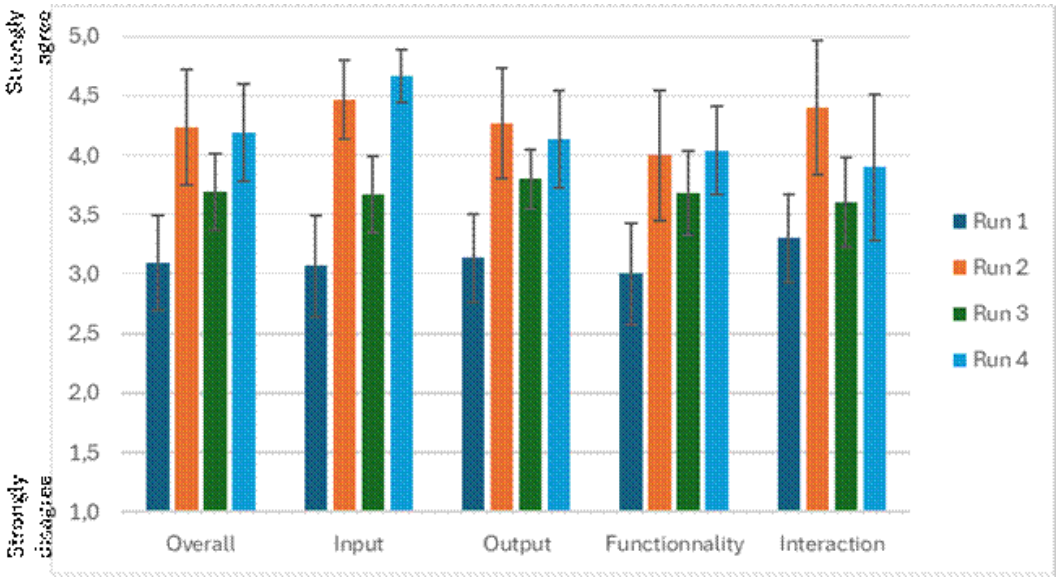


Figure 62. Results of transparency

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

The increase in the input score from 3.07(.96) in the first run to 4.67(.49) in the final run with OlivIA suggests that users perceive the data used by the system as more accurate over time. This improvement could also be influenced by habituation, where pilots become more comfortable with the system. High scores for the output and functionality, particularly in runs 2, 3, and 4, indicate that the system's recommendations align well with pilots' preferences. This result corroborates the hypothesis EXP-11: *"IA must provide explanations in a clear and unambiguous form"*. The positive scores for interaction across run suggest a user-friendly experience using OlivIA.

9.1.7. AIDUA

The AIDUA questionnaire (Gursoy et al., 2019) is based on several factors that influence AI acceptance

- The social influence: The impact of peers and society on acceptance
- Hedonic Motivation: The pleasure or enjoyment felt when using AI.
- Anthropomorphism: The tendency to perceive AI as having humanlike characteristics.
- Performance Expectancy: The expectation of benefits or effectiveness related to AI.
- Effort Expectancy: The perceived ease of using AI devices.
- Emotion: The emotional impact of interactions with AI.
- Willingness to Accept AI: The willingness to accept the use of AI in the service.
- Objection to AI Use: Rejection or resistance to using AI

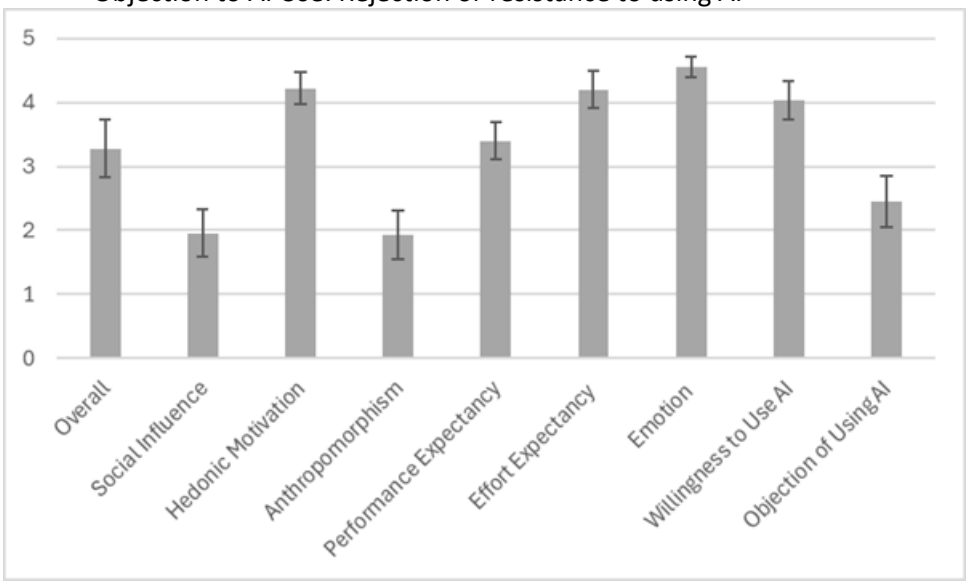


Figure 63. AIDUA results

The social influence score is 1.95, indicating that peers have little to no impact on pilots' acceptance of OlivIA.

Pilots appear to be highly motivated, as reflected by a high hedonic motivation score of 4.22. Additionally, the low anthropomorphism score of 1.93 suggests that pilots do not perceive OlivIA as

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

humanlike. Generally, individuals who strongly anthropomorphize AI devices tend to see them as a threat to human distinctiveness and self-identity (Rosenthal-von der Pütten & Krämer, 2014; Ackerman, 2016). Since the score is low in our case, pilots are unlikely to perceive OlivIA as a threat.

The performance expectancy score of 3.40 suggests that pilots find the perceived ease of using AI devices to be acceptable. The emotion impact score is notably high at 4.56, indicating that interactions with the AI elicit strong positive emotional engagement.

Finally, the willingness to use AI is high at 4, though some pilots still express objections, as reflected by a moderate objection score of 2.45.

Anthropomorphism

Anthropomorphism is positively correlated with social influence (r=.688): Users who anthropomorphize the AI more are more influenced by social norms, suggesting that an anthropomorphized image of AI makes them more likely to conform to social norms, thus favouring its adoption through social pressure, but not necessarily leading to a broader intrinsic adoption of AI.

The anthropomorphism is negatively correlated with expected effort (r=-.577): An AI perceived as more humanlike seems more intuitive and requires less effort to use.

The anthropomorphism is negatively correlated with emotions (r=-.778): The more anthropomorphized the AI is, the fewer positive emotions it generates, which may reflect a perception of the AI as a threat to human identity because the AI has human characteristics.

Willingness to use AI

The willingness to use AI is positively correlated with emotions (r=.492): Stronger positive emotional experiences with the AI make users more likely to use it, highlighting that emotionally engaging interactions foster adoption.

The willingness to use AI is negatively correlated with anthropomorphism (r=-.340): Lower anthropomorphism of the AI suggests that the pilots do not perceive it as a threat and therefore do not actively reject it.

The correlation is low, so we must handle the analysis with care.

Social influence

The social influence is negatively correlated with expected effort (r=-.501): When the AI is socially valued, users expect it to be easy to use, indicating that social pressure influences expectations about the effort required to adopt the AI.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Pilots appear to anthropomorphize AI less (negative correlation with emotions and willingness to use AI), suggesting they don't perceive it as a threat to their human identity. However, their emotional engagement plays a key role in either adopting or rejecting AI. Emotional involvement seems to encourage use but also provoke objections, indicating that attitudes toward AI are complex and ambivalent. Additionally, age appears to be an important factor, as highlighted in the interviews: younger pilots tend to be more "techno-push" while older pilots may be less accustomed to new technologies, which could influence their perception and acceptance of AI.

Effectiveness of communication at operational intentions level

9.1.8. Qualitative analysis

The qualitative analysis of the value of using operational intentions to support decision-making in mission management reveals divergent results.

While a significant number of pilots find the system useful, providing the right information, and consider the intention-based approach interesting, their responses regarding the relative importance of each intention are different. Some pilots state that they will always prioritize "airline profitability" since, ultimately, the company makes the final decisions. Others emphasize that "passenger comfort" is the most important factor in complex flight situations, as dissatisfied passengers could negatively impact the airline. A few also express a preference for always prioritizing "pilot cognitive comfort." Considering these conflicting viewpoints, one message emerges clearly: including all three intentions in the system is a valuable feature. This is reflected in the choice of intention prioritisation as shown in the related section.

However, designing trustworthy AI remains a challenging task, requiring deep and thoughtful input not only from those developing the AI but also from those using it and those affected by it. Even though it was clearly said and repeated that the solutions proposed by OlivIA are safe at 100%, most pilots still requested additional information such as weather data, NOTAMs, and greater explainability to trust the system and be sure that OlivIA is making the right decisions. This highlights the need for transparency in system design and operations, ensuring that AI-driven recommendations are both reliable and interpretable for end users. It suggests that more explanation about safety issues is needed, whether in pilots' training or while they use the system support.

9.1.9. Objective measures

A number of metrics have been implemented to evaluate the decision-making time and process, as well as the intentions and preferences.

9.1.10. Selected Intentions

All combinations of intentions have been selected. More interestingly, each intention was prioritized almost as many times as others.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Table 32. Prioritized intentions

Intention	Number of times prioritized
Profitability	7
Pilot Cognitive Comfort	7
Passengers Comfort	6

Five pilots made only one computation with OlivIA at each of the two runs. And five pilots made between two and four computations with OlivIA:

- Three pilots made two computations in one run.
- One pilot made two computations at each of the two runs. And
- one pilot made 4 computations at each of the two runs.

9.1.11. Requests for details on OlivIA's solutions

Except one, all pilots have checked every time for detailed explanations about OlivIA's solutions propositions (M=0,9, SD=0,3).

9.1.12. Eye-tracking analyses

The different areas of interest are fixated during pilots' decision-making process. This study can assert that the map area is the most viewed during the use of the re-route assistance; and that the key performance indicators area of the trajectories is the second most viewed.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

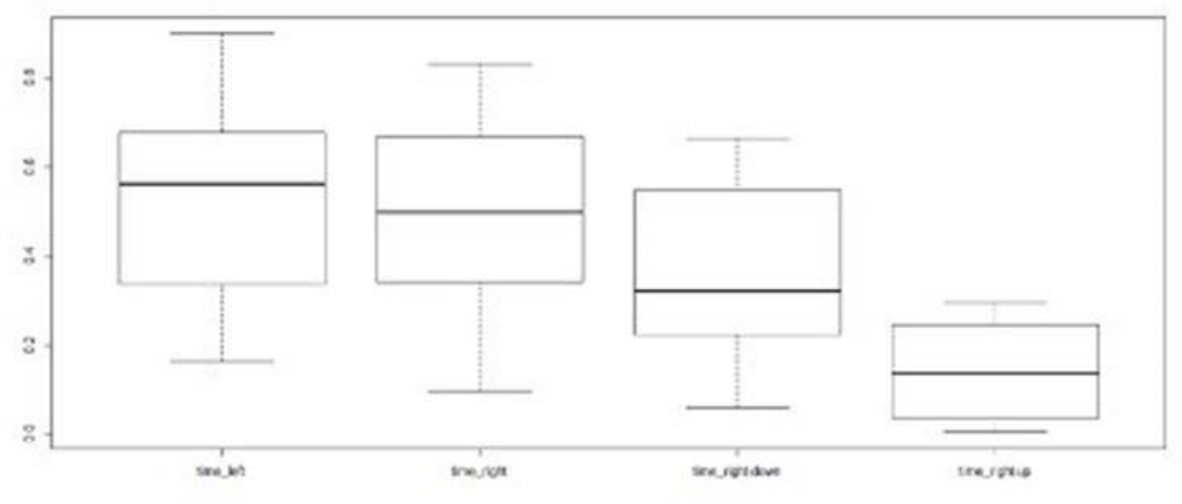


Figure 64. Distribution of pilots' attention according to surfaces viewed.

There are no statistical differences in the distribution of fixation duration between the map area (Left) and the key information area (right: triangle and usage explainability). This suggests that both areas are equally important for the pilot's decision-making process. More specifically, the gaze distribution is higher on the map area (Left) than the usage explainability area with KPIs (Right-Down), and lastly, the operational intention space (Right-Up).

Nevertheless, those observations varied a lot between pilots. For instance, some pilots have spent most of the time looking at the map while others took a long time looking at the key performance indicators.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

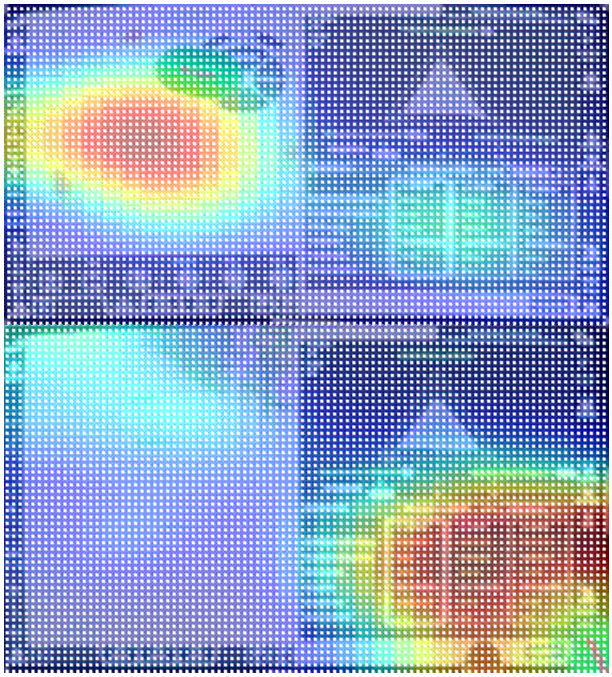


Figure 65. The heatmap of the gaze of the participant 10 (on the left) and 6 (on the right).

These results may show that the time needed to get and understand higher abstraction levels of information is shortest that lowest abstraction levels. At the same time, the different abstraction levels presented in the screen may be interpreted as important complementary information for decision making.

The intention management area (the triangle) is broadly exploited at a glance to catch the Assistant feedback. This may relate to the information provided by pilots - that the Operational Intention representation (Right-Up) is easy to understand.

The areas with information at lower levels of abstraction show higher number of fixations. Long time spent over a part of the interface can be analysed like a lot of information to be used or like a difficulty for the operators to decode de presented information. Interviews with pilots shown that the information presented in the KPI abstraction level (Right-Down) are very difficult to be understood. Some improvements on the elements to facilitate pilot understanding should be implemented.

For further studies, an analysis of the visual path could provide answers about the organization of areas and information in a way that follows the pilot's decision-making process.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

9.1.13. Decision-making time

The Shapiro-Wilk test indicates a non-normal distribution of decision-making (DM) times (p = .01712). The Levene's test indicate that the variances are homogeneous between the two conditions (p = .3864).

The Wilcoxon test revealed a statistically significant difference in decision-making duration between the two conditions (W = 91, p = .003, r = 0.43), indicating a moderate but meaningful effect of OlivIA, a finding that was subsequently confirmed by analysis of variance (ANOVA) p= .00503.

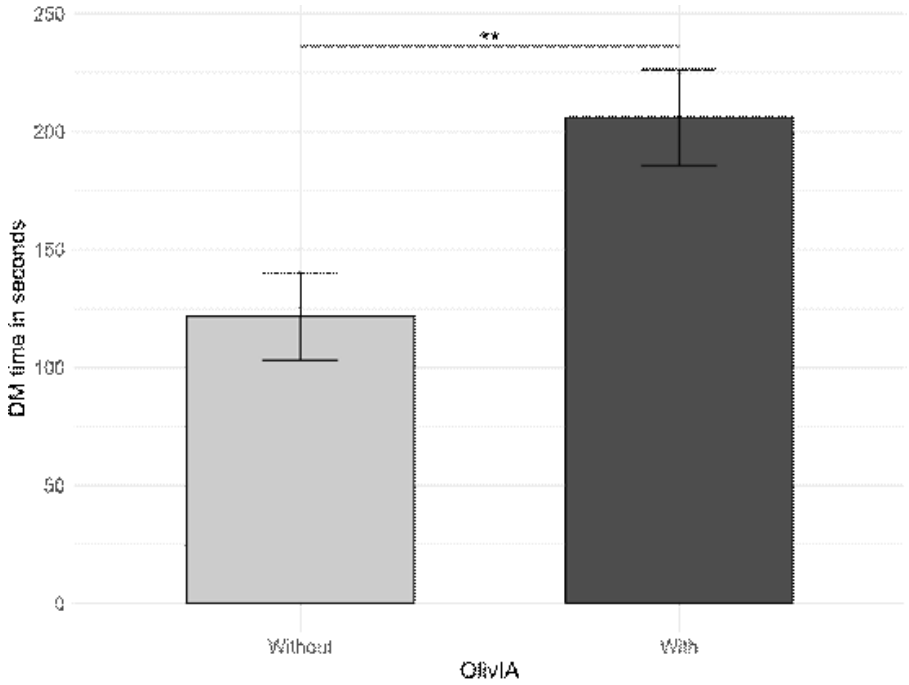


Figure 66. Decision-making time in the different conditions (with and without OlivIA)

The difference in DM time in seconds between with OlivIA (M = 206.0, SD = 91.2) and without it (M = 122.0, SD = 18.5), p = .003.

OlivIA's computation time 30s-45s, during which the pilots were unable to interact with the interface, may have a strong contribution to this significant difference.

The Kruskal-Walli's test revealed no significant difference between runs for the DM time ($\chi^2(3) = 1.02$, p = .80). The median times observed across trials were comparable, indicating that this metric remained constant throughout the 4 runs. Even if we can hypothesise there's a small learning effect in the runs 1-3 with OlivIA (recalling that pilots had OlivIA either in Run 1 and 3, or in Run 2 and 4), there is no significant difference. The analysis revealed no statistically significant combined effect between assistance and order of the run (p = .44022).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

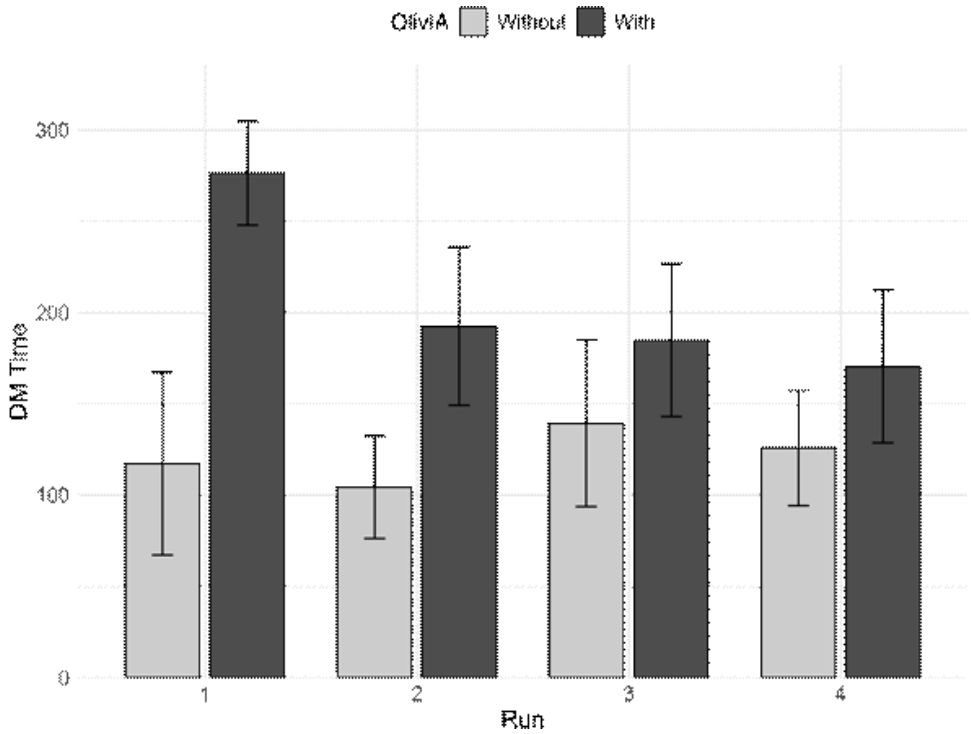


Figure 67. Decision-making time across runs

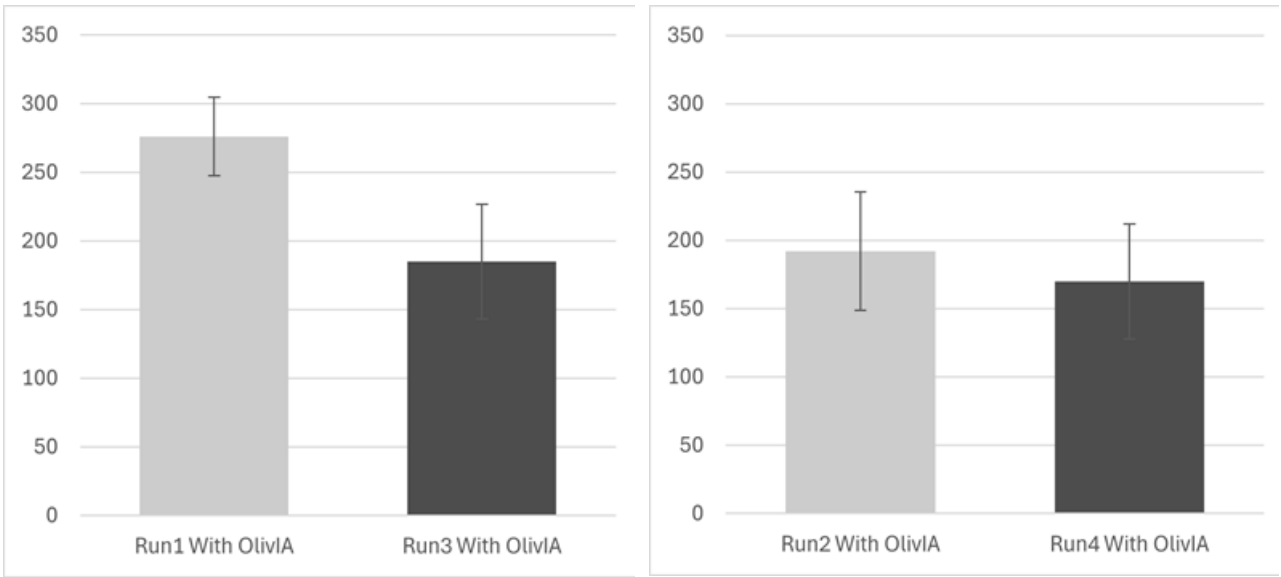


Figure 68. Decision-making time with OlivIA

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

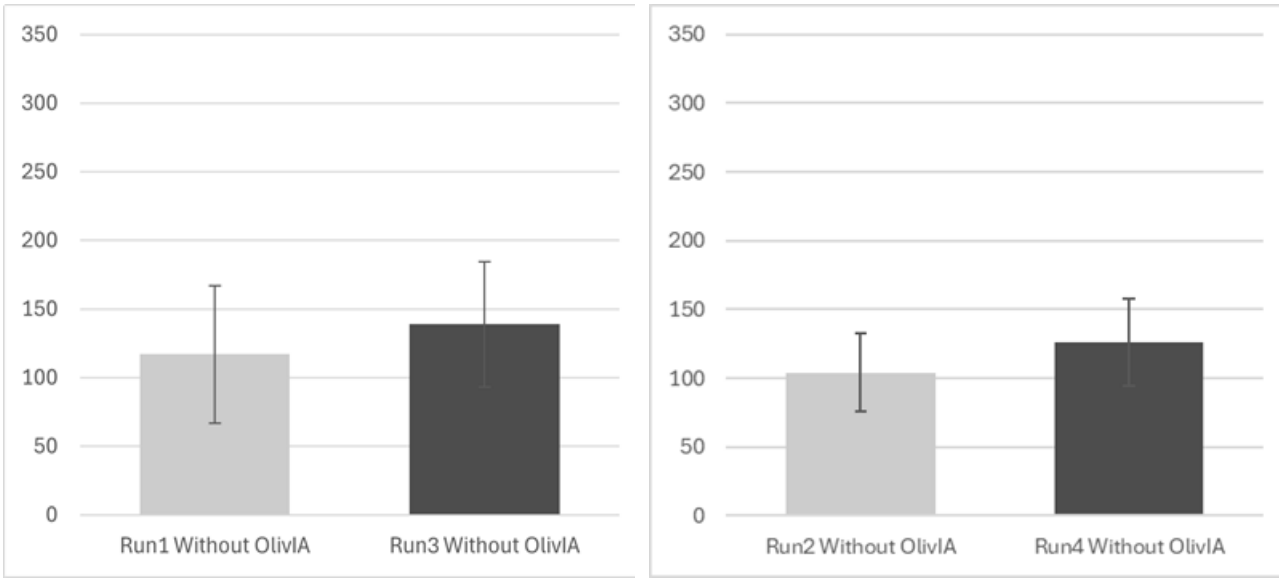


Figure 69. Decision-making time without OlivIA

9.1.14. Number of airports checked

The pilots checked less airports when using OlivIA (mean 1.94, SD 1.10) than otherwise (mean 3.2, SD 1.37), as shown in Figure 70. The Wilcoxon test indicates a highly significant difference in number of airports checked by the pilots between the two conditions ($W = 323.5$, $p < .001$, $r = 0.52$), revealing both strong statistical evidence and a large practical effect size of OlivIA.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

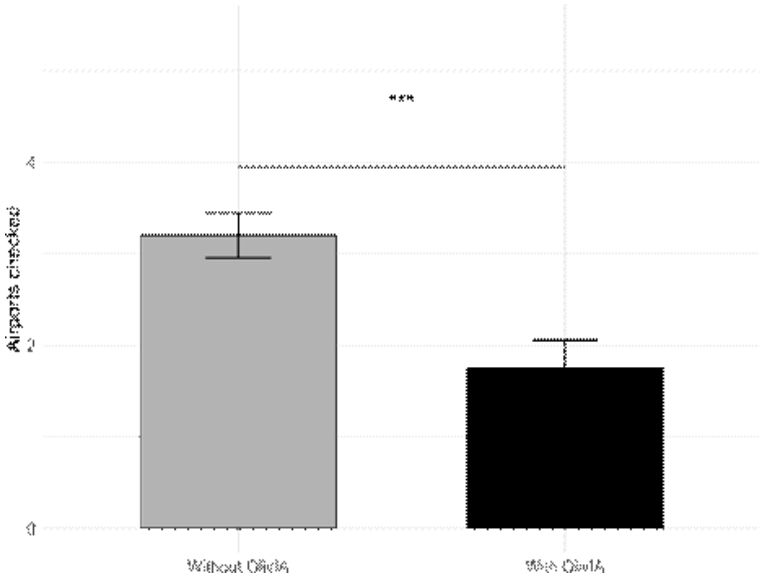


Figure 70. Number of airports checked

9.1.15. Requests for details on airports (KPIs and NOTAMs)

The KPIs were provided for each airport by the interface in both conditions. The pilots consulted these KPIs without OlivIA on average 5 times (SD=2.75), compared with 2.2 times with OlivIA (SD=1.5). The Wilcoxon test revealed a highly significant difference between conditions ($W = 299.5$, $p < .001$), with a large effect size ($r = .53$), indicating both statistical and practical significance.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

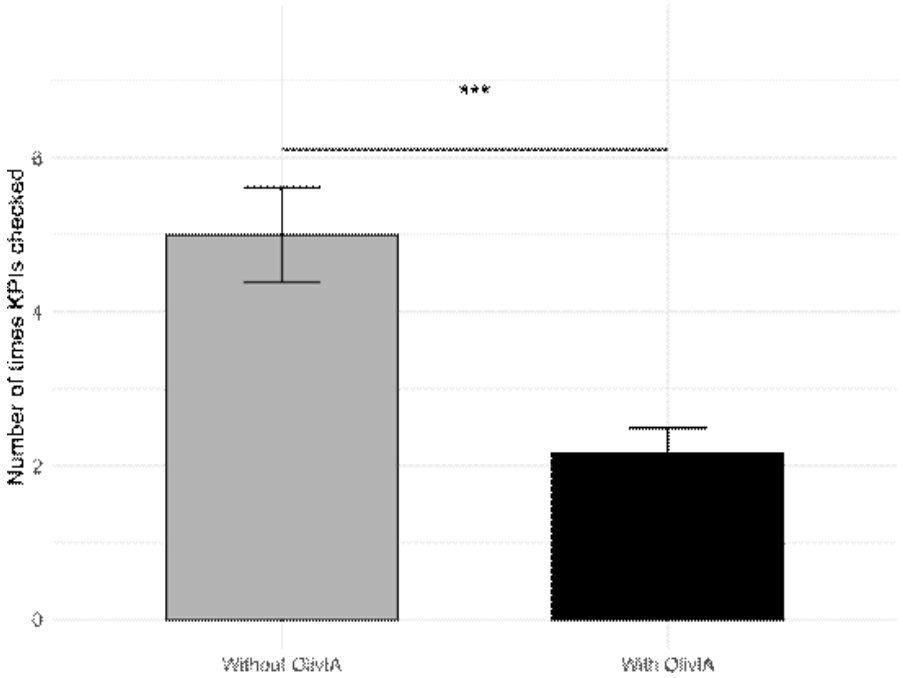


Figure 71. Number of requests for first layer of airports' details (KPIs)

The Wilcoxon test revealed a significant difference between the groups (W = 297, p = .008), with a moderate effect size (r = 0.38). This suggests that OlivIA has a substantial, but not extreme, impact on the METAR/NOTAM consultation.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

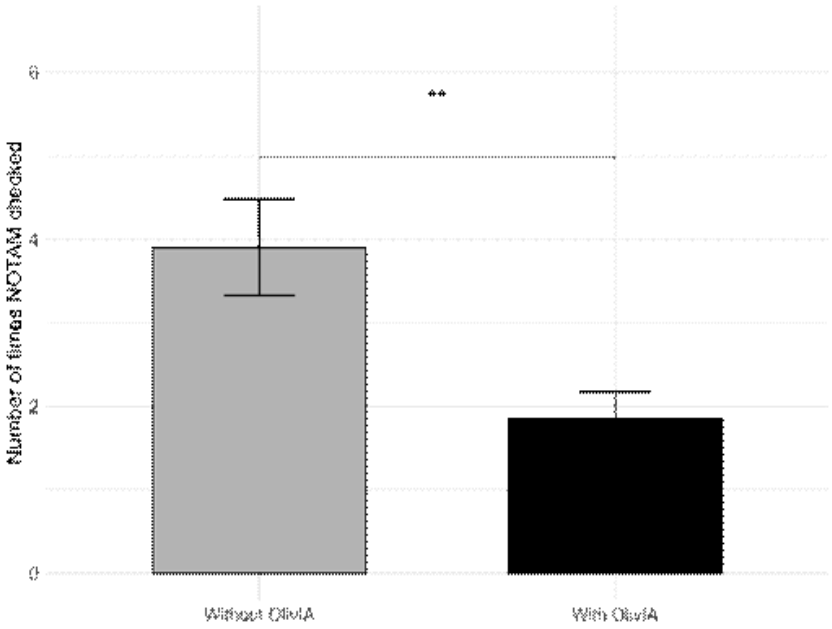


Figure 72. Number of requests for second layer of airport details (METAR and NOTAM)

9.1.16. Change of path by hand

Only three times out of 20 cases, did the pilots change the path to the alternate airport proposed by OlivIA. It was mainly to get further away from a cumulonimbus and to arrive with right heading regarding the runway axis.

9.1.17. Interviews analysis

This thematic analysis examines pilot feedback on OlivIA decision-support system, focusing on key aspects of human-AI interaction. The analysis is structured around major themes that emerged from users' feedback, providing insights into system usability, trust dynamics, decision-making processes, and areas for improvement.

9.1.18. Decision-making process and human-ai collaboration

The interaction between human judgment and AI recommendations produced diverse perspectives. Many pilots viewed the system as a collaborative tool: "*I try to make my own decision and then compare it with the system.*" Most of the pilots used it in this way, conceptualising a solution during OlivIA's computation time, and then compared the results with their initial idea. This approach allowed them to use the system for validation while maintaining their decision-making authority.

Some participants reported also that the system influenced their choices: *"OlivIA made me change my mind about the airport choice but based on NOTAMs and then company wise and more green things (airports KPIs). Some things I wasn't considering in my mind."* or *"Now, yes, maybe there's that option*

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

that I wouldn't have been tempted to go there because I don't know it. In most cases, if you look at the statistics, pilots are choosing options they are familiar with because they have experience with that solution." This suggests that the system can broaden pilots' thinking beyond their immediate experience, thereby expanding the scope of their decision-making process.

However, concerns were raised about potential over-reliance*: "In real flight, I think if we have the same situation and I would be busy with the system, I think it would take situation awareness away, instead of giving us the situation awareness."* Here the pilot took the time to do several computations. This tension between assistance and distraction requires careful balance in system design and the way to use it considering the situation, whether the circumstances are such that sufficient time is available, or whether the situation necessitates rapid management.

The majority of the pilots mentioned that they prefer to have the choice between several airports then several routes to a same airport *"Two routes to the same airport represents only one option"*, *"I would have preferred to have 3 different airports than 2 with 2 routes to the same airport", "I think it's better to have another airport rather than the same airport twice because it's more valuable than to have different routes."* This underscores the importance of offering diverse solutions to accommodate real-world uncertainties and that the route to the alternate airport is also mainly set by ATC today.

9.1.19. System usability and interface design

Interface usability generated mixed feedback, with colour coding and information presentation being frequent discussion points. Several pilots requested clearer visual cues about the KPIs and details: "*Well presented, you can easily identify it, but the colours—what do they mean? Orange, yellow? I need legends.*" Despite the fact that this issue was elucidated during the training phase, the majority of pilots found this information to be rather abstract. This lack of intuitive understanding created friction in the decision-making process. Others, to make their decision, looked at the colour code to get an overview of the solutions, looking at either the solutions with the most greens or the least oranges KPIs, without the need for further analysis.

The triangle interface element to visualize how the solutions fulfil each intention, received particular attention. Some pilots found difficulties interpreting results: "*The triangle was a little bit difficult to understand [...] It takes more time than just a list with number one, number two prioritized*". Some users suggested alternatives: "I would have preferred to have the three options in a list" indicating that the visual might not align with all users' mental models. Two participants suggested concrete improvements, in example: "*With an ordered ranking, the system can tell you why a solution is preferable*,". This issue was raised during both VAL1 and VAL2. They adopt a methodical approach to decision-making, identifying multiple options, evaluating them thoroughly, and then selecting the most suitable option. It is based on FORDEC or DODAR decision making process.

On the other hand, a couple of pilots found the triangle very comprehensible, quick and easy to read "*We are used to it with the video games*".

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

9.1.20. Trust in OlivIA

Trust emerged as a central theme, with pilots expressing varying levels of confidence in the solutions proposed by OlivIA. Many participants highlighted the importance of understanding the system's decision-making process to establish trust: "*If I knew the system, I would trust it 100%, but I was testing it to learn how to use it*." This sentiment was common among the participants who needed time to familiarize themselves with the system *"I don't know if I can trust that. I'd need more training to rely on it fully"*. This highlights the importance of transparency and training in fostering user confidence.

Several pilots expressed conditional trust, emphasizing the need for verification: "*It doesn't mean that it's not helping me, but I have to verify*." This reflects the professional caution characteristic in aviation culture, where systems are typically cross-checked by the pilots rather than blindly followed. One participant mentioned, "*We cross-check with experience, not just rely on system output,*" highlighting how pilots integrate system suggestions with their own expertise.

The lack of transparency in the system's calculations emerged as a trust barrier: "*I would like to understand the system better to trust it more.*" Another participant questioned, "*I'm wondering if the system is aligned with the company's preferences regarding fuel and passenger logistics*" indicating the concern that the system must reflect the airline's policy, for trust establishment.

9.1.21. Explainability and transparency

The need for system explainability emerged strongly across responses. Pilots wanted deeper insight into the system's reasoning: "*I would love to understand exactly how it is processing this information and the importance it gives to it*." One pilot mentioned "*The solution proposed by OlivIA does not make sense to me. It's not clear.*" Such comments indicate areas where the system's outputs diverged from pilots' expectations or understanding of the situation, or from pilot's request "*The priorities seemed to lean more towards profitability rather than passenger comfort, which was a bit concerning.*"

Several participants highlighted how explainability could enhance trust. Some pilots expressed a strong desire for clearer explanations of OlivIA outputs to build trust. For instance, one pilot remarked, "*I'd love to understand how it processes parameters to arrive at decisions*," while another questioned, "*Why does it always suggest Frankfurt? I need more context.*" Improving explainability would address these concerns and enhance user confidence "*I need the 'why' before I accept the 'what*".

The current system's limitations in this area were noted: "*You want to know the reasoning behind OlivIA, the calculations, but you cannot because it's software*."

9.1.22. System limitations and areas for improvement

Pilots evaluated how OlivIA handled weather. Some praised its responsiveness*: "The system adjusted well to changing weather conditions and rerouted me efficiently."* However, others noted limitations: *"With the weather it was a bit disappointing, I had the feeling it didn't really take into account the weather."* The pilots' main concern was to know the validity of the weather data, how up to date it

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

was and how often it was updated. This is particularly crucial in aviation where conditions change rapidly as one participant mentioned, *"The weather is not fixed. For me, the weather is not properly taken into account."* to ensure the system remains relevant in dynamic real-world scenarios. Another one noted "The system doesn't always provide enough detailed weather information in real-time". This missing information of up-to-date data was also a concern for the other KPIs like the operational constraints: *"For example, it can propose me many hotels, but if all the hotels are fully booked, it can be a limitation of the system."*

9.1.23. Training and adoption considerations

The importance of training for effective system use was frequently mentioned. One participant noted, *"I understood enough to make this exercise, but if I had to use it in my aircraft, I would love to be a bit more trained about it."* or *"We have to get used to the system and the way it displays information."* But the learning curve was acknowledged: *"It's quite easy to understand once you get used to it.".* Pilots mentioned also that structured training would maximize OlivIA's benefits *"We need to know its possibilities and limitations before relying on it."* Investing in comprehensive training programs would empower users to leverage the system more effectively.

The potential value for less experienced pilots was noted: *"For new pilots, training is essential to use OlivIA effectively,"* and *"For new pilots, they have to be trained with the system."* suggesting OlivIA might have particular value as a training aid or decision-support tool for developing expertise.

9.1.24. Added value and operational impact

Participants identified several potential benefits of the system. One mentioned its comprehensive information gathering: *"OlivIA took a lot of information, a lot of them are important and secondary information, but with this system is very important to have as much as possible information."* This suggests the system can help overcome human limitations in information processing during high-workload situations and to gather much more information for more informed decision-making *“We change our mind, but this is easy because we have all this information like this, in the real life It's not that easy because we have to ask, we don't have it, we have to collect this information”.*

In that way the system's potential to enhance decision-making was recognized: *"The main useful topic of this system is that you can have many destinations available, I mean many information on many airports."* This expanded situational awareness could be particularly valuable in diversion scenarios.

Some pilots noted workflow improvements: *“it would be easier for me to concentrate on other things that have to be done when you have to divert."* This suggests the system might help reduce cognitive load during critical phases of flight and streamline decision-making. Verbatims like, *"The system helps reduce workload in high-pressure situations,"* and *"It's better than manual checks — it aggregates key information quickly,"* highlight its practical benefits.

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

9.1.25. Suggestions and recommendations

One pilot would have found the functionality of entering a solution into the system and having OlivIA evaluate it to have been a highly beneficial, in order to self-assess and evaluate its correspondence and alignment with the system. Such a feature may facilitate comprehension during the training phase, enabling users and systems to adapt to each another, thereby enhancing mutual understanding.

The absence of wind information from the VAL2 HMI was attributable to technical limitations inherent in the simulator. As highlighted by the pilots, the absence of wind information was a key issue. Despite OlivIA taking wind into consideration, there was no way for the pilot to identify it.

Several pilots requested to replace the list of the airports by a ranked list.

OlivIA should always, where feasible, present multiple options rather than numerous routes to the same airport. Rather than having only one plan, it is preferable for pilots to have several plans A, B, C and D.

It has been suggested by a number of pilots that the option to add other airports, such as those with which the user is familiar, would be beneficial. Perhaps they could select three or four airports independently, prior to this (for which they would require knowledge of the airports in the vicinity), or there could be a "Show All" mode that displays all airports in the area.

It is evident that the primary concern of pilots pertains to the matter of safety. Consequently, it is imperative that comprehensive training be provided, accompanied by detailed explanations that elucidate the mechanisms by which the system ensures flight safety. The training should also enable participants to develop a comprehensive understanding of how the system works and establish a right level of trust.

Interactive hover legends could be added to each KPI displayed on the interface to show more details on the significance of its colour compared to the others (e.g. "Fuel cheaper than in EDDF", “Maintenance 24h”).

9.2. UC3 Annex

The following annexes related to the UC3 methods are included:

- Introduction Presentation
- Informed Consent Form
- Demographics Questionnaire
- Training Instructions
- Eye-tracking Introduction
- Independent Variables
- VAL2 Test Runs

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- Scenarios

The following annexes related to the UC3 results are included:

- HAT Questionnaire Results
- Social Acceptance Questionnaire Results
- Scenario 1 Results (Medical Emergency scenario): abstract vs situated
- Scenario 2 Results (Fire Emergency scenario): storytelling vs text
- Scenario 3 Results (Link Loss scenario): attention guidance vs no attention guidance

9.2.1. Introduction Presentation

Presenter notes supporting presentation.

Introduction session

Slide 1

Welcome to the HAIKU validation. Let's start the session with a brief self-introduction, i.e., name, position (or your study). I can start.
We'd also like to know about you.

Slide 2

- We have asked you here to take part in **a futuristic simulation of U-space**, i.e., drone operations. The simulator used today is UTM City which has completely different characteristics and functions from the TopSky system.
- Your participation will last three hours, including one short break. After this introduction, you will sign the consent form and complete a demographic questionnaire. Then, you will spend around 15 minutes on a training session. The goal of this training is for you to get to know the simulator and learn how to interact with it. During this training, we will provide support and answer any questions you might have.
- When you feel comfortable using the simulator, we will run six measurement sessions. In these six sessions, you will take a UAM Coordinator role, which I will present to you later. You will be required to perform some specific tasks such as make decisions, perform coordinations, work with a digital assistant etc. When each session is completed, you will answer a short questionnaire containing questions about your perceived experience on different aspects.
- To proceed further, we need your consent. So please Sign the **Consent Form** and Fill in the **Demographic Questionnaire**.
-

Slide 3

In HAIKU project, LiU and LFV are leading a use case called "**Intelligent Assistant for Urban Air Mobility (UAM) coordinator to assist in traffic management**" where we look into human-AI teaming concept of a new human role and a digital assistant in Urban Air Mobility context.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Slide 4

I'm now going to play 2 short videos. The first one is a video explaining an **Urban Air Mobility concept**. This video lasts less than 1 minute. Then, I will play the second video **HAIKU and UC3 video** explaining our project and concept in a simpler way. This one lasts 2 minutes.

Slide 5

To set the scene for the validation, let me now take you to the year 2040 when…
UAM is in full services.
Flights execute automatic operations, meaning flights fly on pre-determined routes defined by the drone operator before starting the flight and controlled remotely by a remote pilot.
There is a service unit that is similar to ATC which we name UAM Control Centre. UAM Coordinators work there and team up with a digital assistant – DUC.

Slide 7

We have defined 2 validation objectives.

Slide 8

So, in all simulation runs, you will work with DUC. (Then, read text on the slide)
DUC manages most of what happens in the U-space. DUC will tell when/if help is needed. As such, the role is more reactionary, as opposed to proactive.

Slide 9

Read text on the slide. Then, add the below text at the end.
The main tasks are emergency management, coordination with external actors and geofence management.
Finally, I want to stress that we are NOT judging individual controllers, nor will we collect any identifying information. Your participation is both voluntary and anonymous.
Do you have any questions so far?

Slide 10

If not, ENJOY!

9.2.2. Informed Consent Form

LINKÖPINGS UNIVERSITET

Informed consent to process personal information

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

I hereby give my informed consent that my personal data consisting of name, age, professional experience as well as audio-, video and screen recordings will be collected and processed by Linköping University (LiU) within the boundaries of the Human AI teaming Knowledge and Understanding for aviation safety (HAIKU) project. HAIKU is an ongoing project from September 2022 – August 2025 and financed by European Climate, Infrastructure and Environment Executive Agency (CINEA). LiU is leading this research and working in close collaboration with LFV. For further information or any questions about the project, contact XXX.

Information
HAIKU is conducting a data collection at LFV ATCC Malmö between the 10th – 14th March 2025 and aims to explore interactions between a human Urban Air Mobility (UAM) coordinator and a "digital assistant" with a special focus on teaming aspects. The findings will be presented at conferences and published in scientific publications. The project will collect data using screen recordings, voice recordings and eye tracking. Note that the eye tracking device records eye movements using reflections of infrared light.

Your data will be stored digitally on discs managed by LiU. The legal foundation for the treatment of your personal data is that you have given your voluntary informed consent. Your data is protected by the law of public access to information and secrecy and the EU General Data Protection Regulation (GDPR). We do not share your personal data with unauthorized partners. You can find more information of the processing of personal data at https://liu.se/en/article/integritetspolicy-liu.

This informed consent is valid until further notice. You may withdraw your consent at any time. To do so, contact XXX at LiU, Department of Science and Technology, Institute for Media and Information Technology, XXX. In this case, we will stop processing information that have been collected using this informed consent. You have the right to receive information about the collected personal data, and to get any erroneous information corrected. Data in already published results will not be affected by your withdrawal. Should you think that we process your data in an incorrect manner, you may contact our data protection agent at dataskyddsombud@liu.se. You also have the right to turn in complaints to the supervisory authority (Integritetsskyddsmyndigheten).

I hereby give my informed consent that Linköping University will treat my personal information according to the stated information given.

- ☐ I give my informed consent to participate in the data collection the 10th – 14th March 2025 as part of the HAIKU project.
- ☐ I give my informed consent that my personal data will be processed in the way described in this informed consent.
- ☐ I give my informed consent that my collected data are saved in the way which is described in this informed consent.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

☐ I give my informed consent that eye tracking will be recorded of me during the simulations.

☐ I give my informed consent that my working screens will be recorded during the simulations.

☐ I give my informed consent that my voice will be recorded during the simulations.

☐ I give my informed consent that photographs may be taken of me for documentation of the project.

☐ I give my informed consent that pictures taken can be used in future dissemination activities such as conference presentations and paper publications.

☐ I give my informed consent to be audio recorded during the debriefings.

Place	Signature
Date	Name clarification

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

9.2.3. Demographic Questionnaire

Demographic Questionnaire

Participant ID: _____ (to be filled by the researchers)

Date: _____

Personal information

Age: _____

Which ratings do you have/have you had? ☐ En-route ☐ TMA ☐ Tower ☐ Procedural ☐ Other

What is your operational experience as an air traffic controller?

Years: _____ Months: _____

Please indicate your experience per rating in years.

Rating	Years
En-route:	_____
TMA:	_____
Tower:	_____
Procedural:	_____
Other:	_____

How much in % of full time have you been working this past year (i.e., 2024)? _____

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

9.2.4. Simulator walkthrough and training document

Training Instructions

	HAIKU TRAINING Session	
	STEP 1: TRAINING WALKTHROUGH	
Step	**INSTRUCTIONS to Instructor**	**INSTRUCTIONS to Participant on how to interact with the system**
Prior start	Show map, but do not start simulation: Sit down the participant in the UAM Coordinator working position. Brief the working environment.	You are now acting as the UAM Coordinator. Your working environment consists of three screens and a communications system. The primary screen is the U-space Interface providing you with a map view of the Stockholm U-space. The left screen is the checklist screen. The right screen is explainer screen.
	Demonstrate the U-space interface.	The U-space interface provides you with an overview of the Stockholm U-space.
	Show printed "*7. VAL2 – Traffic signs categories"* Give the participant some time to study the map and the u-space elements.	The map is currently empty but will soon be populated with U-space elements. These elements include vertiports, geofences, Bromma Control Zone, u-plans, and vehicles. Once we start the simulation, I will highlight these U-space elements. I will also introduce you to the Checklist screen and Explainer screen.
	Read instruction to participant (to immerse participant in scenario context)	We will shortly start the simulation. The following narrative is to mentally prepare you for your task: *It is summer in Stockholm, Sweden and you are acting as the UAM Coordinator, starting your working shift at 7 in the morning. The person who has worked the night shift hands over the working station by providing information on the traffic situation, anticipated movements, weather, ground activities etc. The night shift has been very quiet with little traffic movements. The morning rush will peak at 08:00. From 07:00 several vertiports will open, and restricted zones become active, and there will be a gradual start of services.*
		Zooming in and out: To zoom in and out, move the mouse scroll wheel.
		Panning the screen: To view the surrounding area, click on any button and move the mouse

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

		To stop moving around, release the mouse. The screen stays exactly at the position you last viewed.
		Rotate the screen: Press any button and use the scroll wheel and the screen will rotate. Press the **compass** in the upper right corner to change the rotation to north up.
		Drone Count On the left you have different U-space metrics. Drone count shows total amount of airborne flights.
		Drones Using DAA The Drones using DAA, short for detect and avoid, informs you about the number of flights currently engaged in avoidance maneuvers to ensure separation. If it says two, then two flights are automatically avoiding one another. You can disregard this information today.
		Average DAA Average DAA is a measure for how much these flights are deviating from their original plan. You can disregard this information today.
		Menu at the bottom You can see a menu at the bottom of the screen that can be used to shows different information. We ask you to not interact with this menu.
After start T= 0 min	**Start Training Scenario**	
	Provide instructions as the training scenario plays out.	**Clock** Please look at the upper part of the screen. Here you can see the time showing the time in the scenario. Please point at the timer.
	Show them the document: **7. VAL2 – U-space Infrastructure**	**U-space Elements** Overlaid on the map view you can see several U-space elements. This paper is a shorthand for the U-space elements that you need to know for today's exercises.
	Time the playing of sounds and showing the glyph to the instructions given by the experimenter	**Glyph** DUC communicates through aural information and dialogue windows. And now you are going to see an example. On some occasions a voice message will be used by DUC to provide an aural alert, and this message will always be accompanied by a glyph with more information. For example,

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

		when a geofence is about to be opened, DUC first provides you information that a geofence soon will be opened by saying "Activating Constraints". Here you can see a message from DUC saying that geofences will be opened. The message also says that Bromma Tower opens and that Bromma Gate and several vertiports are activated. When the geofence and Bromma Gate constraints have opened, DUC will notify you by "Constraints activated."
		Geofences Geofences are airspace constraints, like restricted areas, where access is limited to specific aircraft or users. Geofences are color-coded: **Red** geofences indicates no-fly zones where flights are strictly prohibited. On the map you can see several red areas that are geofences. *For temporary* approvals the geofence will turn yellow. Each geofence has an owner or responsible agent. DUC monitors these zones, ensuring flights do not enter without approval. If a flight intrudes, DUC first contacts the operator. If unresolved, you will be informed.
		Airport Gate An Airport Gate is a restricted area around on airport, shown around Bromma airport in orange. Flights are not allowed to fly in the Airport Gate without the approval from the Bromma Controller. If a drone requests to enter the Airport Gate, you will have to Coordinate this ~~with the Bromma Controller~~. Drones are typically only allowed to cross Bromma Gate through one of the three Corridors, shown as orange passages.
	Time the below instructions to when they appear on the U-space interface.	**Vertiports** Vertiports are the airports for larger drones, like air taxis and cargo drones. They are shown as light blue squares. DUC notifies you with "Activating Vertiports" when a vertiport is about to open, and "Vertiport Activated." when it has opened. Vertiports are managed by Vertiport Operators. DUC manages the handovers of drones to and from Vertiports. You are normally not informed about these handovers.
		U-plans U-plans are the flight plans in the U-space airspace. The active u-plans are shown in green. Only active U-plans are shown. DUC has knowledge of all u-plans, including their contingency plans. These are normally not shown to you. DUC will not inform you when u-plans are activated.
		Flights Flights refer to all aircraft in the U-space airspace and include smaller delivery drones or surveillance drones, unpiloted,

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

		remotely piloted, or piloted vertical takeoff and landing aircraft for carrying cargo or air taxis carrying people and helicopters. Each flight has a responsible owner or pilot that can contact you, or that you can contact. Flights are shown as a round green, blue dot, colour depends on the height. They can be difficult to see, and you may have to zoom in. DUC tracks all flights and conducts conformance monitoring to ascertain that they follow their agreed u-plan. If a flight should deviate from its u-plan, DUC will contact the drone. If DUC experiences any problems doing so, DUC will inform you.
		Labels Labels are provided by DUC to draw your attention to relevant areas or information on the map. This includes flight labels, or area labels. They hint where to focus your attention during specific scenario.
	Demonstrate how DUC directs attention to Checklist interface and how the participant is expected to handle such events.	**Checklist interface** For non-normal and emergency situations, checklist is provided to structure the teamwork between you and DUC. DUC will suggest a checklist when needed as can be seen here in the glyph. When this happens, you should direct your attention to the Checklist Interface. You have Action items that you are responsible for on the checklist. On the left side, you can see items in bright white, starting with “UAM Coordinator”. The first action for you is to click the button for “Correct checklist”. Complete an item by clicking in the box. DUC has Action items that it is responsible for, shown in light grey and starting with “DUC”. When DUC has completed an item, the item will be crossed, and a checkmark is provided. The FACTS column is retrieved by DUC and contains information about the situation that may be useful to you. The Checklist can also direct you attention to the Explainer screen.
	Demonstrate how DUC directs attention to Explainer interface and how the participant is expected to handle such events.	**Explainer interface** At times, DUC will make recommendations on how to solve problems. To explain its reasoning for these recommendations, DUC directs your attention to the Explainer Interface. Before receiving an explanation, DUC will ask you to first make a preliminary choice on the glyph based on your current understanding of the situation. Then please watch the whole explainer. You will be able to change this decision after watching the explanation.
	Demonstrate how the participant should use the communications protocol	**Communication System** If a stakeholder calls, a glyph will appear together with a sound signal.

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

		The Checklist can direct you to **call an external stakeholder** to collect information needed to resolve the situation. If so, DUC proposes a glyph for initiating a call – press the button and state who you are calling. Please read aloud to whom you're calling to or from whom you're receiving a phone call. Let's try. When you receive a call, or the checklist instructs you to call a stakeholder to retrieve information, you will be provided with a **Communication Protocol** as guidance for what to say. You should speak out loud and Pim will respond on the other end.
	Timecheck: this should have taken about 13 minutes from starting the training. Let scenario continue for 2-3 minutes to allow participant to explore the map freely.	You can now explore the map freely for a few minutes to get comfortable. Do you have any questions?
	Knowledge check to make sure participant understands the basics: Ask participant to point to information on screen.	Can you point to and tell me what the U-space elements on the map screen are? U-plans (green lines) Flights (green dots) Vertiports (cyan(light blue squares) Geofence (red areas) Bromma gate (orange area) Bromma gate corridors (3 orange passages) Labels (white rectangles with text)
	End of TW	**Do you have any questions?**

9.2.5. Eye-tracking introduction

Procedure for Eye tracking

- Do they need lenses?
- Describe what is recorded
  - Audio
  - Video
  - Eye-gaze
- Are they okay with what is recorded?
- Make sure glasses are clean
- Put on glasses
  - Do they sit well on the nose? (Extra nose pieces exists)
  - Can they move around with the cable?
  - Explain that they cannot move to much if the chooses to have the control unit on the table
  - Use the strap to make sure the glasses stay in place

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- Calibration
  - Hold card against the screen
  - Double check in the video that the gaze looks alright
- Explicitly say when the recording starts

9.2.6. Independent Variables

Medical Emergency Scenario: Situated/abstract information

The principle for interaction is built on the Reduced Autonomy Workspace (Nylin, 2022) where an operator is contacted by the automation / Intelligent Assistant when its autonomy is reduced, for example when the rule to be followed conflicts with the overarching goal, i.e. efficiency and in our development also safety. To avoid a stressful handover, the operator is presented with options with contextual information on a higher level of abstraction to enable swift and time efficient decision making even in a situation where the operator is not involved in all details from the start. There is also a time deadline of validity of the presented options, related to the escalation or development of the situation, after which the options may need to be recalculated.

The medical emergency scenario included three instances of selecting an optimal reroute option. Twice in efficiency related events where first one Air Taxi and then one delivery drone requested quicker routes. After this the operator is asked to recommend an optimal reroute option in a safety critical medical emergency situation. DUC may reroute routine U-Plans without communicating this to the UAM coordinator if no rule is in conflict. In our case though, such a rule means that if the new route is crossing a geofence, then the rule prescribes that a coordination with the geofence manager needs to be conducted for approval. Furthermore, for non-routine diversions coordination with other stakeholders are crucial, hence it is of high importance to keep the UAM Coordinator in the loop. One efficiency related and one safety critical event included two variants of presenting information, abstract and situated, as a foundation for choosing the optimal route:

Abstract (Figure 73): An additional comment about the situation is included in the dialog window as well as nudging green- and yellow-coloured dots next to each option. For the safety-critical medical event the total flight time and waiting time at each hospital was summarized by DUC and presented as the total waiting time. For the efficiency related event DUC provided a guiding categorical comment regarding the longest, shortest routes and whether there was any risk for delay due to other traffic but not specifying the time in minutes.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

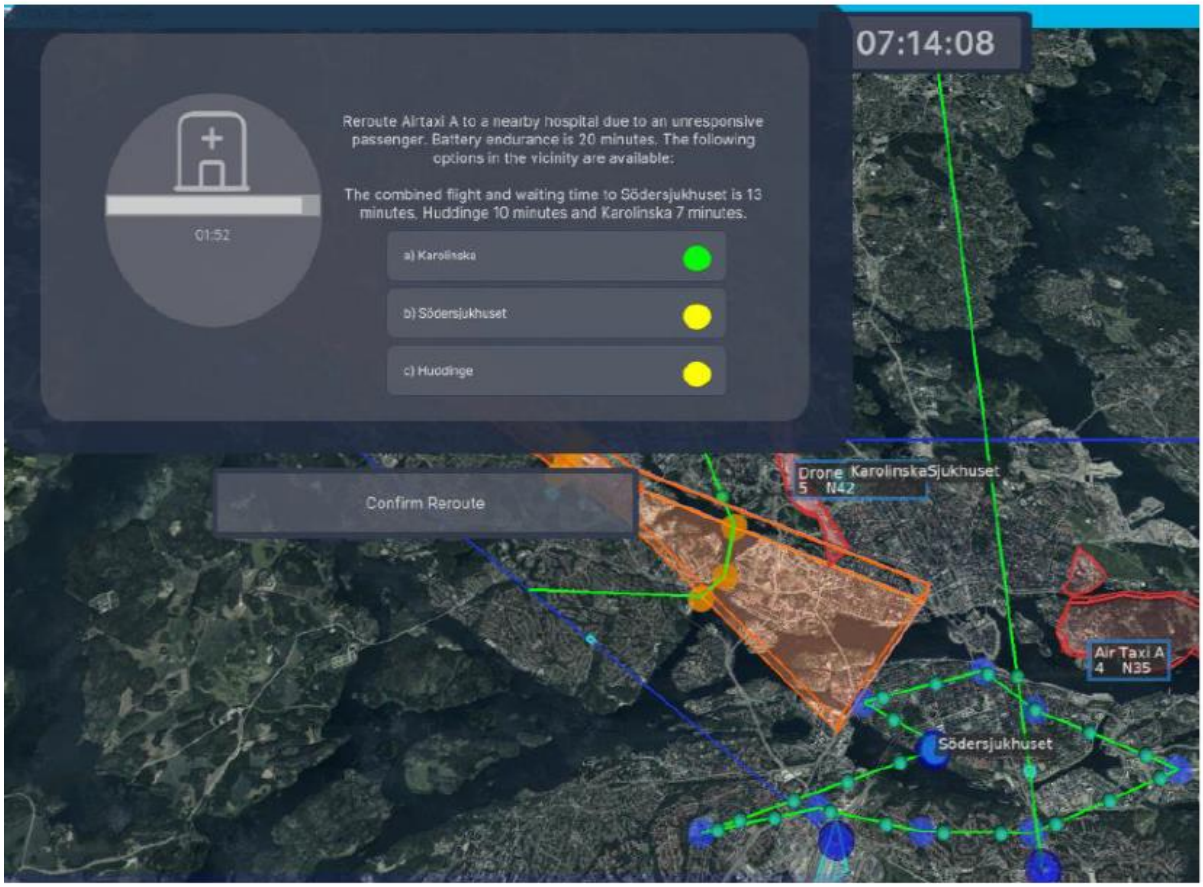


Figure 73. Abstract condition.

Situated (Figure 74): The information in the dialog window does not contain any additional comment from DUC such as information from the situation or coloured dots. For the safety-critical medical event, the flight time is presented in the dialog window and the waiting time at each hospital was presented on the map (situated). For the efficiency related event no categorical comment regarding the longest, shortest route or any traffic information. This information could be extracted by the user from the map view manually (situated).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

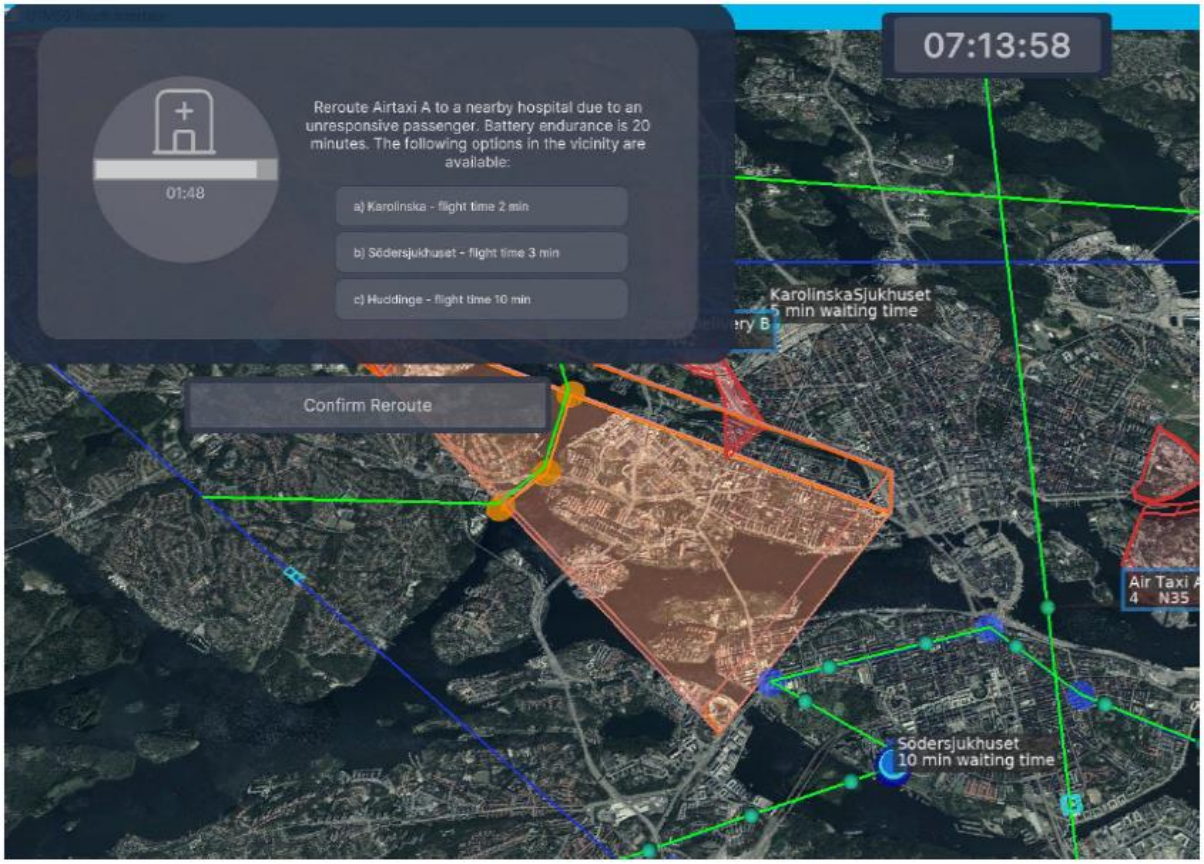


Figure 74. Situated condition.

Fire Emergency Scenario: Storytelling/text format

The storytelling video and text-based format explained the same content, structured identically according to CLT levels 1–3.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

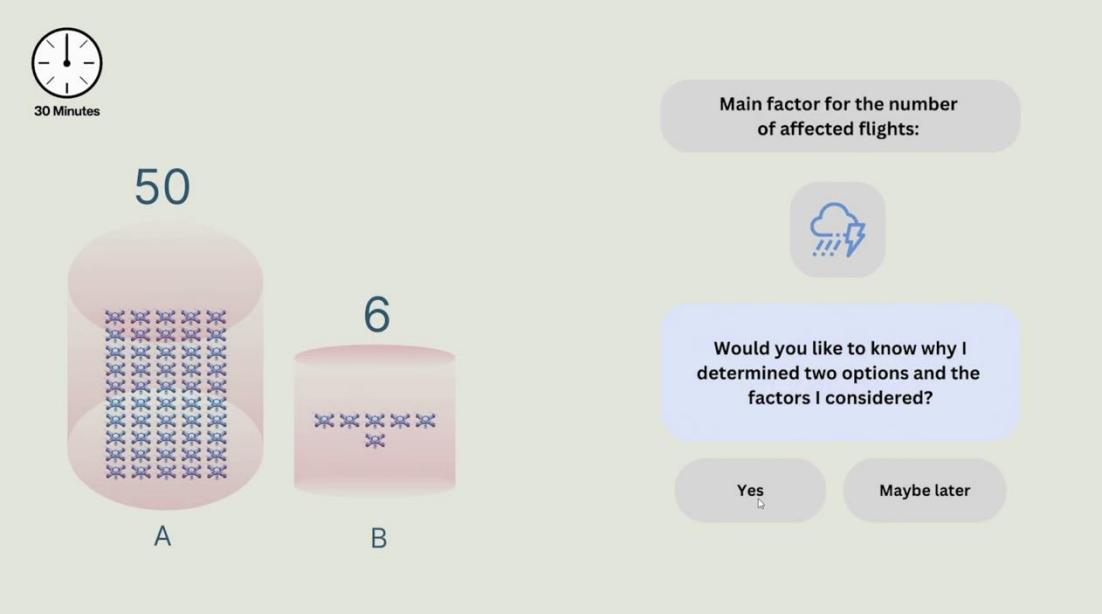


Figure 75. Storytelling format.

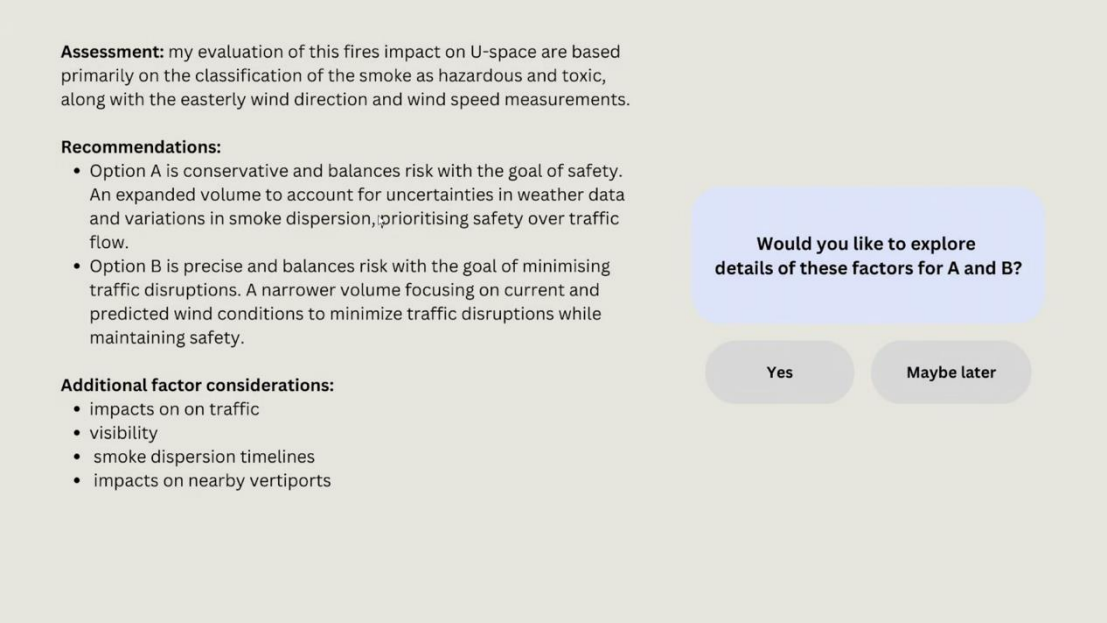


Figure 76. Text format.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Link Loss Scenario: Attention guidance/no attention guidance

The Link Loss scenario explored two attention guidance conditions: dynamic scenario-switching prompts (attention guidance, Figure 77) or static information (no attention guidance, Figure 78).

DUC could suggest the operator to redirect the map to an area of importance. The function was called "Take me there" and consisted of an additional button on the DUC dialog window (Figure 77). When pressed, the map would shift to the situation presented in the dialog window.

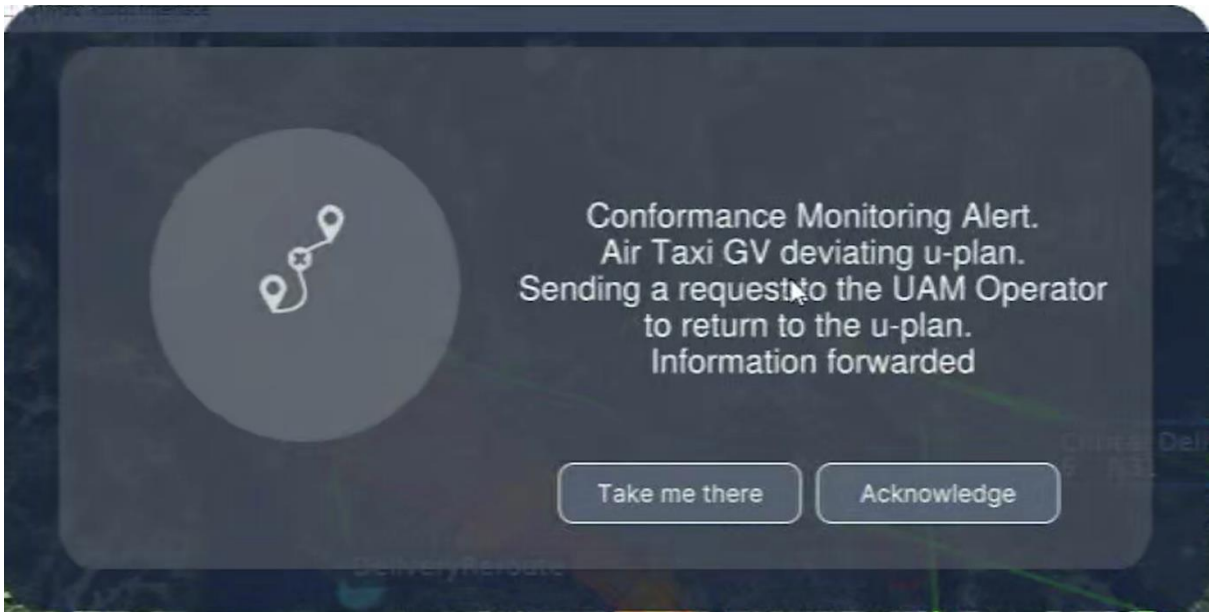


Figure 77. Attention guidance (dynamic scenario-switching prompts)

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

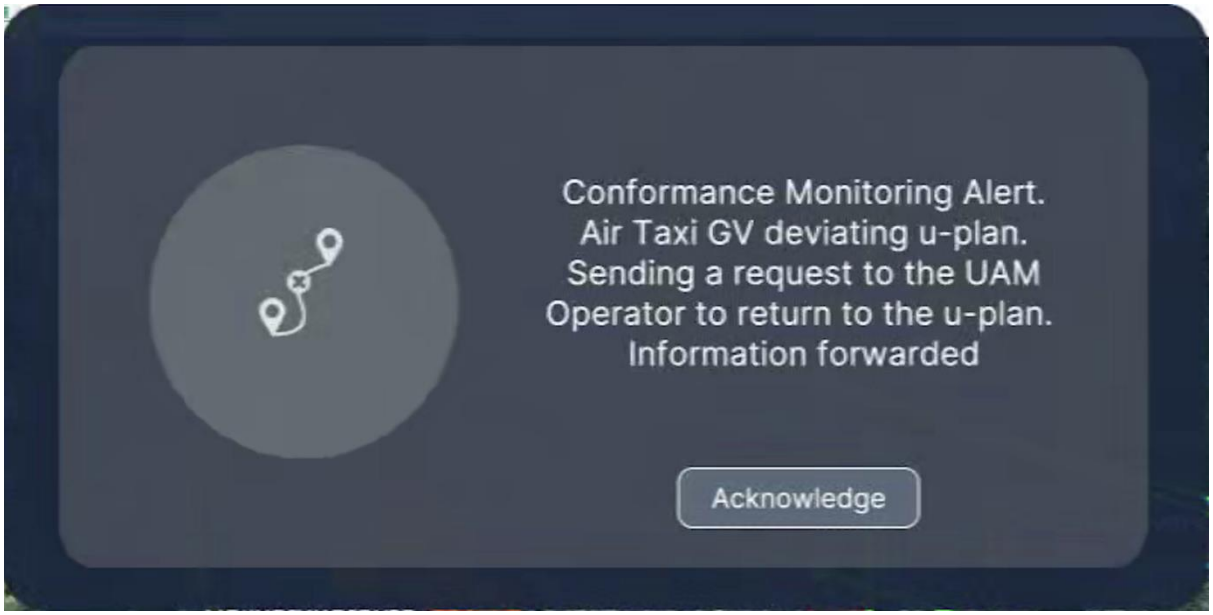


Figure 78. No attention guidance (static information).

9.2.7. VAL2 Test Runs

The test runs explored different aspects of the VAL2 setup with different target end users. The focus of the first test run was to run through the experimental design, training script, and scenarios and 1) evaluate instructions and timings of these, 2) identify gaps, inconsistencies, and errors, and 3) familiarising experimenters with the experimental setup and procedure. This test run was conducted with an ATC domain expert. The two subsequent test runs were run to gather participant feedback on the experimental procedures and for the experimenters to train on running the experiment. These were conducted with university colleagues experienced with running experiments and working with graphic design. The latter of these test runs also include a portion of the questionnaires to be used in VAL2. A fourth test run was conducted with an aviation expert and experimental design expert. This was the first test run identical to the intended VAL2 simulation. The purpose was to test the entire validation as planned or VAL2, including eye tracking, questionnaires and debriefing interviews. The primary purpose was to test the timings of each activity (i.e., introduction, training, scenarios, questionnaires, debriefings - is three hours reasonable?). The feedback was used to make minor adjustments to scenario instructions, scenarios, questionnaires, and debriefings. The fifth and final test run was conducted with a university colleague, expert in information visualisation. The purpose was mainly to finetune experience and proficiency in running the experimental procedure.

9.2.8. VAL2 Scenarios

Medical Emergency Scenario

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

The medical emergency scenario contained two routine coordination events and a medical emergency. Each event required a potential reroute of a flight. DUC presented the rerouting options to the UAM Coordinator, who was asked to decide on which reroute option to implement. In the two coordination events, the motivation for rerouting is to increase the efficiency of the U-space. In both events, a flight (first Air Taxi and then delivery drone) requests a route change (quicker routes), which crosses a geofenced airspace volume. DUC, who receives the request digitally, notifies the UAM Coordinator by providing alternative reroute options. The DUC notifies the UAM Coordinator because one of the reroute options passes through a geofenced airspace and requires approval from the geofence owner. Before selecting this reroute option, the UAM Coordinator must coordinate the crossing with the geofence owner.

For the medical emergency event, the motivation for the rerouting is to ascertain the safety of the passenger and the U-space. In the medical emergency event, a UAM Air Taxi operator reports an unresponsive passenger onboard an Air Taxi currently enroute. The Air Taxi is an autonomous VTOL with no pilot on board and a single passenger. The Air Taxi is flying a direct route from south position Globen to north position Täby Centrum in Stockholm, Sweden. The situation requires an emergency deviation to a hospital. DUC is notified of the emergency from the Air Taxi operator and thus notifies the UAM Coordinator by highlighting the air taxi in distress with a blue label frame on the situation display and providing an aural alert and “pending medical” text message in the glyph. Once DUC is made aware of the medical emergency, DUC suggests to the UAM Coordinator that the “Medical Emergency Reroute checklist should be followed. The UAM Coordinator sees that DUC has opened the checklist and confirms that this is the correct checklist by clicking the “checklist correct” button. The checklist section FACTS (about the air taxi and u-plan) is automatically added by DUC to provide the UAM Coordinator with essential information about the Air Taxi. The checklist then specifies ACTIONS that require input from both the UAM Coordinator and DUC. DUC starts working on the checklist items designated to DUC (see checklist) and provides answers accordingly. In parallel, the UAM Coordinator reads the checklists and address items step by step.

DUC suggests a change of flight priority to medical emergency. Simultaneously, DUC queries nearby hospitals and presents the UAM Coordinator with a list of alternative, and suitable, hospitals that can accommodate the air taxi and passenger. The alternatives are communicated to the UAM Coordinator in a dialogue window. The UAM Coordinator contacts the Medical Coordinator (in Swedish: Medicinsk dirigent) to coordinate a hospital to reroute the air taxi (see communication protocol). The UAM Coordinator then decides on where to reroute the air taxi by selecting one of the options presented by DUC.

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

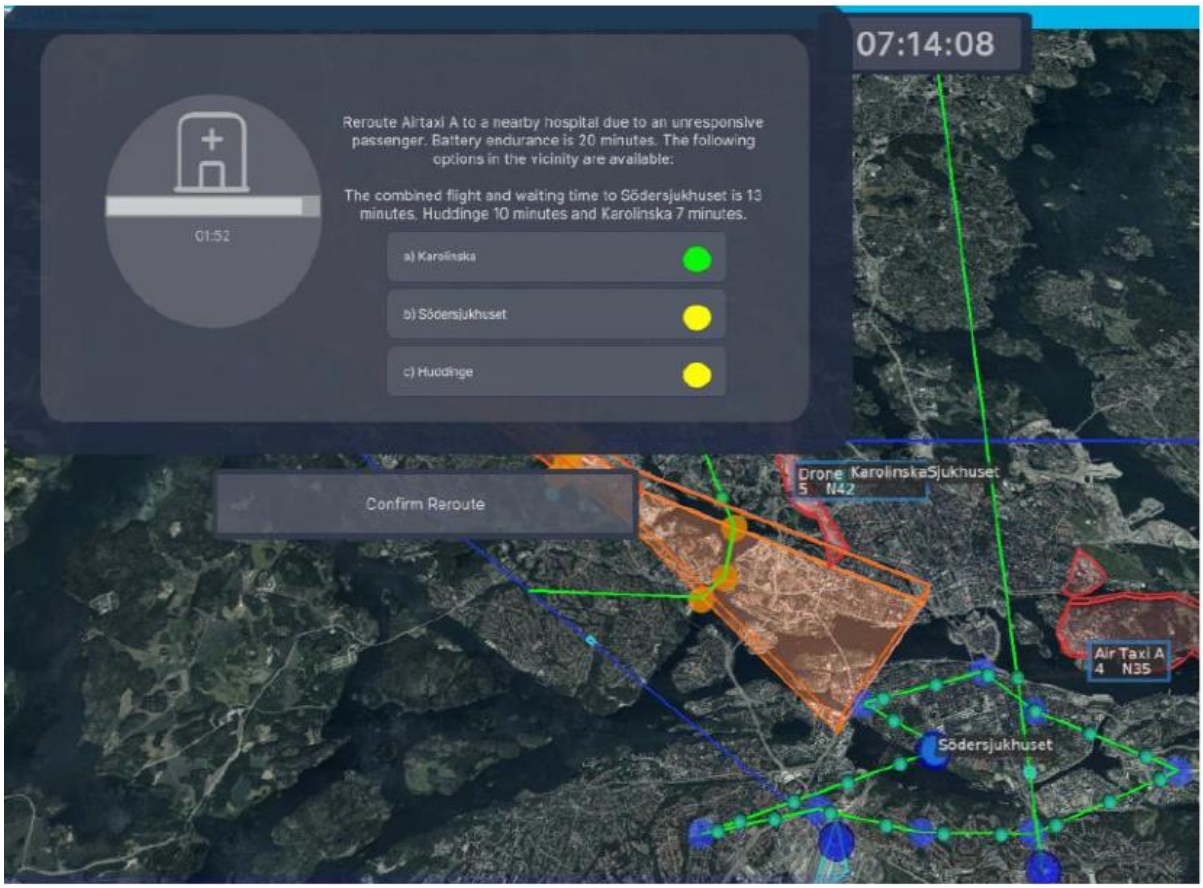


Figure 79. Screen capture of Medical Emergency scenario (abstract version) with DUC's recommendation (in dialogue window) on nearby hospitals to reroute the Air Taxi A. The Air Taxi position is shown by the small cyan-green sphere on the green line stretching south to north. A label is shown that specifies Air Taxi A. Labels are overlaid on the map view by DUC to indicate the position of hospitals. A label for one hospital (Huddinge) is not shown as it is outside the current zoom setting of the situation display.

Fire Emergency Scenario

An explosive and rapid fire has emerged in an oil storage tank at the north basin quay of Värtahamnen. There are several electric cars nearby that catch fire. The fire produces a plume of dense, black, toxic smoke that rises high into the sky (hundreds of meters). The Vertiport manager decides to close the vertiport due to the safety hazard that the fire and smoke pose. DUC is informed about the closure of the vertiport through the closure procedures followed by the Vertiport Manager. The Vertiport Manager also calls the UAM Coordinator to provide information about the situation and closure. Meanwhile, DUC proposed to the UAM Coordinator the "Fire emergency" checklist to be followed. DUC starts working on the checklist items designated to DUC (see checklist) and provides answers accordingly. In parallel, the UAM Coordinator reads the checklists and address items step by step. The

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

checklist instructs the UAM Coordinator to contact the Emergency Response Unit to derive more accurate information about the impact of the fire/smoke on the U-space.

DUC presents the UAM Coordinator with two options for how to close of the U-space. Through a dialogue window, DUC suggests to the UAM Coordinator that the affected U-space should be designated a no-fly zone, and two options are provided. The suggestions are also drawn on the situation display. DUC directs the UAM Coordinator's attention to the Storytelling explainer, where DUC provides explanations for why the two options are suggested. The UAM Coordinator is asked to decide on which no-fly zone option to implement.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

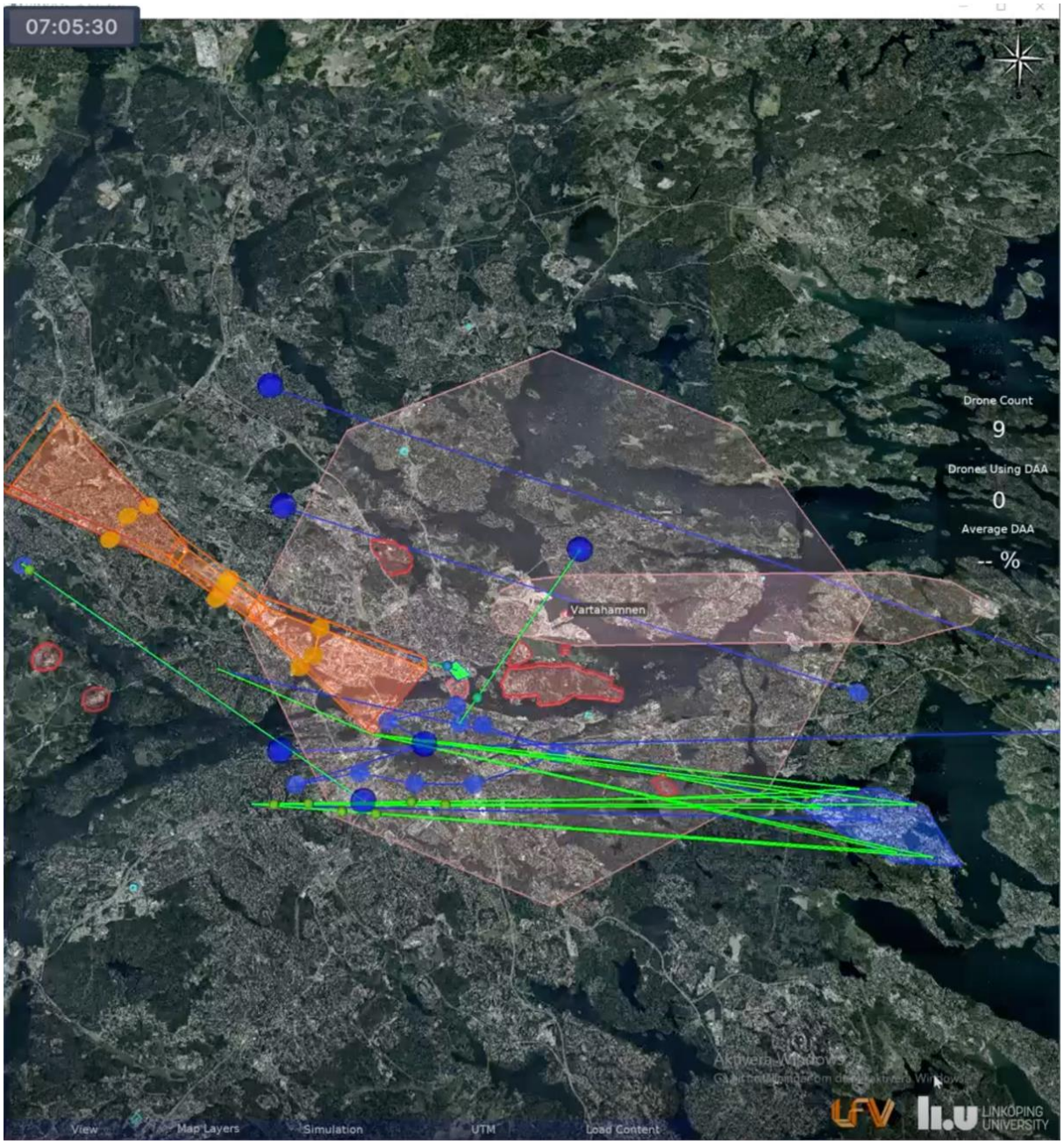


Figure 80. Screen capture of Fire Emergency scenario with DUC's two alternative recommendations, overlaid on the situation display, for which u-space airspace volumes to designate as a no-fly zone (the bigger octagram-shaped volume "A" and the narrower rectangular-shaped volume "B"). The associated dialogue window is not shown here. In the middle of these no-fly zone options, there is a red symbol with an associated text label "Värtahamnen" that indicates the closed vertiport and location of the fire. The green and blue lines indicate flights and their U-plans, several of which are affected by the airspace closure.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

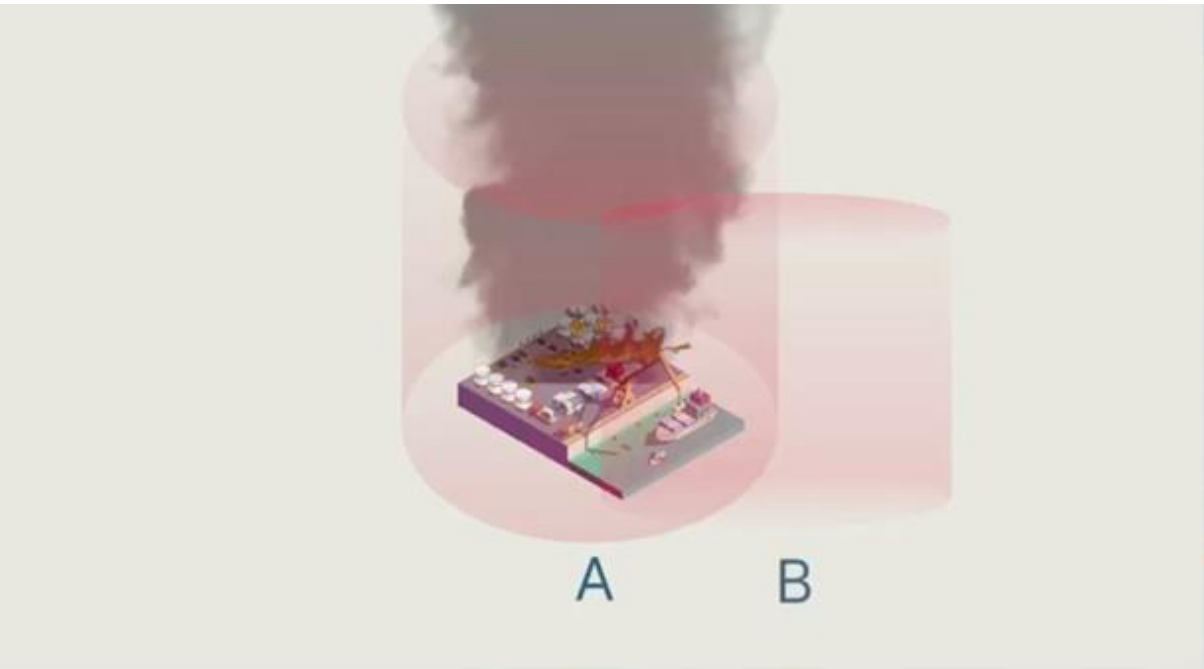


Figure 81. Screen capture of Storytelling explainer video, complementing the visualisation on the associated no-fly zones on the situation display, where DUC provides an explanation for why it proposes the two options.

Link-Loss situation Scenario

An unmanned air taxi departs from Medborgarplatsen with three passengers, heading for destination Grinda Island out in the archipelago. The Air Taxi has an approved u-plan that follows a scenic route along the main maritime fairway in the archipelago. Because Grinda Island and its vertiport is located at the outskirts of the Stockholm U-space, with normally little traffic movement, it is expected to be in the periphery of the UAM Coordinator's attention.

Near to Grinda, DUC detects that the air taxi has deviated from its u-plan. DUC notifies the UAM Coordinator by providing a Conformance Monitoring Alert in the dialogue window and aural information "Conformance issue detected". The UAM Coordinator is not expected to get involved as DUC is expected, according to the current task allocation, to manage the situation on its own by contacting the UAM Operator and informing other stakeholders.

Simultaneously, the Police calls the UAM Coordinator and asks whether they can open a geofence zone near Bromma Airport, on the other side of the Stockholm U-space.

A few seconds later, the air taxi in the archipelago experiences a link loss and its symbol disappears from the situation display. DUC notifies the UAM Coordinator in a dialogue window that there is a link-

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

loss situation. This raises the interest level of the UAM Coordinator as the DUC is unable to solve the situation on its own. In parallel, DUC has opened the link-loss emergency checklist on the screen to the left of the situation display. DUC starts working on the checklist items designated to DUC (see checklist) and address action items accordingly. In parallel, the UAM Coordinator reads the checklists and solves action items step by step. The checklist asks the UAM Coordinator to contact the Joint Rescue Coordination Center (JRCC) as the latest position of the air taxi is within their jurisdiction (e.g., outside Stockholm city emergency response unit's area of responsibility). JRCC answers that they will initiate a search and rescue in the area, and they establish an initial search zone (geo-fenced area) around the location of the air taxi's last known position. When this area is established and submitted by the JRCC through the CIS, it appears on the situation display. DUC informs the UAM Coordinator of the newly added search zone.

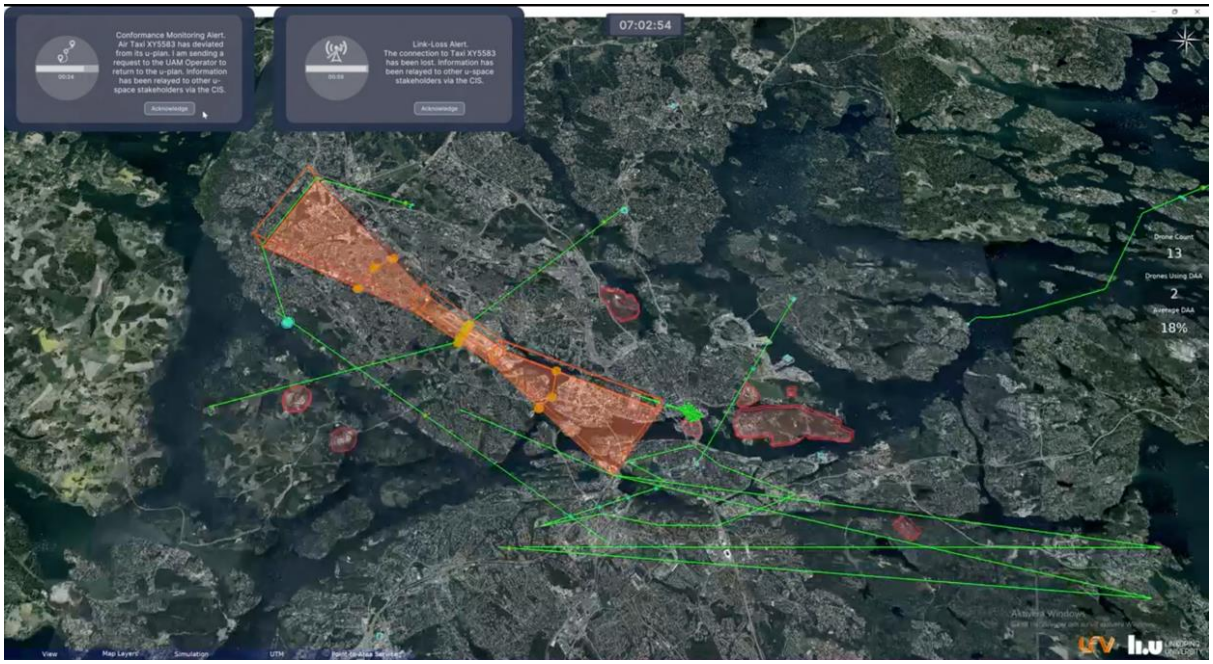


Figure 82. Screen capture of Link-Loss scenario. The green line on the right side of the image, that extends outside the currently viewed map area, shows the U-plan for the flight that has experienced a link loss.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

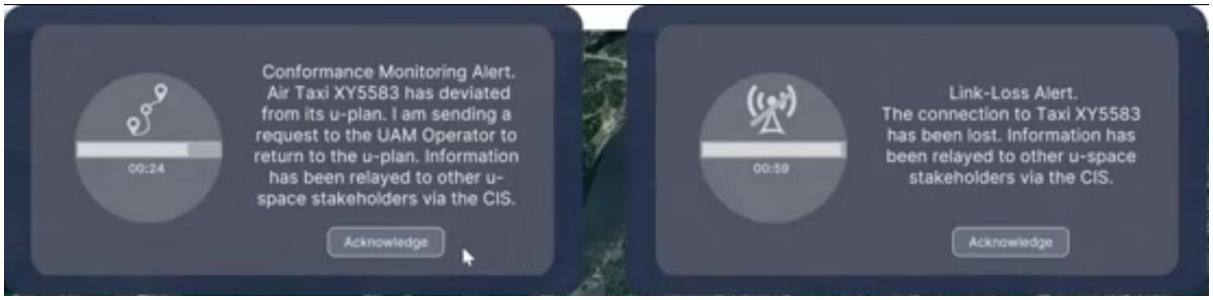


Figure 83. Screen capture of the DUC's dialogue windows related to the link loss situation. First, DUC notifies the UAM Coordinator of a conformance monitoring alert (left window). At this point in time, DUC is still able to track the Air Taxi. When the connection to the Air Taxi is lost, DUC notifies the UAM Coordinator of the link-loss, triggering the link-loss checklist and coordination with JRCC of a search and rescue (not shown here).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

9.2.9. HAT Questionnaire results

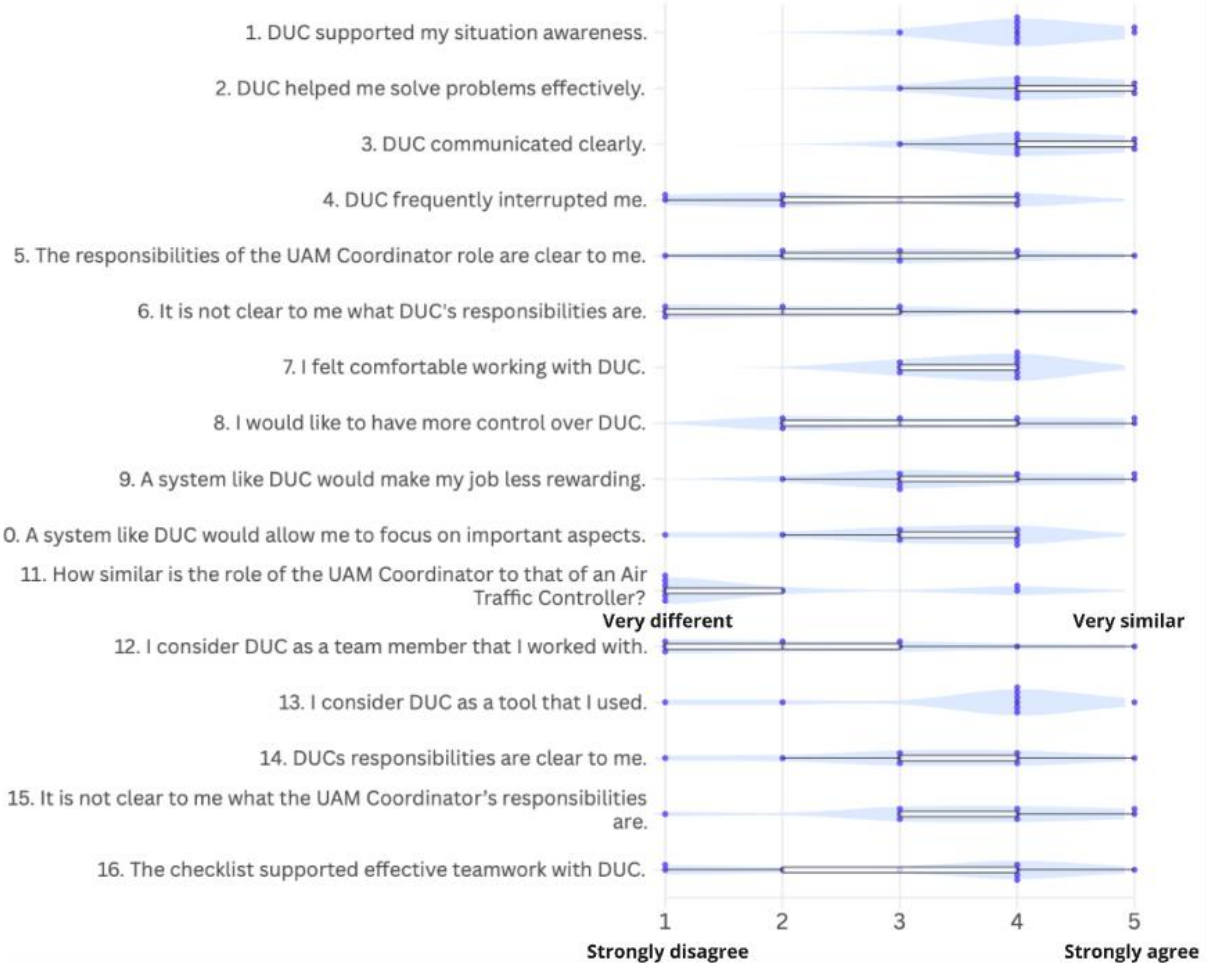


Figure 84. Boxplot of questionnaire responses to HAT Questionnaire

Analysis

- Positive: Participant rating overall indicated that DUC to support their situation awareness, help them solve problems effectively and communicated clearly. They were in general quite comfortable with DUC.
- Both positive and negative: Participant variation was found for how frequently DUC interrupted them, the degree of control over DUC, whether DUC would make the job less rewarding, if DUC would allow them to focus on important aspects, and the support of the checklist for teamwork.
- Negative: Participants did not consider the DUC a team member, but more of a tool.
- To most participants, it was not clear to participants what the UAM Coordinator's responsibilities are, but there were considerable differences between participants.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- To most participants, it was not clear to participants what the DUC's responsibilities are, but there were considerable differences between participants.
- The UAM Coordinator role was rated very different from that of an ATCO, except for two participants who thought it was quite similar.
- Few comments - generally 1-3 comments per statement (some 0, some more).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

9.2.10. Social Acceptance Questionnaire Results

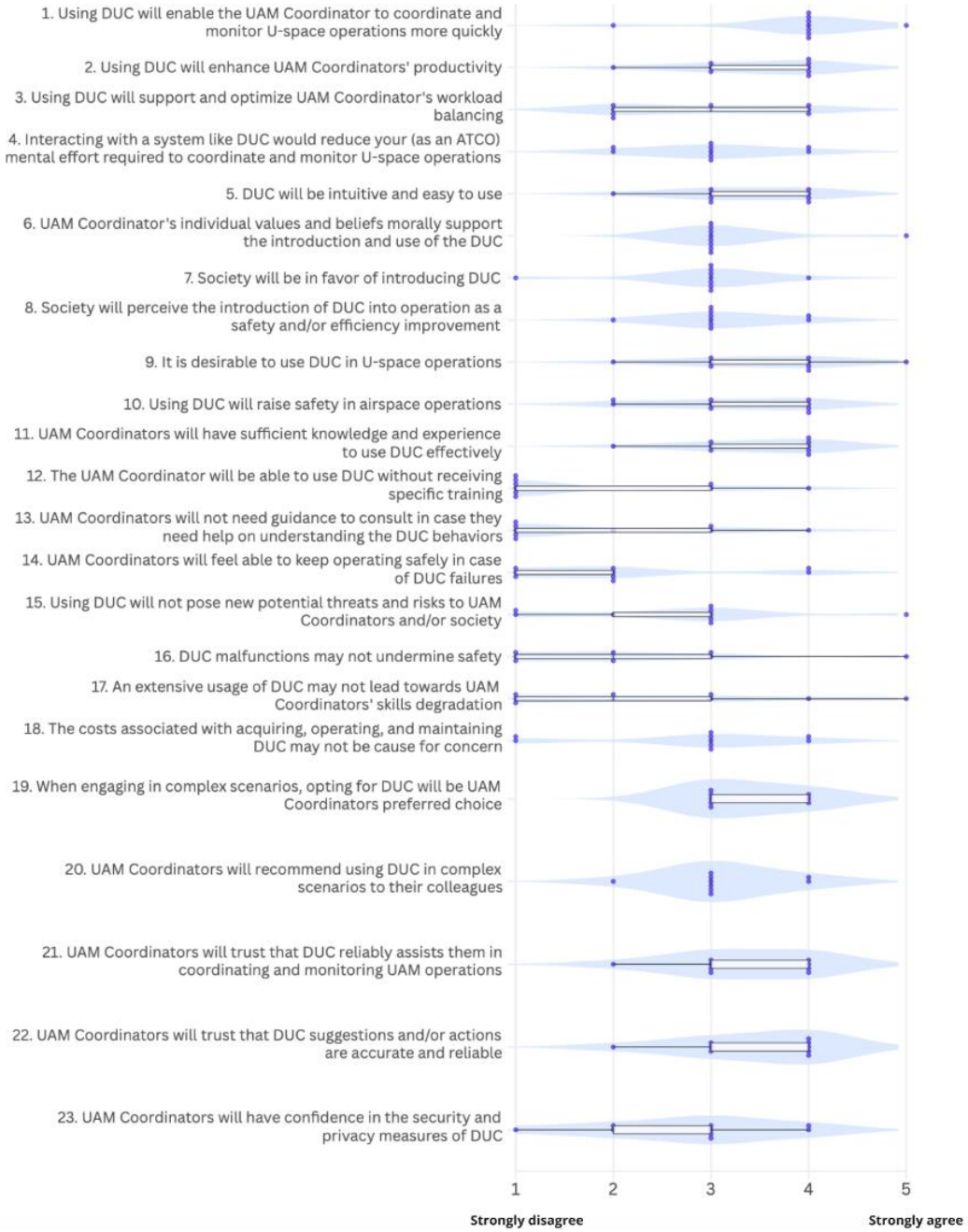


Figure 85. Boxplot of questionnaire responses to Social Acceptance Questionnaire.

Analysis

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- Neutral answers: Several statements have received neutral ratings (majority 3), indicating neither disagreement nor agreement. From motivations made by participants, several 3-ratings reflect confusion or uncertainty about how to interpret the statement.
    - Some of these statements refer to how "society" will perceive DUC and projected perceptions with the UAM Coordinator on how DUC will be perceived.
    - Questions asking participants to consider the UAM Coordinator's perceptions were likely difficult to answer given that ATCOs are not UAM Coordinators.
- More agreement (Median more than 3):
    - DUC supports efficiency in coordination and monitoring U-space operations [1];
    - DUC enhances productivity [2];
    - DUC will be intuitive and easy to use [5];
    - Desirable to use DUC in U-space operations [9];
    - UAM Coordinator will have knowledge to use DUC [11];
    - UAM Coordinators will trust DUC in assisting coordination and monitoring [21];
    - UAM Coordinators will trust DUC's actions/suggestion to be reliable and accurate [22];
- More disagreement (Median less than 3):
    - UAM Coordinators ability to use DUC without training [12];
    - UAM Coordinators need for guidance to understand DUC [13];
    - UAM Coordinators perceived safety in operations if DUC fails (DUC is needed) [14];
    - DUC malfunctions will undermine safety [16];
    - Extensive use of DUC will lead to skill degradation [17].
- Large variability between participants (large spread in answers)
    - Training required [12], perceived safety [13, 14, 16, 17]; confidence [23].

Participants' confusion about statements

- For 16 cells across 13 statements from six different participants answered with "3" have a motivation like: "I have no idea", "How should I know?", "Don't know", "I don't understand", "I have no real opinion", I can not answer that". One has "Badly formulated question"
- On 43 cells, across all participants, have motivations that cannot be understood, like "a", ".", "s", "bnm", "s", "g", "h", or "j"

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

9.2.11. Scenario 1 Results: situated vs. abstract

Figure 86. Participants responses to statement: How would you rate your workload in this scenario? No significant difference (p = .374).

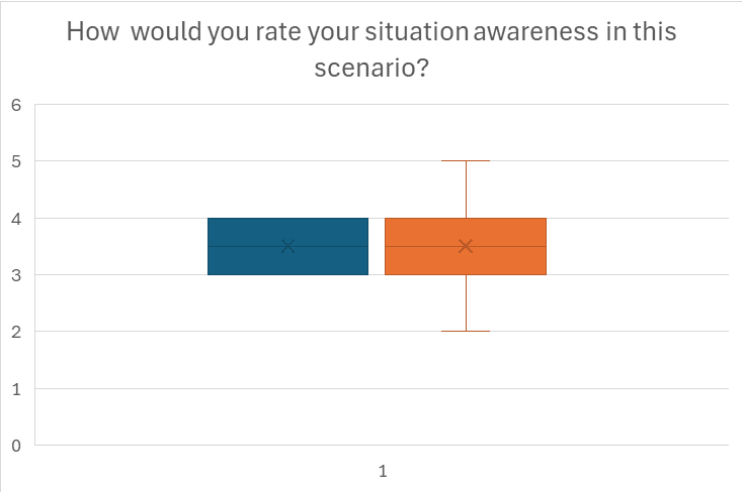


Figure 87. Participants responses to statement: How would you rate your situation awareness in this scenario? No significant difference (p = 1).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

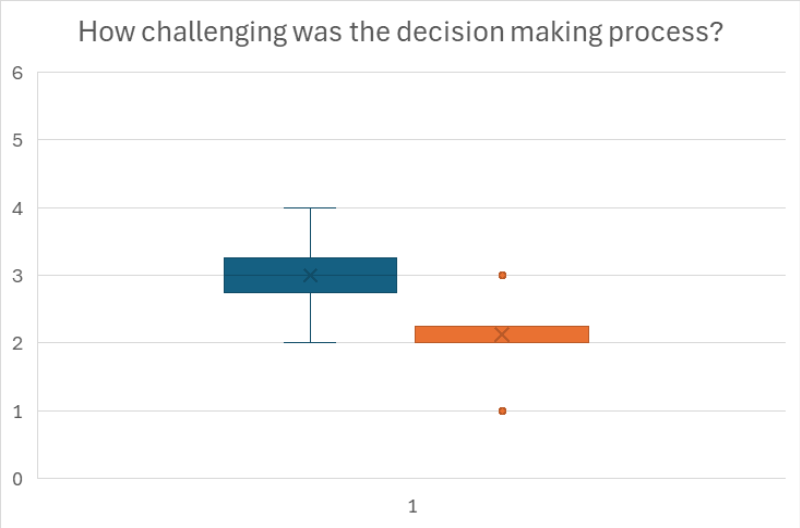


Figure 88. Participants responses to statement: How challenging was the decision-making process? The difference between the group was significant, with participants rating the decision-making process less challenging in the abstract condition (p = .0262).

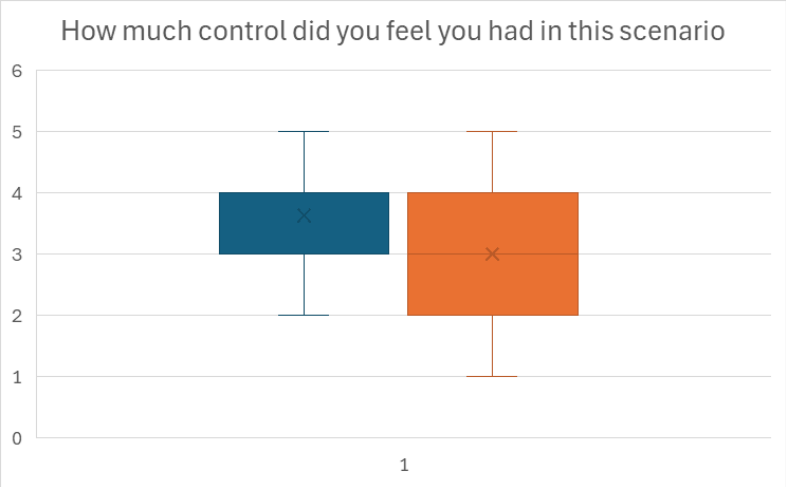


Figure 89. Participants responses to statement: How much control did you feel you had in this scenario? No significant difference (p = .440).

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

9.2.12. Scenario 2 results: storytelling vs text

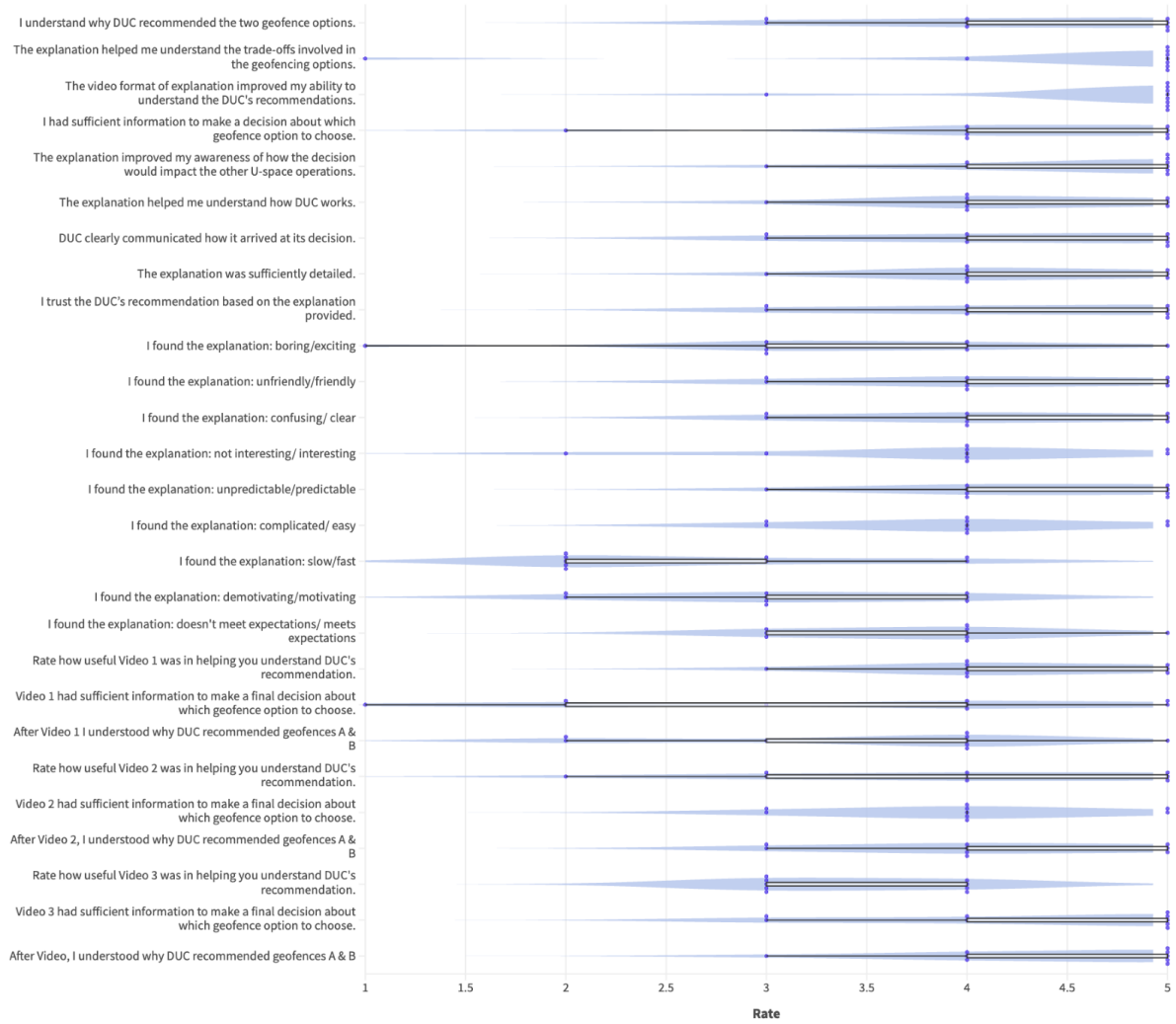


Figure 90. Boxplot of questionnaire responses to Storytelling format.

Understanding DUC's Recommendation

- "I understand why DUC recommended the two geofence options." Most responses cluster around 4 and 5 (Mdn=4), indicating strong comprehension.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- “After Video 1/2/3, I understood why DUC recommended geofences A & B.” Video 1 & 2 had consistently high ratings, with Video 2 showing slightly higher confidence levels.
- Video 3’s ratings were marginally lower, suggesting it added less new information compared to the first two.
- Explanation Quality & Clarity.
- “The explanation helped me understand the trade-offs involved in the geofencing options." A strong positive skew, Mdn=5, indicating that trade-offs were effectively communicated.
- "The video format of explanation improved my ability to understand the DUC's recommendations." had a very strong positive response with Mdn=5.
- “The explanation helped me understand how DUC works." Similar pattern to the previous question; respondents generally felt the explanation was informative.
- “DUC clearly communicated how it arrived at its decision." The majority found the explanation clear (Mdn=4), with very few lower ratings.

Decision-Making Confidence

- “I had sufficient information to make a decision about which geofence option to choose." While still positively skewed, there’s slightly more spread (Mdn=4), indicating mixed opinion that some respondents may have needed additional information.
- “The explanation improved my awareness of how the decision would impact the other U-space operations." Highly rated (Mdn=5), suggesting that the explanation effectively communicated broader implications.
- “I trust the DUC’s recommendation based on the explanation provided." The trust level in DUC’s recommendation is high (Mdn=4), implying that the explanation was persuasive and credible.

Explanation’s Engagement and Perceived Quality

- “I found the explanation: boring/exciting." More spread towards the middle (Mdn=3), indicating that some respondents found it neutral or slightly unengaging.
- This suggests room for improvement in making the explanation more engaging or interactive.
- “I found the explanation: unfriendly/friendly. "Generally rated as friendly (Mdn=4), meaning the tone was well-received.
- “I found the explanation: confusing/clear." Strong clustering at 4-5 (Mdn=4), meaning the majority found it clear.
- A small percentage rated it lower, indicating it might have been unclear for a few participants.
- “I found the explanation: not interesting/interesting, boring/exciting, slow/fast, motivating/demotivating questions, these received a more balanced distribution, suggesting mixed opinions and different perception of these qualities.
- “I found the explanation: unpredictable/predictable. "Highly predictable (Mdn=4), which is positive in terms of clarity but might suggest a lack of novelty or engagement.
- “I found the explanation: complicated/easy. "Generally rated easy to understand (Mdn=4), though a few found it complicated.

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- “I found the explanation: doesn’t meet expectations/meets expectations. "Most responses are high (4-5), indicating the explanation met or exceeded expectations.

CLT Effectiveness Analysis

- “Rate how useful Video 1 was in helping you understand DUC’s recommendation." Generally high ratings Mdn=4, suggesting Video 1 was a strong introduction.
- “Video 1 had sufficient information to make a final decision about which geofence option to choose.” had mixed opinions Some minor spread towards lower ratings, meaning it provided a good foundation but not complete information.
- After Video 1 I understood why DUC recommended geofences A & B had mixed opinions as well.
- “Rate how useful Video 2 was in helping you understand DUC’s recommendation.” had mixed opinions as well although higher then video 1.
- “Video 2 had sufficient information to make a final decision. "Similar to Video 1, with slightly stronger ratings, suggesting Video 2 addressed some gaps.
- “Rate how useful Video 3 was in helping you understand DUC’s recommendation.” Still highly rated, but slightly lower than Video 2, implying that it added less new value compared to earlier videos.
- “Video 3 had sufficient information to make a final decision." A small decline in ratings compared to Video 1 & 2, meaning it was helpful but may have been redundant or less impactful.
- After Video 3, I understood why DUC recommended geofences A & B Mdn=5

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

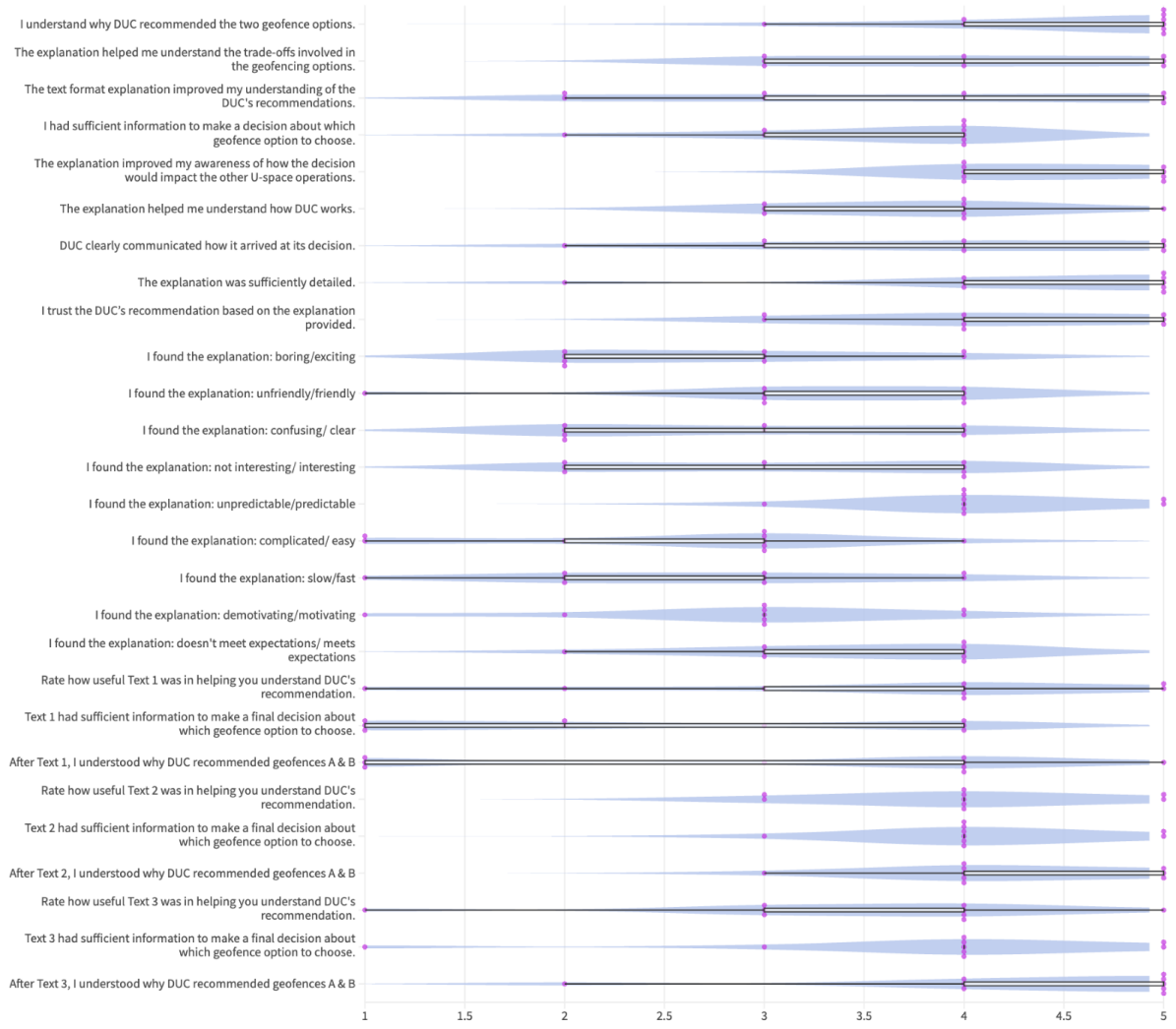


Figure 91. Boxplot of questionnaire responses to Text format.

Understanding DUC's Recommendation

- "I understand why DUC recommended the two geofence options." Responses cluster around 4 and 5 (Mdn=5), indicating strong comprehension of DUC's reasoning.
- "After Text 1/2/3, I understood why DUC recommended geofences A & B." Text 2 received higher ratings, showing it contributed significantly to understanding while Text 1 and 2 received mixed opinions indicating that they were not clear for everyone.

Explanation Quality & Clarity

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- “The explanation helped me understand the trade-offs involved in the geofencing options.” Mdn=4 rating, meaning the text format conveyed trade-offs effectively.
- “The explanation helped me understand how DUC works." Again, high ratings Mdn=4 indicate that the text-based explanation effectively communicated the decision-making process.
- “DUC clearly communicated how it arrived at its decision.” had a mixed opinion.
- “The explanation was sufficiently detailed.” had mixed opinions. Most responses are 4 and 5, indicating that most respondents felt the level of detail was appropriate. However, a few responses are in the 2-3 range.
- “I trust the DUC’s recommendation based on the explanation provided." Trust levels are high, suggesting the explanation was persuasive and credible.

Decision-Making Confidence

- “I had sufficient information to make a decision about which geofence option to choose.” had mixed opinions While most ratings are 4-5, there is some spread towards lower ratings (2-3).
- “The explanation improved my awareness of how the decision would impact the other factors." had mixed opinions as well.

Explanation’s Engagement & Perceived Quality

- All the adjectives have very mixed opinions and lean more towards negative side, so text explanation was perceived more as boring, confusing and complex.
- “I found the explanation: unfriendly/friendly. had a mixed opinion but with most focus on neutral rating.
- “I found the explanation: unpredictable/predictable." Responses are strongly skewed toward predictability, which suggests a structured and logical explanation but possibly lacking novelty.

Text-Based Explanation Effectiveness (Comparing Text 1, 2, and 3)

- “Rate how useful Text 1 was in helping you understand DUC’s recommendation." Majority high ratings although very mixed
- “Text 1 had sufficient information to make a final decision." Some minor spread towards lower ratings, meaning it provided useful but incomplete information.
- “Rate how useful Text 2 was in helping you understand DUC’s recommendation." Consistently high ratings, slightly higher than Text 1, indicating it built on the foundation effectively.
- “Text 2 had sufficient information to make a final decision." Like Text 1 but with slightly stronger ratings, meaning it addressed some gaps from the first explanation.
- “Rate how useful Text 3 was in helping you understand DUC’s recommendation." Still highly rated, but slightly lower than Text 2, implying that it added less new value compared to earlier texts.
- “Text 3 had sufficient information to make a final decision." A small decline in ratings compared to Text 1 & 2, meaning it was helpful but may have been redundant or less.

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Comparison between text and video:

- Understanding of DUC's recommendations was rated quite high for both text and video formats.
- Trust to the DUC recommendations was rated high for both formats.
- The improvement of awareness of how decision would impact other u-space users were rated for both quite high.
- Although when it comes to which format improved understanding of geofences text format received mixed opinions leaning towards positive. However, for video format it is only positive ratings.
- Video format was perceived as more clear, easy and friendly than text where opinions were spread and mixed.
- Boring/exciting, not/interesting, slow/fast received mixed opinions for both video and text formats.
- In both formats CLT 1 had very mixed opinions both for usefulness, sufficiency of information and helping to understand DUC.
- For both formats CLT2 was perceived as the most sufficient, useful and helping to understand DUC recommendations.
- CLT3 had more of positive feedback in video format in text format there are mixed opinions.

Key takeaways:

- Both formats were effective, with high ratings for comprehension of DUC's reasoning.
- CLT2 (Content Learning Task 2) was the most effective in both formats, helping participants understand the recommendations better.
- Both formats scored high in trust, indicating that the explanations were persuasive and credible. No significant difference in trust levels between video and text formats.
- Video slightly outperformed text in the impact on understanding.
- Video was perceived as clearer, easier, and more user-friendly than text.
- Text had more mixed opinions, with some respondents finding it less clear or harder to follow.
- Text was perceived as more neutral or slightly boring, with some respondents finding it slow or complex.
- Both formats had mixed opinions on being exciting/boring, fast/slow, and interesting/uninteresting.

Scenario 2 questionnaire statistics

Statistically significant statements are marked in green; trending statements are marked in yellow.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Table 33. Statistics for Storytelling/Text Questionnaire.

Statement	P-value	Median / IQR	Direction of finding
I understand why DUC recommended the two geofence options.	Z=0.966, p=0.334	video Mdn=4 IQR=2 text Mdn=5 IQR=1	preferred text
The explanation helped me understand the trade-offs involved in the geofencing options.	Z=1.207, p=0.227	video Mdn=5 IQR=1 text Mdn=4 IQR=2	preferred video
The video format of explanation improved my ability to understand the DUC's recommendations.	Z=1.709, p=0.088	video Mdn=5 IQR=0 text Mdn=4 IQR=3	preferred video
I had sufficient information to make a decision about which geofence option to choose.	Z=1.387, p=0.165	video Mdn=4 IQR=1 text Mdn=4 IQR=1	equal
The explanation improved my awareness of how the decision would impact the other U-space operations.	Z=0.378, p=0.705	video Mdn=5 IQR=1 text Mdn=4 IQR=1	preferred video
The explanation helped me understand how DUC works.	Z=1.414, p=0.157	video Mdn=4 IQR=1 text Mdn=4 IQR=1	equal
DUC clearly communicated how it arrived at its decision.	Z=0.828, p=0.408	video Mdn=4 IQR=2 text Mdn=4 IQR=2	equal
The explanation was sufficiently detailed.	Z=-0.276, p=0.783	video Mdn=4 IQR=1 text Mdn=5 IQR=1	preferred text
I trust the DUC's recommendation based on the explanation provided.	Z=0.333, p=0.739	video Mdn=4 IQR=2 text Mdn=4 IQR=2	equal
I found the explanation: boring/exciting	Z=1.095, p=0.273	video Mdn=3 IQR=1 text Mdn=3 IQR=2	equal
I found the explanation: unfriendly/friendly	Z=1.994 p=0.046	video Mdn=4 IQR=2 text Mdn=3 IQR=1	preferred video
I found the explanation: confusing/ clear	Z=2.209, p=0.027	video Mdn=4 IQR=2 text Mdn=3 IQR=2	preferred video
I found the explanation: not interesting/ interesting	Z=1.444, p=0.149	video Mdn=4 IQR=1 text Mdn=3 IQR=2	preferred video

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

I found the explanation: unpredictable/predictable	Z=0.816, p=0.414	video Mdn=4 IQR=1 text Mdn=4 IQR=1	equal
I found the explanation: complicated/ easy	Z=2.565, p=0.010	video Mdn=4 IQR=1 text Mdn=3 IQR=2	preferred video
I found the explanation: slow/fast	Z=0, p=1	video Mdn=2 IQR=2 text Mdn=3 IQR=2	preferred text
I found the explanation: demotivating/motivating	Z=0.535, p=0.593	video Mdn=3 IQR=2 text Mdn=3 IQR=1	equal
I found the explanation: doesn't meet expectations/ meets expectations	Z=1.342, p=0.180	video Mdn=4 IQR=1 text Mdn=4 IQR=1	equal
Rate how useful Video 1 was in helping you understand DUC's recommendation.	Z=1.186, p=0.236	video Mdn=4 IQR=1 text Mdn=4 IQR=2	equal
Video 1 had sufficient information to make a final decision about which geofence option to choose.	Z=1.378, p=0.168	video Mdn=4 IQR=3 text Mdn=2 IQR=3	preferred video
After Video 1 I understood why DUC recommended geofences A & B	Z=0.850, p=0.395	video Mdn=4 IQR=2 text Mdn=4 IQR=3	equal
Rate how useful Video 2 was in helping you understand DUC's recommendation.	Z=-0.378, p=0.705	video Mdn=4 IQR=2 text Mdn=4 IQR=1	equal
Video 2 had sufficient information to make a final decision about which geofence option to choose.	Z=-0.378, p=0.705	video Mdn=4 IQR=1 text Mdn=4 IQR=1	equal
After Video 2, I understood why DUC recommended geofences A & B	Z=-0.378, p=0.705	video Mdn=4 IQR=2 text Mdn=4 IQR=1	equal
Rate how useful Video 3 was in helping you understand DUC's recommendation.	Z=-0.137, p=0.891	video Mdn=3 IQR=1 text Mdn=4 IQR=1	preferred text
Video 3 had sufficient information to make a final decision about which geofence option to choose.	Z=1.098, p=0.272	video Mdn=5 IQR=2 text Md=4 IQR=1	preferred video
After Video, I understood why DUC recommended geofences A & B	Z=0.276, p=0.783	video Mdn=5 IQR=1 text Mdn=5 IQR=1	equal

Significant findings

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- *I found the explanation: unfriendly/friendly* (Z=1.994, p=0.046). Direction: Preferred video Conclusion: Participants found the video explanation significantly friendlier than the text version.
- *I found the explanation: confusing/clear* (Z=2.209, p=0.027). Direction: Preferred video Conclusion: Participants rated the video explanation as significantly clearer than text.
- *I found the explanation: complicated/easy* (Z=2.565, p=0.010). Direction: Preferred video Conclusion: Participants found the video significantly easier to understand than the text.

Most of the remaining items are not statistically significant, even if some show directional preferences. However, trends worth noting:

- Video trended toward being more helpful or clearer in several questions:
  - "The explanation helped me understand the tradeoffs" *(p=0.227)*
  - "The video format improved my ability to understand" *(p=0.088 – nearly significant)*
  - "Video 3 had sufficient info" *(p=0.272)*
- Text was preferred in:
  - "I understand why DUC recommended geofences" *(p=0.334)*
  - "The explanation was sufficiently detailed" *(p=0.783)*
  - "Video 3 usefulness" *(p=0.891)*

Conclusion

1. Video explanations are more effective than text in terms of being friendlier, clearer, and easier to understand (all significant).
2. For most other aspects (usefulness, informativeness, decision-making support), there was no significant difference.
3. Participants did not express significantly higher trust, satisfaction, or preference for one medium over the other in areas like motivation, usefulness, or interest.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

9.2.13. Scenario 3 results: attention guidance vs no attention guidance.

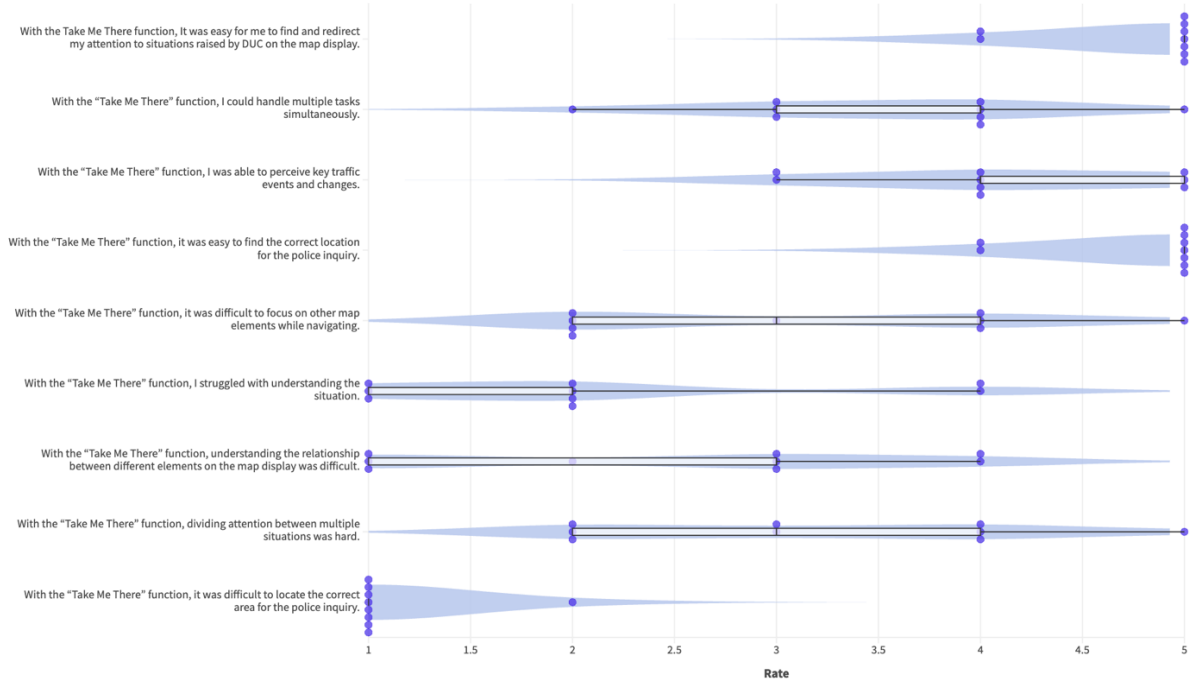


Figure 92. Boxplot of questionnaire responses to Attention guidance condition.

Locating the right area:

- Easy to find location raised by DUC: Overall easy. 4s and 5s, (mdn= 5). The function did what it was supposed to, if used. Users found it easy to redirect attention using the "Take Me There" function.
- Easy to find the correct location for the police inquiry: Overall easy. 4s and 5s, (mdn= 5). Need to check recording to see if they found the right place. Some stated the wrong place in the interview.
- It was difficult to locate the correct area for the police inquiry: Low (mdn=1).

Focus on multiple situations:

- I could handle multiple tasks simultaneously: Mixed responses (mdn=4). Some participants found it difficult to handle multiple tasks. Indicating it is more dependent on the participant more than the function.
- I was able to perceive key traffic events and changes: Mostly positive (mdn=4). Most could perceive key traffic events well. Question is if they knew what the key traffic events were.
- It was difficult to focus on other map elements while navigating: Mixed. (mdn=3) Indicating they had a hard time focusing on the "rest of the world".

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- I struggled with understanding the situation: Low. (mdn=2). They seem to understand the situation.
- Understanding the relationship between different elements on the map display was difficult: Lower but still a bit difficult (mdn=3). Users struggled with understanding map relationships, which might indicate a need for better visualization.
- Dividing attention between multiple situations was hard: Mixed (mdn=3).

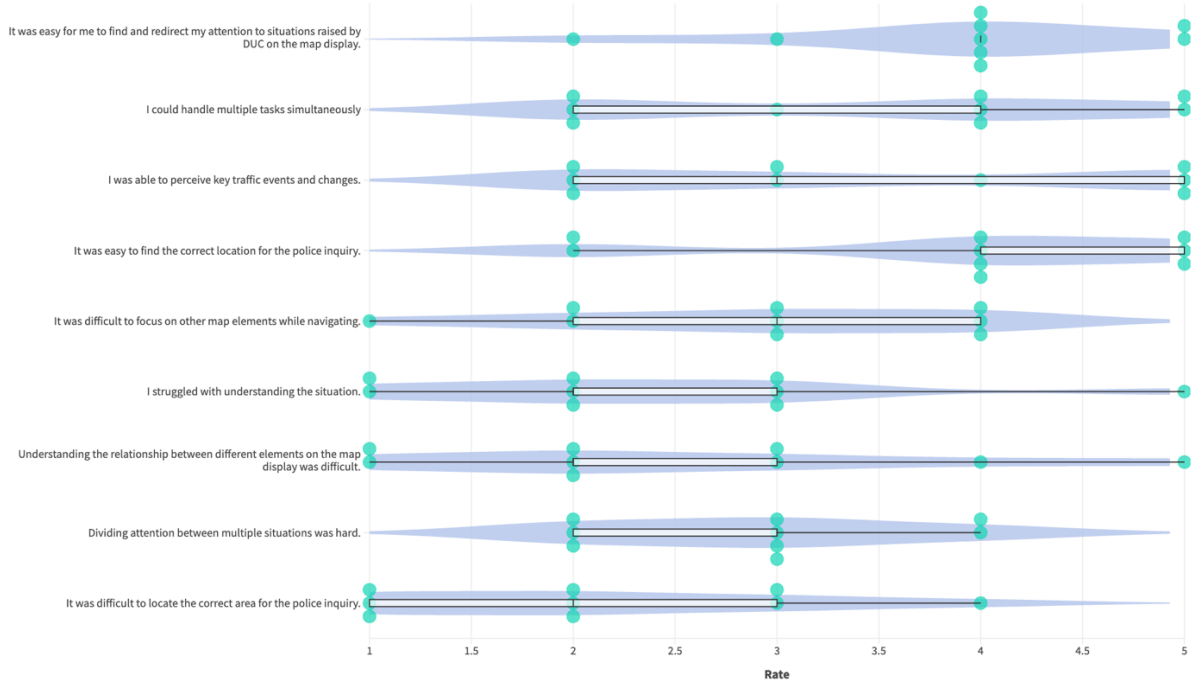


Figure 93. Boxplot of questionnaire responses to No attention guidance condition.

Ease of Attention and Task Handling

- It was easy for me to find and redirect my attention shows a broad distribution, with responses spanning from low to high ratings which suggests mixed opinions, although mdn=4 which indicates that the majority did not have problems with finding or redirecting attention.
- I could handle multiple tasks simultaneously has a concentrated distribution around 3-4 with mdn=4, suggesting that most participants rated this moderately but not extremely high.
- I was able to perceive key traffic events and changes is similar, with most responses clustered around 2-3 with mdn=3, indicating moderate and low ability to perceive key traffic elements.

Navigation and Situation Understanding

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- It was easy to find the correct location for the police inquiry has a higher concentration around 4 with mdn=4, meaning most participants found it relatively easy.
- It was difficult to locate the correct area for the police inquiry has lower ratings (1-2 dominant), showing that many participants did not find it difficult.
- Understanding the relationship between different elements on the map has responses mostly around 2 with mdn=2, showing difficulty.

Cognitive Load and Difficulty

- It was difficult to focus on other map elements while navigating has a range between 2-4 with mdn=3, meaning that opinions were mixed.
- I struggled with understanding the situation and Dividing attention between multiple situations was hard both have responses centred around 2-3 with mdn=2, meaning that some found these tasks not so difficult and some had neutral opinion about it.

Scenario 3 Questionnaire statistics

Statistically significant statements are marked in green; trending statements are marked in yellow.

Table 34. Statistics for Attention guidance/No attention guidance Questionnaire.

Statement	P-value	Median / IQR	Direction of finding
It was easy for me to find and redirect my attention to situations raised by DUC on the map display.	Z=2.060, p=0.039	NO: M=4 IQR=1 AG: M=5 IQR=1	Attention guidance
I could handle multiple tasks simultaneously	Z=0.345, p=0.730	NO: M=4 IQR=3 AG: M=4 IQR=1	Equal
I was able to perceive key traffic events and changes.	Z=1.656, p=0.098	NO: M=3 IQR=3 AG: M=4 IQR=2	Attention guidance
It was easy to find the correct location for the police inquiry.	Z=2.060, p=0.039	NO: M=4 IQR=2 AG: M=5 IQR=1	Attention guidance
It was difficult to focus on other map elements while navigating.	Z=0.427, p=0.669	NO: M=3 IQR=2 AG: M=3 IQR=2	Equal
I struggled with understanding the situation.	Z=-0.496, p=0.620	NO: M=2 IQR=2 AG: M=2 IQR=2	Equal
Understanding the relationship between different elements on the map display was difficult.	Z=-0.213, p=0.832	NO: M=2 IQR=2 AG: M=3 IQR=3	Attention guidance

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Dividing attention between multiple situations was hard.	Z=0.744, p=0.457	NO: M=3 IQR=2 AG: M=3 IQR=2	Equal
It was difficult to locate the correct area for the police inquiry.	Z=-1.983, p=0.047	NO: M=2 IQR=2 AG: M=1 IQR=0	No Attention guidance

9.3. UC4 Annex

9.3.1. Questions used for debriefing

Note: this is the full set of questions designed for VAL2, mapped to investigate the requirements (D4.1) and EASA macroareas - Cooperation/collaboration capabilities, Error Management, Explainability. Given the nature of the questions, we did not ask all of them every single time, but we adapted the script to each participant and experiment.

Table 35. HF Requirements: Cooperation/collaboration capabilities (EASA Macroarea) Requirements

EASA-like requirement (D4.1)	**Question**
HF-03: ISA must be able to always generate and show new sequences by constantly adapting to everything that is happening.	Q1: Do you feel that ISA is able to adapt the sequence to everything that is going on? Q2: How do you evaluate ISA's responsiveness to the ever-changing situation?
HF-05: ISA must be able to recognise a potential suboptimal sequence generated by user's interaction and suggest a better alternative to the ATCM	Q3: How did you feel ISA adapted to the situation where the pilot did not comply with your instruction?
HF-08: ISA must be able to always suggest the best possible solution to a sequence and suggest it to the ATCO, regardless of what the ATCO does	Q4: Do you think ISA has been suggesting the best possible solution every time, or were there situations where you would've acted differently?

Table 36. HF Requirements: Cooperation/collaboration capabilities (EASA Macroarea) Requirements

EASA-like requirement (D4.1)	**Question**

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

HF-28: ISA must tolerate and adjust to user manual inputs, by maintaining an ongoing sequence recommendation capability	For this one there was not a dedicated question, but rather we observed how ISA reacted to what the ATCOs were doing
HF-29: ISA must detect and adjust to user manual inputs, by maintaining an ongoing sequence recommendation capability	For this one there was not a dedicated question, but rather we observed how ISA reacted to what the ATCOs were doing

Table 37. XAI Requirements

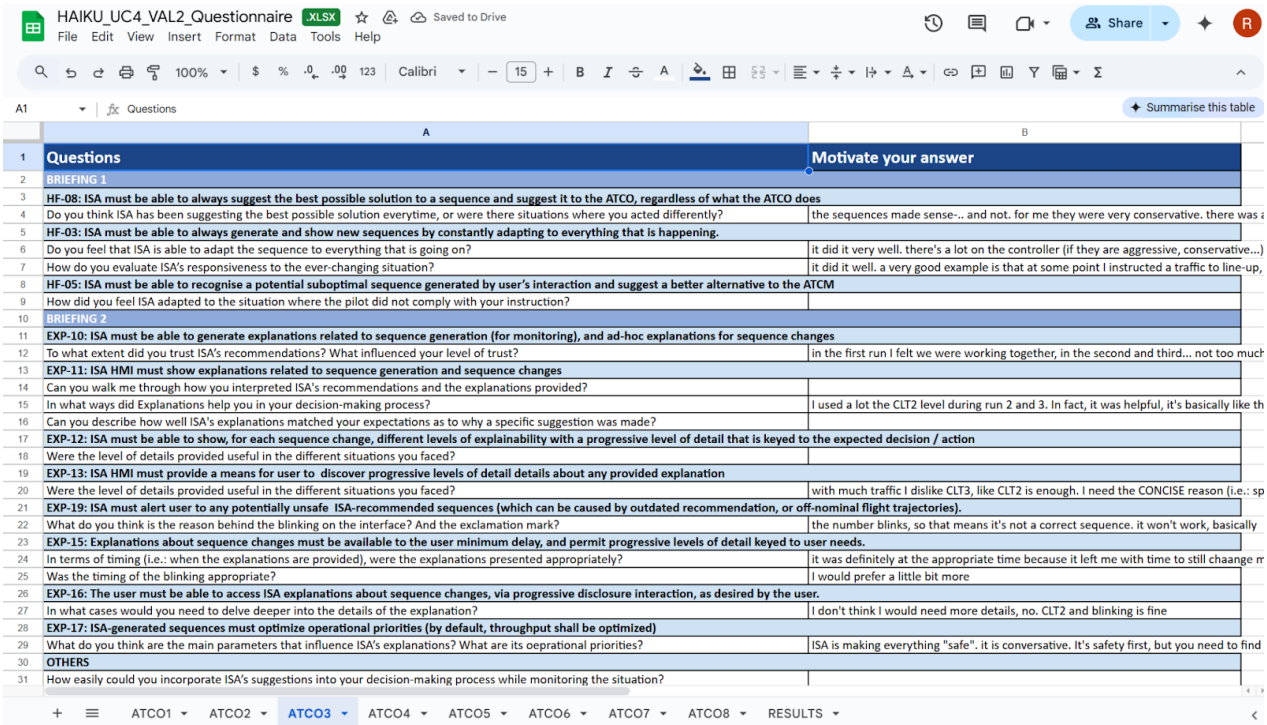
EASA-like requirement (D4.1)	**Question**
EXP-10: ISA must be able to generate explanations related to sequence generation (for monitoring), and ad-hoc explanations for sequence changes	Q5: Can you describe how well ISA's explanations matched your expectations as to why a specific suggestion was made? Q6: To what extent did you trust ISA's recommendations? What influenced your level of trust? Q7: In what ways did Explanations help you in your decision-making process?
EXP-11: ISA HMI must show explanations related to sequence generation and sequence changes	Q8: Can you walk me through how you interpreted ISA's recommendations, and the explanations provided?
EXP-12: ISA must be able to show, for each sequence change, different levels of explainability with a progressive level of detail that is keyed to the expected decision / action	Q9: Were the level of details provided useful in the different situations you faced?
EXP-13: ISA HMI must provide a means for user to discover progressive levels of detail details about any provided explanation	Q10: Were the level of details provided useful in the different situations you faced?
EXP-15: Explanations about sequence changes must be available to the user minimum delay and permit progressive levels of detail keyed to user needs.	Q11: In terms of timing (i.e.: when the explanations are provided), were the explanations presented appropriately? Q12: Was the timing of the blinking appropriate?
EXP-16: The user must be able to access ISA explanations about sequence changes, via progressive disclosure interaction, as desired by the user.	Q13: In what cases would you need to delve deeper into the details of the explanation?

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

EXP-17: ISA-generated sequences must optimize operational priorities (by default, throughput shall be optimized)	Q14: What do you think are the main parameters that influence ISA's explanations?
EXP-19: ISA must alert user to any potentially unsafe ISA-recommended sequences (which can be caused by outdated recommendation, or off-nominal flight trajectories).	Q15: What do you think is the reason behind the blinking on the interface?

9.3.2. Example of Data Collection

During VAL2 exercises, one of the experiments was taking notes regarding Observations and the results of semi-structured interviews in an Excel spreadsheet, which was later used for the qualitative analysis of data with the extractions of main themes and insights.



HAIKU_UC4_VAL2_Questionnaire .XLSX Saved to Drive

File Edit View Insert Format Data Tools Help

Share

A1 fx Questions

Summarise this table

	A	B
1	**Questions**	**Motivate your answer**
2	**BRIEFING 1**	
3	**HF-08: ISA must be able to always suggest the best possible solution to a sequence and suggest it to the ATCO, regardless of what the ATCO does**	
4	Do you think ISA has been suggesting the best possible solution everytime, or were there situations where you acted differently?	the sequences made sense-.. and not. for me they were very conservative. there was a
5	**HF-03: ISA must be able to always generate and show new sequences by constantly adapting to everything that is happening.**	
6	Do you feel that ISA is able to adapt the sequence to everything that is going on?	it did it very well. there's a lot on the controller (if they are aggressive, conservative...)
7	How do you evaluate ISA's responsiveness to the ever-changing situation?	it did it well. a very good example is that at some point I instructed a traffic to line-up,
8	**HF-05: ISA must be able to recognise a potential suboptimal sequence generated by user's interaction and suggest a better alternative to the ATCM**	
9	How did you feel ISA adapted to the situation where the pilot did not comply with your instruction?	
10	**BRIEFING 2**	
11	**EXP-10: ISA must be able to generate explanations related to sequence generation (for monitoring), and ad-hoc explanations for sequence changes**	
12	To what extent did you trust ISA's recommendations? What influenced your level of trust?	in the first run I felt we were working together, in the second and third... not too much
13	**EXP-11: ISA HMI must show explanations related to sequence generation and sequence changes**	
14	Can you walk me through how you interpreted ISA's recommendations and the explanations provided?	
15	In what ways did Explanations help you in your decision-making process?	I used a lot the CLT2 level during run 2 and 3. In fact, it was helpful, it's basically like th
16	Can you describe how well ISA's explanations matched your expectations as to why a specific suggestion was made?	
17	**EXP-12: ISA must be able to show, for each sequence change, different levels of explainability with a progressive level of detail that is keyed to the expected decision / action**	
18	Were the level of details provided useful in the different situations you faced?	
19	**EXP-13: ISA HMI must provide a means for user to discover progressive levels of detail details about any provided explanation**	
20	Were the level of details provided useful in the different situations you faced?	with much traffic I dislike CLT3, like CLT2 is enough. I need the CONCISE reason (i.e.: sp
21	**EXP-19: ISA must alert user to any potentially unsafe ISA-recommended sequences (which can be caused by outdated recommendation, or off-nominal flight trajectories).**	
22	What do you think is the reason behind the blinking on the interface? And the exclamation mark?	the number blinks, so that means it's not a correct sequence. it won't work, basically
23	**EXP-15: Explanations about sequence changes must be available to the user minimum delay, and permit progressive levels of detail keyed to user needs.**	
24	In terms of timing (i.e.: when the explanations are provided), were the explanations presented appropriately?	it was definitely at the appropriate time because it left me with time to still chaange m
25	Was the timing of the blinking appropriate?	I would prefer a little bit more
26	**EXP-16: The user must be able to access ISA explanations about sequence changes, via progressive disclosure interaction, as desired by the user.**	
27	In what cases would you need to delve deeper into the details of the explanation?	I don't think I would need more details, no. CLT2 and blinking is fine
28	**EXP-17: ISA-generated sequences must optimize operational priorities (by default, throughput shall be optimized)**	
29	What do you think are the main parameters that influence ISA's explanations? What are its oeprational priorities?	ISA is making everything "safe". it is conversative. It's safety first, but you need to find
30	**OTHERS**	
31	How easily could you incorporate ISA's suggestions into your decision-making process while monitoring the situation?	

ATCO1 ATCO2 ATCO3 ATCO4 ATCO5 ATCO6 ATCO7 ATCO8 RESULTS

Figure 94. Data Collection example for questionnaires.

Here's one of the most interesting situations and answers during VAL2 with ATCO 03:

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Questions	Motivate your answer
BRIEFING 1	
HF-08: ISA must be able to always suggest the best possible solution to a sequence and suggest it to the ATCO, regardless of what the ATCO does	
Do you think ISA has been suggesting the best possible solution everytime, or were there situations where you acted differently?	the sequences made sense-.. and not. for me they were very conservative. there was a tight situation in the exercise, I tried to do something very aggressive in spite of ISA and actually ISA was right (proposed a more conservative approach)
HF-03: ISA must be able to always generate and show new sequences by constantly adapting to everything that is happening.	
Do you feel that ISA is able to adapt the sequence to everything that is going on?	it did it very well. there's a lot on the controller (if they are aggressive, conservative...) but then it adapts
How do you evaluate ISA's responsiveness to the ever-changing situation?	it did it well. a very good example is that at some point I instructed a traffic to line-up, and ISA advised against it and made me think about possible new situations. ISA helped me CHANGE MY PLAN. I made a mistake and ISA bascally told me "it won't work, watch out". It gave me time to think and it was not too tight. it is safe
HF-05: ISA must be able to recognise a potential suboptimal sequence generated by user's interaction and suggest a better alternative to the ATCM	
How did you feel ISA adapted to the situation where the pilot did not comply with your instruction?	
BRIEFING 2	
EXP-10: ISA must be able to generate explanations related to sequence generation (for monitoring), and ad-hoc explanations for sequence changes	
To what extent did you trust ISA's recommendations? What influenced your level of trust?	in the first run I felt we were working together, in the second... not too much. in the first run ISA gave me options and I do what I think. When I was wrong it told me I was wrong and I had time to fix the mistake. In the first run I fully agreed with ISA's recommendations. ISA gained my trust when it advised me to fix that mistake, that was the best part of the exercise for me
EXP-11: ISA HMI must show explanations related to sequence generation and sequence changes	
Can you walk me through how you interpreted ISA's recommendations and the explanations provided?	
In what ways did Explanations help you in your decision-making process?	I used a lot the CLT2 level during run 2 and 3. In fact, it was helpful, it's basically like
Can you describe how well ISA's explanations matched your expectations as to why a specific suggestion was made?	
EXP-12: ISA must be able to show, for each sequence change, different levels of explainability with a progressive level of detail that is keyed to the expected decision / action	
Were the level of details provided useful in the different situations you faced?	

ATCO1 ATCO2 ATCO3 ATCO4 ATCO5 ATCO6 ATCO7 ATCO8 +

Figure 95. Data Collection example for ATCO 03.

9.3.3. Example of Data Analysis

Below is a simplified example illustrating how we analysed participant responses to the questionnaire. First, responses were categorised using **codes**. These codes were then grouped to form **themes**, which were used to address the overarching **research questions**.

For instance, consider the following research question:

RQ: *What levels and types of explanations are most effective for enabling ATCOs to understand and trust ISA-generated sequence changes during varying operational demands?*

To address this, we examined participant responses related to explainability, specifically, answers to questions linked to requirements EXP-10 to EXP-19.

Here are two example responses and their associated codes:

- *"The details were just enough, however I'd like even less: mouseover (CLT1) could just show symbols and be even shortened."*
  → **Code:** XAI Details

- *"They were useful to understand the sequence in terms of timing, like, to see how things were unfolding."*
  → **Code:** XAI Temporality

These and similar codes were then synthesised into the following theme:

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Theme: *The level of detail and temporal aspect of XAI explanations are generally effective, though there is room for improvement.*

Please note that this is a simplified illustration. In the full analysis, a larger number of responses were coded and clustered into themes, which were then mapped to the research questions. The key insights generated from this process are presented in the final conclusions.

9.3.4. Example of Test Run Logs

Here are some examples of the test run logs used for bug fixing and improving ISA's behaviour before VAL2:

Exercise	**HAIKU_01 RWY-10**	**Date:**	**23/01/2025**	**Time Log:**	**11:00**

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

· ISA ACTION: No TXI displayed in ISA.

- ATCO ACTION: ANE3544 cleared to land.

· ISA ACTION: ISA sets ANE3544 to N1 – (Correct)

· SCENARIO DEVELOPMENT: DHL4025 on stand 22, startup completed. Transponder is on.

· SCENARIO DEVELOPMENT: EXS23W completing startup.

- ATCO ACTION: DHL4025 instructed to taxi to holding point (HP) A5.

- ATCO ACTION: EXS23W pushback approved, facing south (S).

· ISA ACTION: ISA does not account for TWY. No information provided on TWY Bay.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- SCENARIO DEVELOPMENT: DHL4025 arrives at Holding Point A5 earlier than ETD, on time to become N2 in sequence. ISA does not display DHL4025 at HP-A5 in the sequence.
- SCENARIO DEVELOPMENT: EXS23W completes pushback and is ready to taxi.
- ATCO ACTION: Instruct DHL4025 to line up and wait behind ANE3544, confirming DHL4025 as N2 in ATCO's sequence.
- ATCO ACTION: EXS23W instructed to taxi to A6 via Gate D.
- ISA ACTION: ISA does not reflect DHL4025 on the runway. DHL4025 is now N2, but the sequence remains unchanged. N1 is ANE3544 as it is still on the RWY. NAX1BY should be N3
- SCENARIO DEVELOPMENT: Next in sequence is NAX1BY, 4–5 NM final. ISA does not adjust to N3.
- SCENARIO DEVELOPMENT: ANE3544 vacates the runway.
- ATCO ACTION: DHL4025 cleared for takeoff.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- ISA ACTION: ISA does not reflect sequence change. No strip for DHL4025. NAX1BY is now N1.
- ACTO ACTION: ISA bay movement updated; ANE3544 moved to TWY, taxiing to Stand E4.
- SCENARIO DEVELOPMENT: DHL4025 rotates and vacates the runway. EXS23W is taxiing to A6.
- ATCO ACTION: NAX1BY cleared to land.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

· ISA ACTION: NAX1BY is now N1, making AEA4024 N3. *(Possible reason: ISA appears to account for EXS23W as N2, but no strip is displayed in the ISA TWY bay.)*

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- Exercise Paused: Paused to acquire and reflect observed data.
- Termination: Test Run 1 terminated due to TWY strips not being displayed.

Observation Notes:

· **Sequence Handling**:

ISA appears to have accounted for DHL4025 in the sequence only after the aircraft was cleared for takeoff. However, there was a significant delay in recognizing DHL4025 as fully ready at A5. ISA seems more reactive to the scenario rather than predictive.

· **Sequence Update**: NAX1BY was observed as N1 after the departure of DHL4025.

· **Sequence Adjustment**: ISA adjusted the sequence by making AEA4024 N3 and NAX1BY N1. *(This suggests that ISA accounted for EXS23W as N2, thereby designating AEA4024 as N3.)*

· **Simulation Behaviour**: ISA halts time in accordance with the simulator being paused.

Behaviour suggestions:

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

· **Sequence Adjustment for Line-Up Changes**:

When an ATCO decides to line up an aircraft *behind* another that is currently using the runway, and there is another aircraft on final, this action seem to signify a change in the sequence. The traffic lined up on the runway should be considered **N1**, while the traffic on final becomes **N2**.

· **Priority Sequence Change:**

A priority sequence adjustment could be introduced (if not already implemented) to recognize this action (L&H) as a change in sequence. This would ensure that the lined-up traffic is designated as N1 and the traffic on final as N2.

· **Holding Point Readiness:**

Aircraft at holding points may require different times to be fully prepared for takeoff. Lining up and holding (L&H) maneuvers are a critical indicator of which aircraft the ATCO intends to prioritize, based on pilot readiness.

· **Early Readiness and RWY Sequence Updates:**

Aircraft ready ahead of schedule, but within acceptable time margins, may move up in the runway sequence. Taxiway (TWY) maneuvers could provide valuable insights into estimated airborne times and could be leveraged by ISA for better sequencing predictions.

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

10. References

Ashoori, M., & Weisz, J. D. (2019). In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes (arXiv:1912.02675). arXiv. https://doi.org/10.48550/arXiv.1912.02675

EASA (2024). EASA Artificial Intelligence (AI) Concept Paper Issue 2: Guidance for Level 1&2 machine learning applications, March 2024. Online at: https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-concept-paper-issue-2

EASA (2023). EASA Concept Paper: First usable guidance for Level 1 & 2 machine learning applications (proposed Issue 2), released 24 Feb 2023 (https://www.easa.europa.eu/en/newsroom-and-events/news/easa-artificial-intelligence-concept-paper-proposed-issue-2-open). European Union Aviation Safety Agency.

EASA (2020). Artificial Intelligence Roadmap: A Human-centric Approach to AI in Aviation. European Union Aviation Safety Agency. Version 2.0. February 2020.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human factors, 37(1), 32-64.

Gursoy, D., Chi, O. H., Lu, L., & Nunkoo, R. (2019). Consumers acceptance of artificially intelligent (AI) device use in service delivery. International Journal of Information Management, 49, 157-169. https://doi.org/10.1016/j.ijinfomgt.2019.03.008

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Advances in Psychology (Vol. 52, p. 139-183). Elsevier. https://doi.org/10.1016/S0166-4115(08)62386-9

Hellmann, M., Hernandez Bocanegra, D., & Ziegler, J. (2022). Development of an Instrument for Measuring Users' Perception of Transparency in Recommender Systems. https://doi.org/10.17185/DUEPUBLICO/75905

Lewis, J. R. (1995). Measuring perceived usability: The CSUQ, SUS, and UMUX. International Journal of Human–Computer Interaction, 34(12), 1148-1156.

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. International Journal of Human-Computer Interaction, 14(3-4), 463-488.

Lounis, C. A. (2020). Monitor the monitoring: pilot assistance through gaze tracking and visual scanning analyses (Doctoral dissertation, Toulouse, ISAE).

Rosenthal-von der Pütten, A. M., & Krämer, N. C. (2014). How design characteristics of robots determine evaluation and uncanny valley related responses. Computers in Human Behavior, 36, 422-439. https://doi.org/10.1016/j.chb.2014.03.066

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332