EXPLORATORY RESEARCH

sesar JOINT UNDERTAKING

Deliverable 5.2 Report on Human-Machine Distributed Situation Awareness

Deliverable ID:	**D5.2**
Dissemination Level:	**PU**
Project Acronym:	**AISA**
Grant:	**892618 — AISA**
Call:	**H2020-SESAR-2019-2**
Topic:	**SESAR-ER4-01-2019**
Consortium Coordinator:	**FTTS**
Edition date:	**17 June 2022**
Edition:	**00.01.01**
Template Edition:	**02.00.03**

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Authoring & Approval

Authors of the document

Name / Beneficiary	Position / Title	Date
Ruth Häusler Hermann / ZHAW	Task 5.1/5.3 Leader	31 March 2022
Celina Vetter / ZHAW	Project Associate	31 March 2022
Kristina Samardžić / FTTS	Project Associate	05 April 2022
Dorea Antolović / FTTS	Project Associate	05 April 2022
Ivan Tukarić / FTTS	Project Associate	05 April 2022
Manuel Roth / ZHAW	Project Associate	31 March 2022
Luca Tensfeldt / ZHAW	Project Associate	15 April 2022
Mia Bazina / FTTS	Project Associate	15 April 2022
Jennifer Burkhalter / Skyguide	ATCO, Project Associate	30 April 2022

Reviewers internal to the project

Name / Beneficiary	Position / Title	Date
Tomislav Radišić / FTTS	Project Coordinator	05 May 2022
Chrisoph Herberth / Skyguide	ATCO, Project Associate	06 May 2022
Keiko Moebus / Skyguide	Head of Human Factor, Project Associate	06 May 2022

Reviewers external to the project

Name / Beneficiary	Position / Title	Date

Approved for submission to the SJU By – Representatives of all beneficiaries involved in the project

Name / Beneficiary	Position / Title	Date
Tomislav Radišić/FTTS	Project Coordinator	16 May 2022

Rejected By – Representatives of beneficiaries involved in the project

Name and/or Beneficiary	Position / Title	Date

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Document History

Edition	Date	Status	Name / Beneficiary	Justification
00.00.01	31/03/2022	Initial Draft	Ruth Häusler Hermann	New document
00.00.02	27/04/2022	Draft	Ivan Tukarić	New sections added
00.00.03	05/05/2022	Final draft	Tomislav Radišić	First final draft
00.00.04	12/05/2022	Comments integrated	Celina Vetter	Comments integrated to final draft
00.01.00	16/05/2022	First Issue	Ruth Häusler Hermann	First Issue
00.01.01	17/06/2022	First revised version	Ruth Häusler Hermann	First revision of the document

AISA

AI SITUATIONAL AWARENESS FOUNDATION FOR ADVANCING AUTOMATION

This experimental study is part of a project that has received funding from the SESAR 3 Joint Undertaking under grant agreement No 892618 under European Union's Horizon 2020 research and innovation programme.

Preface

View of an Air Traffic Controller

Wednesday 4th of May in the afternoon at a control sector in Zurich. The sky is full of cumulus nimbus cells, and I have 20 aircraft on frequency which need to be guided safely through the airspace and around the weather cells. I am an Air Traffic Controller working in the Area Control Centre in Dubendorf, Zurich for 12 years but each time I have a situation like this my pulse rises and I am focused and fully aware that I am not allowed to make any mistakes. As an ATCO my main job is to guide airplanes safely and efficiently through the air. In stressful situations such as poor weather and heavy traffic, it is hard to always keep the situation awareness as high as desired. In situations with a high level of stress, a machine which supports us would be a great benefit and an additional safety net. At Skyguide we have a great set of tools which already work fine and support us in all kinds of situations. The future though points in the direction of Artificial Intelligence and that's what is described in the following paper.

Jennifer Burkhalter, ATCO at Skyguide

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Abstract

This report presents the results of two simulation experiments performed with an AI-based situation awareness system (AI SA system) developed in the AISA project to check the accuracy of the AI SA system's estimations and predictions and its capability to contribute to human-machine team situation awareness. It represents the AISA project deliverable 5.2 *Report on Human-Machine Distributed Situation Awareness* and contains four topical sections – described below – that address requirements to fulfil the project tasks 5.1 *Comparison of SA between AI and ATCO* and task 5.3 *Human performance in distributed SA*. The task 5.2 *Risk assessment of AISA* is covered in a separate deliverable D5.1 *Risk assessment report*.

- Topical section 1: Measurement of ATCO situation awareness and scanning behaviour
- Topical section 2: Comparison of human and machine situation awareness
- Topical section 3: Exploration of human-machine team situation awareness and its impact on human performance
- Topical section 4: Accuracy of AI SA system's estimations and predictions and its level of situation awareness

Two simulations were conducted with licensed Air Traffic Controllers working as radar executive. Situation awareness was assessed with multiple methods. The probe technique was applied to compare compare human and artificial situation awareness. ATCOs' experience with AI-based machine situation awareness (receiving "AI SA inputs") and its impact on performance were explored. Post hoc simulations with data collected in experiment 1 were conducted to assess the accuracy of AI SA systems' estimations and predictions.

Main findings per topical section are:

1. ATCOs with preserved situation awareness have characteristic scanning behaviour: Their gaze is less fixed on aircraft or conflicts, and they filter out more effectively non-critical information than ATCOs with degraded situation awareness do.
2. Partial agreement of human and machine situation awareness on conflict detection. Both human and AI SA system missed conflicts (false negative) and named conflicts that were not present (false positive). The AI SA system is better at monitoring non-obvious/unexpected aspects (e.g., non-conformances).
3. ATCOs detected some conflicts earlier and solved them faster when they received AI SA inputs compared to working without AI SA inputs. Input modality (oral messages) was inadequate due to distraction and additional workload.
4. Successful automation of 46 out of 57 en-route air traffic monitoring tasks. Accuracy of Machine Learning module predictions for CD tested (70%): Partly results were inaccurate, and predictions were partly inconsistent. Plausibility checks on CD module's inputs and outputs were successful.

Limitations reduce the validity of situation awareness measurement for ATCOs (use of an unfamiliar simulation tool), the significance of the results (exploration of human-machine team situation awareness was done with early-stage implementation of the AI SA system and with inadequate design of HMI inducing additional workload on ATCOs).

The results generally support the proof-of-concept system of the AISA project in its ability to accomplish en-route air traffic management tasks. Further improvement of accuracy is needed for

EUROPEAN PARTNERSHIP

Co-funded by the European Union

machine learning modules. Accuracy per se is not sufficient, considerable effort needs to be spent on solutions on how to integrate machine situation awareness. A long anticipation span is desirable for optimisation but does not comply with ATCOs' need for prioritization of tasks and information. The HMI of the future AI SA system will need distinctive ways of informing ATCOs about aspects of higher or lower urgency. Half of the participating ATCOs were willing to trust future AI-based tools – even after partially unfavourable experiences with an AI-based SA system in the experiment, about one third is neutral and a fifth is negative about including AI in tools.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Executive Summary

AI situation awareness system is an AI-based *machine situation awareness system* that was developed to accomplish en-route air traffic monitoring tasks. Its capability to fulfil that purpose as well as the usefulness of machine situation awareness as a contribution to human-machine team situation awareness were subsequently investigated in two experimental studies using human-in-the-loop simulation with ESCAPE Light from EUROCONTROL. The settings included only single ATCOs in the role of radar executive, no radar planner was involved. The research focuses on whether machine situation awareness would be capable of developing situation awareness and subsequently sharing it with ATCOs.

An initial experiment was conducted in November 2021 with 20 licensed ATCOs. The aim was to assess the individual ATCOs' situation awareness and subsequently compare it with the artificial situation awareness. Multiple methods were used to measure the ATCOs' situation awareness: subjective rating (SASHA_Q), gaze analysis by eye-tracking, and implicit measurement of performance. Simulation data gathered in the initial experiment was then used as input for the AI situation awareness system to process and generate machine situation awareness. This was done with probe technique using SPARQL queries for specific aspects of the situation. This step was necessary because the AI situation awareness system could not compute machine situation awareness in real-time at the stage of project-level implementation.

A second experiment was conducted in January 2022. The SPARQL query outputs of the AI situation awareness system were translated into oral inputs and given to 16 licensed ATCOs. Only three of these ATCOs had also been involved previously in the initial experiment. In one scenario, participants were able to freely interact with pilots (interactive condition). For the rest of the scenarios, ATCOs observed and implemented actions that were previously recorded in experiment 1 ("watch only" condition). This was necessary because pre-calculated AI situation awareness inputs would not match with ATCOs' manipulations. ATCOs' situation awareness was assessed with the same methods as in the initial experiment, with the addition of probe technique (SASHA_L). This allowed for a direct comparison between machine and ATCO situation awareness on identical queries about specific aspects of the situation. In addition, the contribution of the AI situation awareness system to human-machine team situation awareness and human performance was explored. ATCOs were asked to judge the usefulness of AI situation awareness inputs and to provide feedback on their experience interacting with the AI-based tool.

After the completion of the second experiment, the accuracy of the AI situation awareness system was further improved by implementing the remaining tasks (46 out of 57, including an additional one) for en-route air traffic monitoring (AISA project level implementation of machine situation awareness). Further simulations were conducted in April 2022 to precisely quantify the accuracy and functionality of the AI situation awareness system's estimations and predictions at the project-level stage of implementation. Based on the data collected in the initial experiment, AI situation awareness system's estimations and predictions for machine situation awareness were re-calculated. They were then compared with data from the simulated scenario progress. Any further ATCO interventions were excluded. This made it possible to check the correctness of machine situation awareness using precise data. In addition, the sensitivity (type II error: false negative) and the probability of false alarms (type I error: false positive

The results of the initial experiment showed variations in ATCOs' situation awareness based on gaze analysis for prioritisation of attention. In the second experiment, the ATCOs' situation awareness was

EUROPEAN PARTNERSHIP
Co-funded by the European Union

generally not as complete as artificial situation awareness outputs from an early implementation stage version of the AI situation awareness system. The latter – aside from generally being more complete – suffered also from false alarms and misses. Performance in the condition *with artificial situation awareness inputs* (second experiment) was slightly better for some of the scenarios, but worse for others compared to performance in the condition *without artificial situation awareness inputs* (first experiment). The majority of the participating ATCOs evaluated artificial situation awareness inputs at the early stage of implementation as being rather irrelevant. What they appreciated the most were the artificial situation awareness inputs about non-conformances of aircraft to their instructions.

Accuracy of the AI situation awareness system at the project-level of implementation (with most KG monitoring tasks implemented) showed to be high. Moreover, the KG monitoring tasks performed successfully in different traffic scenarios which proved their robustness. Several indicators of situation awareness degradation were defined based on which KG monitoring tasks were evaluated and compared to human situation awareness. Besides KG monitoring tasks, machine learning module predictions are compared to actual values. The majority of KG tasks deal with the conflict detection ML module which is introduced and evaluated through various methods. The results showed that most of the conflict detection module predictions are accurate. Type I error is persistent and present throughout different analyses methods. Furthermore, conflict detection module predictions are compared to human conflict detection and resolution. The conflict detection module predictions did not change their accuracy regardless of the criterion distance (separation minimum violation without ATCO action vs. predicted minimum distance of 12 NM or less).

It may be concluded that the proof-of-concept system defined in the ConOps of the AISA project can successfully handle selected en-route air traffic monitoring tasks with sufficient accuracy. Further progress in accuracy is needed for the future AI situation awareness system, especially with the machine learning modules. Moreover, ATCOs' reactions and feedback to artificial situation awareness inputs in experiment 2 showed that the design of the future human-machine interface needs to consider that today executive controllers need a lot of concentration for radio communication and are focused on a short to intermediate span of anticipation due to traffic complexity and dynamic. In contrast, machine situation awareness can include large anticipation spans, that may be used for optimisation. A large anticipation span is desirable if accuracy in artificial situation awareness inputs is provided reliably, and changes foreseen in the flight plan are considered by machine situation awareness.

The proof-of-concept system does not function in real-time – experiment 1 data and CD module outputs must be manually exported and converted to RDF format before the automated tasks are applied to the data. Analysis of processing times showed that the system is potentially capable of real-time operation, which is interesting for future development system purposes. The processing time of a single graph, representing a snapshot of a single traffic situation, is similar to the refresh rate of an ATCO's working position.

System components, operation and outputs form a relevant and accurate representation of the traffic situation being processed. To objectively assess the situation awareness attained by the system, an existing AI system situation awareness framework was applied. The specific architecture and capabilities of the AI situation awareness system rate highly on the scale presented in the framework.

Methodological limitations of the human-in-the-loop simulations were rooted in the lack of familiarity that ATCOs had with the simulation tool. This may have created additional workload, impaired the use of long-term memory content such as mental models and schemas for situation awareness, and, in

EUROPEAN PARTNERSHIP
Co-funded by the European Union

some cases, frustrated ATCOs. These effects altogether lowered the validity of the measured human situation awareness. Another limiting factor was the stage of implementation of the AI situation awareness system by the time of the second experiment. It would have been favourable to have *artificial situation awareness inputs* from a fully implemented artificial situation awareness system available to investigate human-machine team situation awareness and human performance. Nevertheless, using human-in-the-loop simulation was useful as it provided data for later simulation to quantitatively check the plausibility of the accuracy of the project-level implementation of the artificial situation awareness system's predictions and estimations.

Future research and development should further improve the accuracy of the artificial situation awareness system in accomplishing en-route air traffic monitoring tasks. Finding an optimal balance between training and test datasets and between sensitivity and the number of false alarms is key. From a safety point of view, it is favourable to ensure high sensitivity for threats of loss of safe separation at the costs of false alarms. From operational experience, false alarms will undermine trust in the artificial situation awareness system.

Keeping ATCOs aware of the situation, of impending threats and future trends is necessary for safe and efficient air traffic. Allowing active involvement to ensure readiness and skilfulness in reaction is probably the most challenging part for the design of future ATC. How to combine best machine and human situation awareness is currently investigated by the SESAR HORIZON 2020 projects MAHALO (Modern ATM via Human/Automation Learning Optimisation) and TAPAS (Towards and Automated and exPlainable ATM System) focusing on comprehensibility and acceptance of AI-based tools in ATC.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Acknowledgements

We thank EUROCONTROL Innovation Hub for the provision of the simulation program ESCAPE Light, and Philippe Bouchaudon as point of contact for adaptations and add-ons as well as support.

Our special thanks go to the 33 ATCOs from Skyguide for their interest and willingness to participate in the experiments and their tolerance to wear all the measurement equipment.

With her Master Thesis on “Acquisition of Situation Awareness and Performance Parameters from En-Route Air Traffic Controllers and Comparison with AI's Situation Awareness” Celina Vetter has greatly contributed to the methodology of eye-tracking analysis applied in the study and to the results for this report.

We also want to thank our two subject matter experts Jennifer Burkhalter and Christoph Herberth from Skyguide for their tremendous efforts, commitment, and patience in explaining. They were important enablers and actors in our simulation experiments.

List of Abbreviations

Abbreviation	Definition
A/C	Aircraft
ACC	Area Control Centre
ADS-B	Automatic Dependent Surveillance–Broadcast
AI SA system	Artificial intelligence situation awareness system developed and evaluated in AISA project
AI SA	Artificial Situation Awareness
AIXM	Aeronautical Information Exchange Model
AOI	Area of Interest
ATC	Air Traffic Control
ATCO	Air Traffic Control Officer
ATM	Air Traffic Management
BVP	Blood Volume Pressure
CD	Conflict Detection
CM	Conflict Manager
CVT	Computer Vision Tool
CWP	Controller Working Position
DST	Dynamic Scanning Tool
ECAM	Electronic Centralised Aircraft Monitoring
EICAS	Engine Indication and Crew Alerting System
ESCAPE	EUROCONTROL simulation capabilities and platform for experimentation
ET	Eye Tracking
FIXM	Flight Information Exchange Model
FL	Flight Level
HDG	Heading
HMI	Human-Machine Interface
HST	Horizontal Scanning Tool
ISA	Instantaneous Self-Assessment
KG	Knowledge Graph
MAHALO	Modern ATM via Human/Automation Learning Optimisation
ML	Machine Learning
MTCD	Medium-Term Conflict Detection
QoS	Quality of Service

EUROPEAN PARTNERSHIP
Co-funded by the European Union

RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
ROC	Rate of Climb
ROD	Rate of Descent
SA	Situational Awareness
SASHA_L	SASHA on-Line (Solutions for Human-Automation Partnerships in European ATM) probe technique
SASHA_Q	SASHA questionnaire (Solutions for Human-Automation Partnerships in European ATM) self-rating
SC	Skin Conductance
SD	Standard Deviation
SHACL	Shapes Constraint Language
SI	Situation of Interest
SME	Subject Matter Expert
SPARQL	SPARQL Protocol and RDF Query Language
STCA	Short Term Conflict Alert
STCD	Short-Term Conflict Detection
TAPAS	Towards and Automated and exPlainable ATM System
TRA	Temporary Reserved Area
TTFF	Time to First Fixation
UML	Unified Modelling Language
XFL	Exit Flight Level

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Table of Contents

Preface 4

Abstract 5

Executive Summary 7

Acknowledgements 10

List of Abbreviations 11

1 Introduction ***21***

1.1 Summary on AISA Project 21

1.2 Background of the AISA Project 21

1.2.1 AISA Status Reached 22

1.2.2 Motivation 29

1.2.3 Object of Investigation 30

1.2.4 Stages of Implementation of the AI SA System 30

1.3 Research Questions 31

1.3.1 Human Situation Awareness 31

1.3.2 Human Compared to Artificial Situation Awareness 31

1.3.3 Human-Machine Team Situation Awareness and Human Performance 32

1.3.4 Accuracy of Artificial Situation Awareness 32

1.3.5 Summary of Research Questions 33

2 Theory ***35***

2.1 Human Situation Awareness 36

2.1.1 Aspects of Situation Awareness Concept 37

2.1.2 Methods to Measure Situation Awareness 41

2.1.3 Attention and Gaze Behaviour 42

2.1.4 Effects of Workload and Stress on Situation Awareness 43

2.2 AI and Machine Situation Awareness 44

2.2.1 Automation and Monitoring 44

2.2.2 Machine Learning 45

2.2.3 Machine Situation Awareness 45

2.3 Human-Machine Team Situation Awareness 47

2.3.1 Distributed Human-Machine Team Situation Awareness 48

2.3.2 Comparison of Human and Machine Situation Awareness 48

2.3.3 Aspects of Human-Machine Collaboration for Situation Awareness 49

3 Methods ***50***

3.1 Experiments 50

3.2 Simulation Tool 52

3.3 Scenario Descriptions 53

3.3.1 Experiment 1 53

3.3.2 Experiment 2 54

3.4 Materials 55

EUROPEAN PARTNERSHIP

Co-funded by the European Union

3.5 Experimental Manipulation 56
3.6 Participants 59
3.7 Measurement of Artificial Situational Awareness 60
3.7.1 Knowledge Graph and Task Analysis 60
3.7.2 Analysis of Conflict Detection ML Module Predictions Regarding Situations of Interest 65
3.7.3 Analysis of Conflict Detection ML Module Predictions Regarding Conflicts 66
3.8 Measurement of ATCO Situation Awareness 66
3.8.1 Subjective Rating for Situation Awareness 66
3.8.2 Gaze-Based Analysis of Situation Awareness 66
3.8.3 Implicit Performance Measurement for Situation Awareness 68
3.8.4 Probe Technique for Situation Awareness 68
3.8.5 Scale Scores for ATCO Situation Awareness 69
3.9 Measurement of Performance and Control Strategies 70
3.9.1 Behavioural Analysis of the ATCO 70
3.9.2 Processing Behavioural Observation Data in R 71
3.10 Workload Measurement 72
3.10.1 Subjective Rating 72
3.10.2 Biometrical Analysis 73
3.10.3 Blood Volume Pulse 74
3.10.4 Skin Conductance 74
3.10.5 Transformation of Biometrical Data for Analysis 74
3.11 Statistical Analysis 75
4 Results 76
4.1 Results on Human Situational Awareness 76
4.1.1 Descriptive Results on Situation Awareness Measures 76
4.1.2 Correlational Results on Situation Awareness Measurement Methods 79
4.1.3 Group-Level Analysis for ATCOs with Preserved and Degraded SA 81
4.1.4 Performance and Workload 86
4.2 Results on Comparison of Human and Machine Situation Awareness 89
4.3 Evaluation of AI SA's Contribution to Human-Machine Team Situation Awareness and Performance 94
4.3.1 Evaluation of ATCO's Performance Based on Behavioural Coding 94
4.3.2 Evaluation of Artificial Situation Awareness Based on Questionnaire Answers 96
4.4 Results on the Accomplishment of Artificial Situational Awareness 100
4.4.1 Results of Knowledge Graph and Task Analysis 100
4.4.2 Results on Conflict Detection ML Module Predictions Analysis Regarding Situations of Interest 102
4.4.3 Results on Conflict Detection ML Module Predictions Analysis Regarding Conflicts 103
4.4.4 Results on Levels of Situational Awareness 105
4.4.5 Discussion on Results of Accomplishment of Artificial Situational Awareness 109
4.5 Results on AI SA System Performance 109
4.6 Robustness and Generalisability of the AI SA System 111
4.6.1 Independence of the Conflict Detection ML Module Predictions Regarding Situations of Interest 111
4.6.2 Independence of the Knowledge Graph and Task Analysis Accuracy from the Scenario 113

EUROPEAN PARTNERSHIP
Co-funded by the European Union

5 Discussion 116

5.1 Methodological Aspects and Limitations 118
5.1.1 Experimental Design 118
5.1.2 Simulation Software 119
5.1.3 Artificial Situation Awareness 119
5.1.4 Limitations Related to Participants 122
5.1.5 Behavioural Coding 123
5.1.6 Biometrical Analysis 123
5.1.7 Eye Tracking Analysis 124

5.2 Summary and Conclusion 126

5.3 Outlook 129

6 References 130

7 Glossary 139

Appendix A Extracts of the SHAPE Questionnaire 141

A.1 SASHA_Q 141

A.2 SASHA_L 141

Appendix B Description of Computer Vision Tool (CVT) 143

Appendix C Graphs 146

C.1 Karolinska Sleepiness Scale 146

C.2 Stress 147

C.3 Satisfaction 149

Appendix D Data Frame for Count – Experiment 1 150

Appendix E Data Frame for Conflict Comparison – Experiment 1 151

Appendix F Mean Reaction times data frame of experiment 1 152

Appendix G Data Frame for Checkbox – Experiment 1 153

Appendix H Data Frame to Compare Time – Experiment 1 154

Appendix I Data Frame for Number of Events per Call & Number of Conflict Solutions – Experiment 1 155

Appendix J Data Frame for Pearson Correlation 156

Appendix K Correlation of Conflict Detection Module Input Values Deviation with Prediction Error 157

Appendix L Experimental Plan of the AISA Project 158

L.1 Overall objectives 158
L.1.1 Knowledge graph and reasoning engine 160
L.1.2 Machine learning modules 166
L.1.3 Overall evaluation objectives 169

EUROPEAN PARTNERSHIP
Co-funded by the European Union

L.2 Experimental Approach and Research Questions.....171
L.3 Variables.....174

EUROPEAN PARTNERSHIP
Co-funded by the European Union

List of Tables

Table 1: Example of statistical data for plausibility check of conflict prediction 25

Table 2: List of AI SA tasks and their automation status 26

Table 3: Conditions for AI system awareness 46

Table 4: AI system awareness level classification 47

Table 5: Knowledge represented in the AI SA KG (from D2.1) 49

Table 6 Between-Group Comparison for effect of AI SA support 50

Table 7: ISA results for experiment 1 and 2 (N= 20; 16) 56

Table 8: Manipulation check for subjects' handling skills with ESCAPE Light rated by the SMEs during experiment 1 and 2 (N= 20; 16) 57

Table 9: Descriptive analysis for participants in experiment 1 59

Table 10: Descriptive analysis for participants in experiment 2 59

Table 11: Summary of objective situation awareness indicators and the conditions for comparison between the machine and human 63

Table 12: Definition of events in the scenarios of experiment 1 73

Table 13: Pearson correlations between different categories of E1S2 scenario in experiment 1 80

Table 14: Pearson correlations between different categories of E2S2.1 scenario in experiment 2 80

Table 15: Dwell time and frequency comparison of ATCOs with preserved and degraded situation awareness 83

Table 16: Pearson correlation of performance measures 87

Table 17: Comparison of query answers between ATCOs and AI SA system (N= 16 ATCOs) 90

Table 18: Comparison of conflict solution by ATCOs from experiment 1 (N=18) and 2 (N= 14) 95

Table 19: Evaluation of the AI SA inputs by the ATCOs regarding relevance (N=16= 96

Table 20: Type I Error and Type II Error results 102

Table 21: AI SA condition fulfilment estimate 105

Table 22: AI SA awareness level estimate 107

Table 23: Type I Error and Type II Error results (7.5NM) 112

Table 24: Type I Error and Type II Error results (12.5NM) 112

Table 25: Summary of Type I error and Type II error results 113

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Table 26 Overview: Research questions and summary of results 116

Table 27: Correlation of Deviation of Input Variable Values with Conflict Detection Module Prediction Error..... 157

Table 1 The overall system research questions, technical objectives, and results..... 159

Table 2 The KG and its subsystem research questions, objectives, and results 161

Table 3 The ML modules research questions, objectives, and results..... 166

Table 4 The SA research questions, objectives, and results..... 169

EUROPEAN PARTNERSHIP

Co-funded by the European Union

List of Figures

Figure 1: Conceptual diagram of the proof-of-concept machine situation awareness system 22

Figure 2: Conceptual diagram of the system including ATCOs and AI situation awareness system 48

Figure 3: Overview of experimental setup regarding the approximate timing of measurements 51

Figure 4: Comparison of subjective workload (ISA) and SME rating for ESCAPE Light handling skill “label handling” in experiment 1 and 2 (N= 20; 16) 57

Figure 5: Mental capacity absorbed to handle new system (ESCAPE Light) (N= 20; 16) 58

Figure 6: How comfortable did you feel with the ESCAPE Light simulation software? (N= 20; 16) 58

Figure 7: ET Gaze with Gaze History 67

Figure 8: Static AOI Creation of Screens 67

Figure 9: Comparison of human and system SA 69

Figure 10: Scoring key for SASHA_Q 69

Figure 11: ISA workload rating 73

Figure 12: Boxplots for mean scale scores for SASHA_Q for all interactive scenarios in experiment 1 (N=20) and 2 (N=16) 77

Figure 13: Most challenging events and conflicts from all scenarios from experiment 1 (N=18)* 78

Figure 14: Dwell time compared to conflict detection for E1S2 scenario (N= 9) 79

Figure 15: Gaze duration and gaze count for aircraft per scenario of ATCO with preserved (green circles) and degraded situation awareness (red circles). Green aircraft labels indicate aircraft participating in a conflict (symbol marking on aircraft label for respective conflict); yellow aircraft needed a flight level change and black aircraft did not require much attention. 84

Figure 16: Cumulated distribution of relative changes of the median heart rate 88

Figure 17: Cumulated distribution of relative changes of the median skin conductance 89

Figure 18: How much do you trust 1) the automation implemented in Skyvisu and 2) AI SA inputs at work in future? 98

Figure 19: Did AI SA inputs support your situation awareness overall? (N= 16) 98

Figure 20: Did AI SA inputs support your decision making? (N= 16) 99

Figure 21: Objective count of occurrences of preserved SA 101

Figure 22: Objective count of occurrences of degraded SA 101

Figure 23: The overall number of occurrences of degraded human SA 102

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Figure 24: Distribution of the initial and final prediction analysis results 103
Figure 25: Comparison of conflict detection ML and human performance 104
Figure 26: Runtime per graph vs Number of graphs in scenario 110
Figure 27: Median of runtime per graph for number of aircraft 110
Figure 28: Median of runtime per graph for scenarios 111
Figure 29: Distribution of the initial and final prediction analysis results (7.5NM) 112
Figure 30: Distribution of the initial and final prediction analysis results (12.5NM) 113
Figure 31: Objective count of occurrences of preserved and degraded situation awareness (E1S1) 114
Figure 32: Objective count of occurrences of preserved and degraded situation awareness (E1S2) 114
Figure 33: Objective count of occurrences of preserved and degraded situation awareness (E1S3) 115
Figure 34: Objective count of occurrences of preserved and degraded situation awareness (E1S4) 115
Figure 35: Accuracy of Tobii Glasses 3 125
Figure 36: Steps for the screen detection of the Computer Vision Tool 143
Figure 37: Information on ET-Recording in CVT 144
Figure 38: Information on Screen recording in CVT 145
Figure 39: Sleepiness before and after experiment 1 146
Figure 40: Sleepiness before and after experiment 2 147
Figure 41: Stress level before and after experiment 1 148
Figure 42: Stress level before and after experiment 2 148
Figure 43: Satisfaction of ATCOs for each scenario of experiment 1 149
Figure 44: Satisfaction of ATCOs for each scenario of experiment 2 149
Figure 1 Conceptual diagram of the proof-of-concept machine situation awareness system 160
Figure 2 Experimental design to evaluate the effect of AI SA support on human performance 174

EUROPEAN PARTNERSHIP
Co-funded by the European Union

1 Introduction

The use of AI-tools are beneficial to air traffic controllers (ATCOs) if they were capable of making accurate predictions, offering recommendations for optimal service, and monitoring air traffic and compliance. The human element has been identified as a relevant factor limiting increases in air traffic management (ATM). Human performance affects overall system safety and effectiveness in air traffic control (ATC)

This chapter provides information on the AISA project and its background. It explains the motivation and research questions used for evaluation of the AI SA system.

1.1 Summary on AISA Project

The SESAR 2020 project AISA – AI Situational Awareness Foundation for Advancing Automation – is an exploratory research project associated with digitalisation and automation principles for air traffic management (ATM). It investigates the use of artificial intelligence to generate situation awareness for en-route air traffic monitoring tasks and its contribution to distributed human-artificial team situation awareness. The AISA project has developed an AI situation awareness system (AI SA system) with a reasoning engine that is based on domain-specific knowledge graphs, and which interacts with machine-learning modules for conflict detection, 3D trajectory prediction and traffic complexity estimation. What AISA requires to have situation awareness was previously operationalized in 57 en-route air traffic monitoring tasks that were identified in the concept of operations for the project. From these, 46 tasks were selected for implementation within the scope of the project.

A proof-of-concept system was built and tested in experiments using low-fidelity human-in-the-loop simulations. Inputs from AI were presented to ATCOs – mimicking real-time human-machine interaction – in an interactive as well as “watch only”-type of simulation to test the accuracy of machine situation awareness and to investigate its comparability and compatibility with human situation awareness.

1.2 Background of the AISA Project

AI is considered a promising solution to augment the capacity and safety in ATC and an important prerequisite to implement integrated solutions rather than having several tools for single purposes in place (e.g., for conflict detection and avoidance, for conformance management, etc.). However, it is important that ATCOs can understand and share situation awareness with AI SA system.

The concept of operations states three different descriptions of the system in regard to time horizon (project level, vision 2035, vision 2050). The goal of the AISA project is to integrate different data sources and to reach artificial situation awareness regarding en-route air traffic monitoring tasks that is comparable to ATCO situation awareness.

The innovative approach combines a reasoning engine with machine learning (ML) modules to ensure comprehensibility of the AI SA output on artificial situation awareness for ATCOs and to check the plausibility of ML modules’ estimates and predictions.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Throughout the course of developmental stages, the integration of different sources of data are automatised. It must be pointed out that the evaluation study comparing ATCO situation awareness with artificial situation awareness used an early stage of AI SA system implementation (Section 1.2.4). Nevertheless, this effort is considered favourable in terms of fast failing and learning in the design and development process.

1.2.1 AISA Status Reached

As stated in the previous chapter, the AISA concept of operations contains different descriptions of the AI SA system regarding the time horizon – vision 2035, vision 2050 and the project level description. While the two "vision" system descriptions offer the "big picture", the goal of the third was to describe the proof-of-concept knowledge-based system which was to be developed in this project. The purpose of that proof-of-concept system was not to reach the technology readiness level of a real-time system, but to explore the feasibility of such a system by developing and testing its most vital components. A conceptual diagram of the proof-of-concept system is shown in Figure 1.

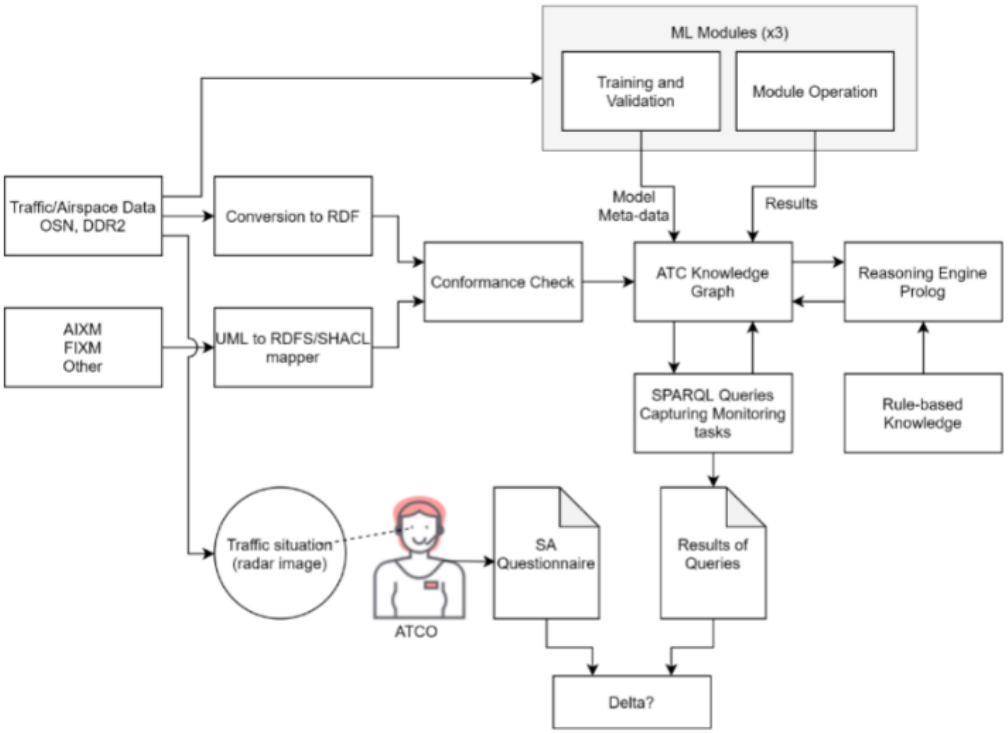


Figure 1: Conceptual diagram of the proof-of-concept machine situation awareness system

Several notable points were added to the description of the proof-of-concept system – instance data would not be automatically generated, that the scale of the system would be (geographically and temporally) limited, that user interface would be limited, and that the system would not be real-time. Those points are important for the overall system and informed the development of sub-systems. The following sub-chapters are intended to show which parts of the initial proof-of-concept system were successfully implemented and to explain discrepancies between it and the final version.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

1.2.1.1 Knowledge Graph

Populated by knowledge from diverse sources such as the ESCAPE Light simulator and the developed machine learning modules, the knowledge graph (KG) serves as the central part of the system. To add the knowledge to the graph, an underlying architecture had to be constructed. As envisioned, aeronautical exchange models (AIXM, FIXM) were used because they already contain semantic information on air traffic and aeronautical concepts. Since these concepts are described in Unified Modelling Language (UML), a mapper was developed to transfer them to the Resource Description Framework Schema (RDFS) and Shapes Constraint Language (SHACL).

The "UML to RDFS/SHACL" mapper takes a chosen subset of an exchange model and produces equivalent RDFS vocabulary/SHACL constraints. The vocabulary of a graph describes which types of concepts may be used in a KG, while SHACL constraints describe structure (properties, type, and range of values etc.). Not all concepts necessary for the AI SA KG system are described in AIXM and FIXM– additional vocabulary and constraints were added directly to RDFS and SHACL, as the need arose. These changes were planned for the concept of operations in D2.1 (AISA Consortium, 2020a).

The creation of graph vocabulary and structure was a prerequisite for the addition of specific traffic situation data. Even though the initial plan was to create these data instances manually, the scope of work required certain parts of the process to be automated. This offers an additional advantage as it brings the proof-of-concept system closer to the real-time system, which will need to employ automation to achieve real-time operation. The most important example of automation was the translation of data exported from the ESCAPE Light simulator to RDF files.

Instance data was created for traffic situations inside a single en-route sector within the Swiss airspace, LSAZM567, ranging from FL355 to FL999. A Swiss sector was chosen because ATCOs from Skyguide were tasked with system evaluation. Traffic information is based on AIRAC 1907, from which a single day (namely 4 July 2019) was chosen for evaluation purposes – the traffic on that day was deemed to represent pre-COVID traffic well. The chosen date was not used to train the conflict detection machine learning module.

The data was divided into 2 groups – static and dynamic. Static data includes all knowledge which does not change with time or changes rarely, such as sector border coordinates, ML module training statistics, etc. These were all collected to a graph named "default" and added to each exercise.

All ATCOs were given identical ESCAPE Light traffic scenarios to solve. Exported data was later converted to an RDF graph every time the ATCO issued a clearance. If there were no clearances, the pause between timestamps/graphs was a maximum of 30 seconds. Dynamic data for each timestamp was stored to a graph named "g(number)", where the initial timestamp's RDF graph being named g0. This encompasses data such as flight data, airport data, machine learning module outputs, and other data.

1.2.1.2 Rule-based Knowledge

Serving as a counterpart to the factual knowledge shown in the Section (1.2.1.1), rule-based knowledge is added to the system to be executed on top of the factual knowledge (i.e. data) represented in the KG. It represents the rules of ATC which the ATCOs apply to traffic situations and events to gain situational awareness and control traffic. The implementation of rule-based knowledge was initially planned for SWI-Prolog (a free implementation of the Prolog programming language), but most of the AI SA tasks were found to be simple enough for the implementation to be done in the Java

EUROPEAN PARTNERSHIP
Co-funded by the European Union

programming language. A single task was implemented in Prolog to compare the results and confirm they are equivalent, but a decision was made to complete the already under-way implementation in Java to ensure system homogeneity.

No matter the implementation, rule-based knowledge requires the facts stored in the KG. Those facts were accessed by using SPARQL Protocol and RDF Query Language (SPARQL) in Java classes. A more detailed description of AI SA task implementation in Java will be given in the section Automated Tasks (1.2.1.4). What is important to note is that each task (and its underlying rules) was developed through the work of subject matter experts (SMEs) to ensure that conclusions/outputs were reached in a proper way and that they conveyed appropriate information to the ATCO.

1.2.1.3 Machine Learning Modules

While some data sources – aeronautical information publications, Base of Aircraft Data (BADA) – offer information that does not need to be checked for correctness by the user, information pertaining to future states (such as weather forecasts or position predictions) is not as trustworthy. ATCOs may use high integrity information to confirm the plausibility of traditional system tool predictions, so the same approach was adopted to confirm the results of machine learning (ML) modules. The nature of ML modules produced another plausibility confirmation method – since the modules were trained on known datasets, scenario data that was used could be checked against those datasets so the system may flag outlying data.

In all, three machine modules were developed:

[1] Trajectory prediction module

[2] Conflict detection (CD) module

[3] Complexity assessment module

4D trajectory prediction module uses a two-step process. A neural network is trained for static aircraft track prediction without the time domain so it can predict a granular static track of the flights. The second step is the combination of the actual aircraft state and the predicted track (aircraft-fixed pattern) in order to determine a concrete future position in 100 NM. The databases OpenSky Network for ADS-B (Automatic Dependent Surveillance–Broadcast) data as well as the DDR2 EUROCONTROL database for flight which is described in D3.1 (AISA Consortium, 2021a) were used.

Conflict detection module revolves around the concept of Situation of Interest (SI). An SI would be when an aircraft pair is expected to intersect while having a horizontal separation lower than a pre-defined separation (10NM) and breach the vertical separation minima (1000 ft). Two approaches have been developed: the Static mode predicts SI and their safety metrics once an aircraft enters the airspace and the Dynamic mode makes the predictions throughout the aircraft's evolution within the airspace. Safety metrics are represented by the Minimum Distance, distance to reach the Minimum Distance and the time to reach the Minimum Distance for each aircraft pair as described in D3.2 (AISA Consortium, 2021d).

Air traffic complexity estimation module uses a novel solution that utilises ATCO tasks which are defined depending on the traffic situation. The model uses air traffic situation characteristics in order to determine ATCO tasks for each aircraft, which can then be expressed as a unique traffic situation signature and graded according to the methodology developed in a previously conducted study. Coefficients of the linear regression model are used to determine the contribution of a task to the

EUROPEAN PARTNERSHIP
Co-funded by the European Union

overall complexity. This model is also able to identify the most complex aircraft, the one that causes the highest complexity rise, and check how changes in its parameters affect the overall complexity described in D3.3 (AISA Consortium, 2021b).

As shown in Section 1.2.1.4, the conflict detection module is the most frequently used of the three ML modules. The module itself is not integrated into the rest of the KG system yet, but this did not present a problem since both input and output conversions of the model were successfully automatised. Necessary data was first converted from its raw form to the form used by the module, predictions were generated and converted into prepared Resource Description Framework (RDF) nodes and included in appropriate KG.

Statistical data representing the training dataset of the module was added to the static graph of each scenario, so it could be used to check the plausibility of conflict predictions. SPARQL queries access both the statistical data on each aircraft in a conflict prediction pair and the current value for each parameter to be checked – altitude, speed, track, latitude, longitude, and vertical rate. The difference between the current value and training data mean value is then expressed using standard deviation ranges seen in Table 1. To analyse how much the current state values diverge from the mean value of training data, four categories were introduced: 1σ, 2σ, 3σ, and the "over 3σ" category. These categories correspond to the sigma band where the current state values are. Therefore, for every aircraft pair there are 8 different sigma band categories (altitude, speed, heading and vertical rate for each aircraft). The Table 1 below shows the statistics for these categories from the conflict detection module training data. In this example data about Airbus A320 is shown. Except minimum and maximum value for each category, the table also shows distribution of values in a way as to present what is the limit value in which there is 25-50-75% of the whole data values. Multiple correlation analysis used to check for correlation between the ML module sigma band category and the predicted minimum distance accuracy showed that these variables are not statistically related.

Table 1: Example of statistical data for plausibility check of conflict prediction

A320	altitude	geo altitude	ground speed	latitude	longitude	track	vertical rate
count	1832175	1832175	1832175	1832175	1832175	1832175	1832175
mean	36732.36	38087.62	440.4109	47.11373	8.605677	210.2047	12.72612
sd	866.6321	1658.498	39.17293	0.45306	1.015303	100.618	211.4941
min	35500	875	69	46.10445	6.95448	0	-3136
25%	36000	37400	412	46.79892	7.740672	130.7967	0
50%	36725	37950	437	47.16289	8.530602	236.8071	0
75%	37025	38775	469	47.47416	9.426407	293.763	0
max	44950	128000	576	47.88332	10.48886	359.8943	3136

Part of the conflict detection process is monitored by checking the conflict module input data statistics. It is also necessary to check and follow the outputs of the conflict module. This system function was divided into two separate tasks, one to grade conflict predictions when they appear and another to

EUROPEAN PARTNERSHIP
Co-funded by the European Union

follow the conflicts as the scenario develops. Since the separation was not planned in the initial task list formation, the final task list contains a total of 58 tasks (described in 1.2.1.4 and D4.4).

The tasks which grade conflict module predictions functions by performing a "sanity check", meaning it calculates previous and current distance between the aircraft in the pair to determine if they are diverging. If the task concludes the aircraft are diverging, no additional checks are performed because the prediction is incorrect. If the task determines the aircraft are converging and the predicted conflict is indeed possible, then it checks if the predicted distance to conflict point (made for the first aircraft in the pair) is similar to the distance that aircraft would cross if it continued flying at its current speed. This check might ascertain that the prediction is unrealistic if it requires the aircraft to fly much faster than it currently is to reach the conflict point.

The task meant to continuously check conflict predictions was developed, but a technical issue concerning SPARQL queries prevented its completion and testing. It was meant to store all conflict predictions to a single output graph and then access that graph to check i) what is the actual distance between the aircraft, ii) if the ATCO has made any changes to either aircraft trajectory. The actual distances would be updated until the aircraft start diverging and then compared to the predicted minimal distance. The task would also consider ATCO actions which could affect the accuracy of the original prediction. The difficulty with performing the task was not in the task Java class, which was completed, but in the inability of SPARQL queries to update the value of current distance in each timestamp. Because of time constraints, this problem was not solved, and the task was left uncompleted. Checking the predictions of the conflict prediction module is thus performed by the two other tasks.

1.2.1.4 Automated Tasks

The AISA ConOps D2.1 (AISA Consortium, 2020a) offers a list of 57 tasks to be automated, divided into 11 categories. Table 2 shows that list and includes an additional task and a column to designate the status of implementation for each task.

Table 2: List of AI SA tasks and their automation status

Task category	Task	Status
1. CONFORMANCE MANAGEMENT	1.1. Check that aircraft is climbing/descending towards cleared FL	Completed
	1.2. Check that aircraft is at cleared FL	Completed
	1.3. Check that aircraft is maintaining FL	Completed
	1.4. Check that aircraft is turning towards/opposite of cleared heading	Completed
	1.5. Check that aircraft is at cleared heading	Completed
	1.6. Check that aircraft is maintaining current heading (different than cleared heading)	Completed
	1.7. Check that aircraft is accelerating/decelerating towards cleared speed	Completed
	1.8. Check that aircraft is flying at cleared speed	Completed

EUROPEAN PARTNERSHIP
Co-funded by the European Union

	1.9. Check that aircraft is maintaining current speed (different than cleared speed)	Completed
	1.10. Check that aircraft is flying towards cleared point	Completed
	1.11. Check that aircraft is at cleared point	Completed
	1.12. Check that aircraft current ROC/ROD is lower/higher than cleared	Completed
	1.13. Check that aircraft is maintaining cleared ROC/ROD	Completed
	1.14. Check that aircraft is increasing/decreasing towards cleared ROC/ROD	Completed
	1.15. Check that aircraft is following the 3D trajectory	Completed
	1.16. Check is the deviation from 3D trajectory is within tolerance	Completed
	1.17. Check that aircraft is following the 4D trajectory	Not completed
	1.18. Check is the deviation from 4D trajectory is within tolerance	Not completed
2. DETECT INCOMING PLANNED FLIGHTS	2.1. Check that aircraft is close to sector boundary	Completed
	2.2. Check that aircraft is approaching sector boundary	Completed
	2.3. Check that aircraft altitude is within the altitude band of the sector	Completed
	2.4. Check that aircraft altitude is approaching the sector altitude	Completed
3. ASSUME, IDENTIFY, AND CONFIRM FLIGHT	3.1. Check that aircraft is incoming	Completed
	3.2. Check that aircraft is planned	Completed
	3.3. Check that aircraft has sent the initial call (via datalink)	Completed
	3.4. Confirm that aircraft can be assumed	Not completed
4. ASSESS IF EXIT CONDITIONS ARE MET	4.1. Check that aircraft is flying towards the exit point	Completed
	4.2. Check that aircraft will reach the exit point on the required FL	Completed
	4.3. Check that aircraft will reach the exit point at the expected time	Not completed
5. CONFLICT MANAGEMENT	5.1. Check all aircraft pairs for conflict (ML module)	Completed
	5.2. Check plausibility of the predicted conflicts	Completed
	5.3. Check which conflicts are to occur within the sector	Completed
	5.4. Rank conflicts based on urgency	Completed
	5.5. Check conflict module inputs against training data	Completed

EUROPEAN PARTNERSHIP
Co-funded by the European Union

6. EXECUTE AIRCRAFT'S PLAN	6.1. Detect aircraft that have to climb/descend to requested FL	Completed
	6.2. Detect aircraft that have to climb/descend to exit FL	Completed
	6.3. Detect aircraft that will reach top of descent within the sector (ML module)	Not completed
	6.4. Detect if planned trajectory passes through restricted airspace	Completed
7. TRANSFER AIRCRAFT	7.1. Check which aircraft need to be transferred	Completed
	7.2. Check if change of frequency is issued to aircraft (via datalink)	Completed
	7.3. Change aircraft status to transferred	Completed
8. MAXIMISE QUALITY OF SERVICE	8.1. Detect direct-to candidates	Completed
	8.2. Determine military airspace availability	Completed
	8.3. Check suggestion for shortened RBT	Not completed
9. WORKLOAD MONITORING	9.1. Track current number of assumed aircraft	Completed
	9.2. Track number of conflicts and potential conflicts	Completed
	9.3. Determine future number of sector entries	Completed
	9.4. Determine sector air traffic complexity (ML module)	Not completed
	9.5. Determine plausibility of traffic complexity assessment	Not completed
10. IDENTIFY MISSING INFORMATION	10.1. Identify aircraft with possible equipment degradation	Completed
	10.2. Check situation at destination airport	Completed
	10.3. Check situation at alternate airports	Completed
	10.4. Monitor adverse weather areas	Completed
	10.5. Monitor restricted airspace	Completed
	10.6. Infer missing information	Not completed
11. MONITOR STATUS OF ATC SUB-SYSTEMS	11.1. Monitor performance of ATC conflict detection module	Not completed
	11.2. Monitor performance of complexity assessment module	Not completed
	11.3. Monitor performance of trajectory prediction module	Not completed

As mentioned in the machine learning module section, the separation of conflict prediction assessment functions into two tasks resulted in an additional task – task 5.5. With that in mind, 46 out of 57 total tasks (~81%) were completed and tested. The small subset of tasks which were not completed is comprised mostly of machine learning module-related tasks. The rest are miscellaneous tasks which

EUROPEAN PARTNERSHIP
Co-funded by the European Union

had proven too complex to define and code, given the time constraints of the project, and should be revisited later.

Each task consists of SPARQL queries which access the flight data stored in the KG and functions which use rule-based knowledge to connect the current traffic situation parameters with appropriate outputs. More information on task creation and SPARQL queries can be found in D4.4 (AISA Consortium, 2021e). Since tasks range from simple to complex (e.g., checking if the aircraft is at cleared flight level vs checking if aircraft trajectories pass through military airspace while its active), number of SPARQL queries and task functions may differ. Auxiliary functions were collected in a single Java class and added to the KG system class hierarchy, which enabled their use in multiple tasks without repeating the code multiple times.

1.2.1.5 AISA System Operation

The AI SA KG system works by first erasing the contents of the KG (to avoid mixing with previously used data), initialising tasks, loading the RDF graph with the static data to the KG, and then looping through the following steps:

1. loading a single dynamic data graph to the KG,
2. running all selected tasks and storing outputs in graph form to the KG,
3. (optional) retrieving, filtering, and printing outputs.

The dynamic data graphs are loaded in ascending order, starting with g0. As already described, this represents the evolution through the timeline of the exercise. Since some tasks require data or results from the previous timestep, they are not performed for g0 or in instances when a flight first appears (if it is not present in g0). No additional coding is required for this – SPARQL queries simply return nothing if some part of the query cannot be completed, and thus no outputs are generated by the task functions.

An additional piece of code was added at the beginning and the end of the loop to store the beginning/end times – their difference shows how long each loop (the processing of 1 graph) took. The same was later done to calculate the full runtime for each exercise, which also enables the calculation of runtime per graph.

1.2.2 Motivation

The main reason AISA was founded was to determine whether it is possible to combine human and machine situational awareness in a way that they complement each other and create a distributed situation awareness. In a world of increasing automation, the contribution of AI SA would be to automate certain monitoring tasks in en-route operations. It is important that this narrow and specific scope, which requires major reliability, can depend on transparency and the generalisation of the system used.

Unlike machine learning systems (e.g., deep neural networks) that basically function as a black box, a reasoning engine is capable of explaining the results it provides and how it obtained them. Another novelty that this system brings into ATC is the possibility to check results obtained from ML systems for inconsistencies and improbable results. This correlates closely to the way a human determines whether something is faulty or not.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

An artificially intelligent system with situation awareness has the purpose of increasing safety by essentially bringing an additional safety net into the equation. It will also serve as another team member that is constantly and consistently observing the situation and can be relied upon to improve team situational awareness. The system performs all those tedious monitoring tasks that take away the ATCO's time and attention capacity, and it does so with high reliability. It is aware of the situation at hand, in its own way, and its state which allows it to be a part of the team and team situation awareness. One of the main benefits that arises with implementation of such a system is its interoperability which is a by-product of KG usage for the purpose of data management. AI SA also has great potential to increase sector capacity since the automation of some monitoring tasks enables the introduction of other automation systems which will alleviate ATCO's workload as well.

1.2.3 Object of Investigation

The goal of the studies conducted and presented in this report is to evaluate the capability of the AI SA system to gain situation awareness for en-route air traffic monitoring tasks and to explore its contribution to shared situation awareness and human performance. Two human-in-the-loop experiments were conducted with the proof-of-concept AI SA system that analyse how ATCOs build situation awareness and investigate the comparability and compatibility of artificial situation awareness to human situation awareness. The experiments evaluated the adequacy of the AI SA system's concept to accomplish monitoring tasks and investigated how – from the point of viewparticipating ATCOs –it can be improved. They also laid the basis for further simulations that quantify the accuracy of the AI SA system's predictions and estimations.

1.2.4 Stages of Implementation of the AI SA System

Note that two stages of the AI SA system implementation were available at different times of the evaluation. They are called stage I and II and differ in the number of tasks implemented for monitoring en-route air traffic.

- Stage I implementation was in use at the time of experiment 2 in January 2022 and had 8 tasks implemented. AI SA system outputs had to be filtered by hand to be provided as AI SA inputs to ATCOs in experiment 2.
- Stage II implementation was accomplished in April 2022, where the AI SA system reached the project-level scope of implementation comprising 46 tasks for en-route air traffic monitoring. Stage II implementation of AI SA system is tested in simulations based on the experimental data from human-in-the-loop simulations to quantify the accuracy of estimations and predictions of the AI SA system.

Results on the comparison of query outputs by AI SA and answers from the ATCOs to the same queries as well as ATCOs' judgements regarding the usefulness of AI SA inputs are related to an early stage of implementation of the AI SA system.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

1.3 Research Questions

Research questions and later the results are grouped according to topical sections. A chronological order is inherent to this structure in respect to the stages of implementation of the AI SA system (see 1.2.4) that had been investigated.

1.3.1 Human Situation Awareness

A first group of question is dedicated to characterising ATCO situation awareness and address methodological aspects of measuring human situation awareness.

Question 1.1: What characterises ATCOs' scanning patterns and priorities?

For this purpose, scanning characteristics of ATCOs with "preserved" situation awareness were compared to ATCOs with "degraded" situation awareness by means of gaze analysis using eye-tracking data concerning aspects for situation awareness.

Question 1.2: Are different measures for situation awareness (self-ratings, queries, gaze-based analysis, and implicit measurements) significantly interrelated according to their meaning?

A moderate intercorrelation is expected among the different measures for ATCO situation awareness from literature. For this purpose, a score was assigned to each method (overall SASHA_Q score across all subscales; overall score for SASHA_L for different queries; score based on speed, accuracy and time aspects for gaze analysis and implicit performance measures (e.g., dwell time, time of conflict detection, conflict solutions) (see 3.8.5). For these scores intercorrelations were calculated with Pearson's r correlation.

1.3.2 Human Compared to Artificial Situation Awareness

Human and artificial situation awareness were compared regarding correctness and comprehensiveness. This was done with an early stage of implementation of the AI situation awareness system (stage I) (see 1.2.4).

Question 2.1: Are artificial and ATCO situation awareness comparable?

This question was investigated by comparing ATCO and AI SA system answers to identical queries about specific aspects of the situation across different scenarios of experiment 2.

Question 2.2: Can the AI SA system provide inputs to situation awareness that ATCOs were not aware of?

Based on the comparison of ATCO answers and AI SA system outputs it is investigated if machine situation awareness can provide information to ATCOs that they are not aware of in terms of explicitly named in their answers to the queries.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

1.3.3 Human-Machine Team Situation Awareness and Human Performance

Is human-machine team situation awareness beneficial for human performance? How do ATCOs react to AI SA inputs for shared situation awareness. This is investigated with a comparison of objective performance across two work conditions ("with" and "without AI SA input") and by means of ATCOs' subjective judgements of AI SA inputs. As with research question on Human Compared to Artificial Situation Awareness (1.3.2) investigations were done with an early stage of implementation of the AI situation awareness system (stage I).

Question 3.1: Is human performance enhanced by adding machine situation awareness?

To address this question implicit performance measure for ATCO situation awareness was used (time to detect conflict) was compared across the conditions "without AI SA inputs" (experiment 1) and "with AI SA inputs" (experiment 2).

Question 3.2: Do ATCOs evaluate artificial situation awareness inputs as useful and trustworthy contribution to human-machine team situation awareness?

Question 3.3: Do ATCOs use artificial situation awareness inputs for their situation awareness and decision making?

To answer these question ATCOS' judgements for each AI SA input provided at the end of a scenario were analysed for each AI SA input in all scenarios of experiment 2.

1.3.4 Accuracy of Artificial Situation Awareness

Is the AI situation awareness system based on a KG system and associated machine learning modules capable of accomplishing the assigned tasks for en-route air traffic monitoring in an accurate, comprehensible (transparent) and generalisable manner? For the analyses of the questions concerning accuracy of artificial situation awareness the project level of implementation for AI situation awareness system (stage II) is used (see 1.2.4).

Question 4.1: Can the monitoring tasks be applied to the KG to achieve situational awareness?

To answer this question, a list of objective requirements for situational awareness was made. The outputs of the tasks were analysed to check if all requirements were fulfilled, and no faulty conclusions were made. To check for generalisability, the accuracy of the results was compared across different scenarios.

Question 4.2: Does the CD machine learning module provide accurate results regarding situations of interest?

To answer this question, the initial and final predicted minimum distance and time to minimum distance were compared to the actual measured values. The initial predictions were observed before any clearances by the ATCOs were issued. For example, the first time the conflict detection ML module makes a prediction about two aircraft which did not receive any ATCO input by that moment is an initial prediction. After all ATCO clearances were made, and no further changes to the trajectories were done, the final predictions were observed.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Question 4.3: Does the CD machine learning module provide accurate results regarding conflicts?

This question extends the previous one. To compare how accurate is the CD ML module regarding only traffic that would violate separation minima, ML module predictions are compared with the ATCO recognition and resolution of conflict. It was also observed if the ML module predicted distances correspond to the distances after ATCO issued resolution actions e.g., does the CD module predicted distance increase after ATCO have issued clearance which resolves the conflict?

Question 4.4: Does the AI SA system check the status of its sub-systems?

Tasks were developed to assure the AI SA system can detect how the actual traffic data compares to the training data of the CD ML module and the validity of that module's predictions. This enables the system to have a degree of self-awareness.

1.3.5 Summary of Research Questions

Category	Research Question	Section
Human Situation Awareness	Q.1.1. What characterises ATCOs' scanning patterns and priorities?	4.1.3.2 Comparison of ATCO Groups for Gaze-Based Analysis of
	Q.1.2. Are different measures for situation awareness (self-ratings, queries, gaze-based analysis, and implicit measurements) significantly interrelated according to their meaning?	4.1.2 Correlational Results on Situation Awareness Measurement Methods
Human Compared to Artificial Situation Awareness	Q.2.1. Are artificial and ATCO situation awareness comparable?	4.2 Results on Comparison of Human and Machine Situation Awareness
	Q.2.2. Can the AI SA system provide inputs to situation awareness that ATCOs were not aware of?	4.2 Results on Comparison of Human and Machine Situation Awareness
Human-Machine Team Situation Awareness and Human Performance	Q.3.1. Is human performance enhanced by adding machine situation awareness?	4.3.1 Evaluation of ATCO's Performance Based on Behavioural Coding
	Q.3.2. Do ATCOs evaluate AI SA inputs as useful and trustworthy contribution to human-machine team situation awareness?	4.3.2 Evaluation of Artificial Situation Awareness Based on Questionnaire Answers
	Q.3.3 Do ATCOs use AI SA inputs for their situation awareness and decision making?	4.3.2 Evaluation of Artificial Situation Awareness Based on Questionnaire Answers

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Accuracy of Artificial Situation Awareness	Q.4.1. Can the monitoring tasks be applied to the KG to achieve situational awareness?	4.4.1 Results of Knowledge Graph and Task Analysis
	Q.4.2. Does the CD machine learning module provide accurate results regarding situations of interest?	4.4.2 Results on Conflict Detection ML Module Predictions Analysis Regarding Situations of Interest
	Q.4.3. Does the CD machine learning module provide accurate results regarding conflicts?	4.4.3 Results on Conflict Detection ML Module Predictions Analysis Regarding Conflicts
	Q.4.4. Does the AI SA system check the status of its sub-systems?	4.5 Results on AI SA System Performance

EUROPEAN PARTNERSHIP
Co-funded by the European Union

2 Theory

This chapter provides the theoretical background for human and machine situation awareness and outlines a conception for distributed human-machine team situation awareness.

The term situation awareness is often used in ergonomic research, operational and training practice. It describes how people, but also entire socio-technical systems, become and remain coupled to the dynamics of their environment (Stanton et al., 2017). Human situation awareness corresponds to an internalised mental model of the current state of environment that is used to base decisions on, take necessary actions, and/or adapt plans, and from which projections to the future are made. The selection and amount of information perceived and processed to build and maintain situation awareness depend on the salience of external cues, the relevance of information for task accomplishment and the availability of prior knowledge in long-term memory. Humans have to recognise or remember what is important for the task at hand. They have to know what, where and when to look for information. At a higher level of skill progress, people know subtle signals that are meaningful for successful regulation of action control. Workload obviously is an important factor for human situation awareness, as multiple competing tasks limit cognitive resources available for updating the mental image of the situation und for anticipating future demands to adapt to.

In comparison to this machine situation awareness is generated in a more comprehensive manner of information processing: All available information gets considered in processing and all queries for relevant issues (e.g. non-conformance, conflict detection and prediction etc.) are verified. The capability of artificial situation awareness to accomplish monitoring tasks depends on the implemented tasks to be accomplished in processing, and on the available information. Products of data processing can get stored and used for later purpose. The correctness of estimations and predictions can be ensured by plausibility checks on the input data and outputs.

A loss of situation awareness may result in human error – for instance, if an ATCO does not notice a pilot's non-conformance with a speed limitation. Consequently, this may lead to a steady reduction of separation and a lack of necessary countermeasures to ensure safe separation distance. In the cockpit, checklists support monitoring and control actions in all phases of flight. They support pilots' adequate situation awareness and completion of important action steps. This approach, however, is not applicable to ATC to the same extent. Guidelines for dealing with unusual and emergency situations in ATC exist. But the dynamic nature of air traffic prevents a high level of standardisation in ATC that goes beyond the use of standard phraseology in radio communication. Therefore, scanning techniques represent an important element to ensure adequate situation awareness.

How do ATCOs make sure they do not miss important information? Attention may get absorbed by tasks or may be focused exclusively on one problem, while diminishing the capability to overview the whole situation. To support ATCOs in their situation awareness machine situation awareness might be a useful complement and beneficial for safety and efficiency in future ATC and it might allow for more automation to keep ATCOs' workload at a manageable level.

To achieve distributed human-machine team situation awareness presupposes that machine and human situation awareness are compatible and complement each other. How does the output of machine situation awareness need to be designed and timed that it effectively supports human situation awareness without creating unjustifiable additional workload to ATCOs? Although this aspect

EUROPEAN PARTNERSHIP
Co-funded by the European Union

is out of the scope of the exploratory research of an AI-based machine situation awareness in the AISA project, this aspect is key to the effectiveness of any human-machine team situation awareness.

The combination of human and machine situation awareness provides redundancy to monitoring tasks. To achieve a diverse redundancy capable for effective countermeasure against loss of situation awareness due to lacking or flawed information would further require that human and machine situation awareness were fed by different, independent sources of information.

2.1 Human Situation Awareness

Situation awareness is defined as “the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future” (Endsley, 1988, p. 97). It is composed of the levels perception, interpretation of the perceived information, and anticipation how the situation will change in future (Endsley, 1995b). To avoid confusion with the levels of awareness an AI-based system can achieve (Jantsch & Tammemäe, 2014a) described in Section 2.2.3, the situation awareness levels will be called *Endsley’s levels of situation awareness*.

Situation awareness (SA) can be thought of as an internalised mental model of the current state of the operator's environment. All incoming data from the many systems, the outside environment, fellow team members, and others must all be brought together into an integrated whole. This integrated picture forms the central organising feature from which all decision-making and action takes place” (Endsley, 2006, p. 528). The concept includes both aspects of the process of gaining situation awareness and the resulting product–a state of awareness–that will initiate further search for information or reasoning to extract meaning or to project to the future.

Situation awareness places high demands on mental resources and processing. The conscious analysis of a problem or the situation at hand and the direction of attention necessary for that purpose are executive functions performed by the working memory. With experience, elements of situation awareness become part of schemas and mental models in long-term memory. This allows for top-down processing of information that is less demanding on mental resources as it replaces bottom-up processing of external information by recognition from memory. Fast recognition of familiar situations and reliance on routines free mental capacity. Mental processing for the development of situation awareness and for the control of actions is substituted by a domain-specific, hierarchically organised repertoire of “if-then-procedures” for familiar tasks and situations. This has been identified as recognition-primed decision-making, a characteristic of experts working in their context of naturalistic decision making (Zsambok & Klein, 1997). Consequently, and relevant for the measurement of situation awareness: Not all aspects of situation awareness need to be conscious, nor do aspects need to be persistently kept in mind. Some information is externally accessible and does not load memory (Durso & Sethumadhavan, 2008). And it can immediately trigger contents from long-term memory that offer possible interpretations for the situation and future states and outcomes from experience. In addition, the knowledge about relevant situations is connected to solutions that proved to be adequate for those situations (‘if-then productions’ that specify when a (cognitive) act should take place; Anderson, 1982). Based on these capabilities humans can make quick and effective decisions in complex situations that would otherwise overload the capacity for mental processing.

Situation awareness is the nub of the matter for adaptation to situation requirements and chose adequate control strategies and problem solutions (Haeusler et al., 2012). It is justified to emphasise on this aspect of human and system performance, even if the theoretical concept of situation

EUROPEAN PARTNERSHIP
Co-funded by the European Union

awareness has faced considerable criticism (e.g., Dekker, 2015; Flach, 1995). The following chapter addresses different facets of the concept situation awareness, including shared or team situation awareness.

2.1.1 Aspects of Situation Awareness Concept

This chapter explains important aspect of human situation awareness. For each aspect that is outlined below, the last paragraph explains how the topic is related to ATCOs' work and reflects on how ATCOs could be supported be machine situation awareness.

Forming an adequate mental picture on what is happening and what could be in future is based on processes on the level of perception, interpretation, and projection (Endsley, 1988). For example, if an ATCO notices that two aircraft are on the same altitude, this corresponds to level 1 situation awareness – the perception. If the ATCO gets aware that this causes a conflict, because these aircraft have crossing flight paths, that represents level 2 situation awareness about the comprehension of the meaning in relation to the goals. Being able to anticipate the future positions of aircraft and figuring out the remaining distance and time until minimum separation will be lost, if aircraft continue on the present heading and altitude, is an example of level 3 situation awareness–projection.

2.1.1.1 Endsley's Level 1 Situation Awareness: Perception

Perception is about encoding information from the outer world to become consciousness about them. Directing and focusing attention on task-relevant aspects is key to human perception, as bottlenecks limit the overall capacity for attention and hence the amount of information and speed of processing (Matthews et al., 2000). This is emphasised by Jones' and Endsley's (1996) findings that the majority (76%) of situation awareness errors of pilots occur on the level of perception. Information processing is improved, if information is presented in different modalities, as they then may be processed in parallel (Wickens, 2002). Different attentional sources exist for different modalities of the senses (Wickens, 1984). Limited attention is a challenge for information intake and situation awareness that might be mitigated with machine situation awareness.

For skilful performance, novices need to learn subtle signals that convey highly relevant information about for situation awareness. Instruction on effective scanning techniques (e.g., Palma Fraga & Kang, 2021). Lack of cue salience can decrease scanning performance, but can be compensated partially by ensuring sufficient knowledge is provided in training for appropriate scanning (Schaninger & Hofer, 2004).

Challenges for situation awareness on the level of perception are rooted in the selectivity in attention. Goals are an important mean to direct information to relevant aspects of the situation for task accomplishment. This might lead to *inattentional blindness*, when important information gets ignored, when it does not fit into the focus of attention, even if it is readily available.

On the situation awareness level of perception ATCOs need to extract information from labels attached to aircraft that are moving on the radar and take up information from radio communication with pilots. The labels themselves are constantly moving and the information within the labels may change. Attention plays an important role in perception. Because attention capacity is limited, not all information can be taken up. Prioritisation and direction of attention are therefore important. Expert ATCOs can easily filter out irrelevant information and focus quickly on the most important. Machine situation awareness can be used to ensure that vital information does not get missed. However, it is

EUROPEAN PARTNERSHIP
Co-funded by the European Union

required to have adequate filtering to notify only about relevant aspects for maintaining adequate situation awareness. That way machine situation awareness can mitigate the threats of workload, attention fixation and tunnel vision, distraction, and fatigue to degrade ATCO situation awareness.

2.1.1.2 Endsley's Level 2 Situation Awareness: Interpretation

On a second level interpretation of the gathered information is accomplished. The meaning and significance of information need to be recognised considering the task to be performed and in the specific context of operation. This process is enhanced by top-down processing of information that is stored in and retrieved from memory as activation in neural networks. In ATC multiple pieces of information and their interactions need to be considered (e.g. differences in aircraft speed and the resulting reduction of separation) and integrated to make an appropriate decision and perform well. Jones and Endsley (1996) found that 20% of the situation awareness errors of pilots occurred on this level.

On the situation awareness level of interpretation experienced ATCOs can immediately recognise the situation upon single and mostly subtle cues of the situation. They generate expectations. In contrast, novices need to process the information they perceived to infer the meaning. Due to their experience and routine, expert ATCOs possess more refined mental models and a broad repertoire of schemas about situations that may occur. Developing a correct understanding of the available information can sometimes still be challenging. Automation tools can help interpretating information and issuing alerts, when the situation is critical. Examples are Short-Term Conflict Detection (STCD) or Medium-Term Conflict Detection (MTCD) for ATCOs or Boeing's Engine Indication and Crew Alerting System (EICAS) or Airbus' Electronic Centralised Aircraft Monitoring (ECAM) for pilots. The capability of machine learning technology to recognise patterns might offer ATCOs support in interpreting information about complex processes regarding air traffic monitoring tasks (e.g., options to optimise service for pilots). It might offer a second opinion or help novice ATCOs with insufficiently developed mental models. Or it might prevent expert ATCOs form overreliance on mental models. However, for distributed human-machine team situation awareness to be useful, it would be necessary that the plausibility of the ML inferences can be checked and that underlying assumptions can be traced back (comprehensibility and transparency).

2.1.1.3 Endsley's Level 3 Situation Awareness: Projection

Level three situation awareness is the projection of current trends to the future to foresee what might come and what could change. Jones and Endsley (1996) found that 4% of situation awareness errors that pilots committed occurred on that level. This requires mental simulation and inferential reasoning.

Mental simulations as well as other processes involved in storing, integrating, and maintaining the current internalised mental mode up to date rely on working memory. Its capacity is limited to 7 plus/minus 2 chunks (Miller, 1956). A single junk can integrate a great amount of information if it is organised by meaningful associations. For reasons of manageability, the number of "system elements" should not exceed seven in system design. However, under stress, working memory capacity gets further reduced (Luethi et al., 2009). If capacity limit is exceeded, critical information can be forgotten or wrongly remembered and integration of information in situation awareness level 2 and 3 can fail. Cowan conceptualises working memory in his model of cognition as an activated subset of long-term memory (1988). Especially novices are affected by working memory limitations in their situation awareness, while experts rely more strongly on their long-term-memory for situation awareness (Sohn & Doane, 2004).

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Schemas and mental models[1] are elements of long-term memory that substitute working memory processes for interpretation and projection of information. They direct attention to critical cues, provide expectations and interpretations and advise on appropriate actions as a single-step connection (Endsley, 1995c), what Anderson called "if-then productions" (1982). Experts have learnt highly detailed classifications for relevant situations. Their superior memory stores allow sophisticated understanding of critical cues and more precise situation awareness. They spend ongoing active effort to project likely or high consequence events and create contingency plans to avoid or quickly deal with possible negative events (Endsley, 2018).

Top-down goal-driven processing of information enhances the efficiency of information processing for situation awareness, as the information is prioritised according to goals and criticality. It operates in close interaction with bottom-up processing in which salient cues activate appropriate goals and models. This interplay allows attention to focus on information relevant for the task at hand and the choice of a suitable strategy. This selectivity in human awareness for information might be challenging to incorporate in human-artificial team situation awareness. Human beings by nature need to focus and switch focus in the service of the task they are accomplishing. Bringing in additional information (machine situation awareness) must consider the human need to focus and stay focused.

Level 2 and 3 situation awareness provide expectations about what information needs to be looked for, when, and where. They drive attention to information cues (level 1 situation awareness) that serve to confirm or deny expectations on situation awareness level 2 and 3. That way, understanding drives attention and perception (top-down) and perception (re-)shapes understanding (bottom-up). However, preconceptions may lead to situation awareness level 2 errors. And confirmation bias – a tendency to exclusively look for information confirming one's expectations – may inhibit maintaining an accurate situation awareness. This can be a trap for experienced ATCOs and lead to *professional blindness*. Characteristics of long-term memory of experts allow for more differentiated categorisations of the information perceived.

On the situation awareness level of interpretation expert ATCOs have expectations from experience stored in long-term memory. To project from the current situation to the future is a challenging task with regard to a three-dimensional space with multiple aircraft and environmental dynamics (e.g., airspace availability, weather etc.). ATCOs accomplish this task with 2D radar screen and their mental models. As the entire system is very dynamic, the ATCOs need to reassess the situation regularly to detect changes, to recognise impacts on the future and to decide on necessary actions or change of plans. Machine situation awareness could protect expert ATCOs from degraded situation awareness due to anchoring effects by their experience and confirmation bias, that lead people to miss important information that would be available to them but does not fit into their expectations. ATCOs could use machine situation awareness to cross-check the validity of their situation awareness. Workload, interruptions, and distractions are threats to human situation awareness. Machine situation awareness could serve as a backup and as a second opinion. Its pattern recognition is not biased by frequency or recency of events, their possible interpretations and future trends as it is the case in

[1] Mental models refer to memory structures that contain information about the purpose and form of a system (e.g., an aircraft), explanations of system functioning, system states, and predictions of future states (Rouse & Morris, 1985), whereas schema are prototypical states of the mental model, e.g. patterns of states of relevant system elements for various classes of situation (Endsley, 2018).

EUROPEAN PARTNERSHIP
Co-funded by the European Union

humans. Availability of memory content for retrieval depends on the frequency and recency of exposure.

2.1.1.4 Shared and Team Situation Awareness

Team situation awareness is about ensuring that every team member has access to information, understanding and anticipation relevant for his or her tasks. Therefore, team situation awareness consists of the situation awareness of the individual team members. The term “team” refers to a collaborating group of people having shared goals (Brannick & Prince, 1997). Teams foster redundancy for safety-critical tasks in terms of quality checks of each other’s work and share task load to keep workload at a manageable level.

Team situation awareness reflects the degree to which every team member possesses the situation awareness required for his or her responsibilities (Endsley, 1989) whereas shared situation awareness implies a focus on the degree to which team members have the same situation awareness on shared situation awareness requirements (Endsley and Jones, 1997). So, team members have distinct and overlapping contents of situation awareness. High team performance is expected to result, if each team member has good situation awareness on his/her duties and have congruent situation awareness on shared situation awareness elements (Endsley & Robertson, 2000). Shu and Furuta (Shu & Furuta, 2005) promote the aspect of mutual awareness. Also, from the socio-technical system point of view there is shared awareness between people and systems they interact with. Even if team members have access to the same information, they might have different (unshared) situation awareness due to differences in roles, goals, tasks, and experiences (Stanton et al., 2017). To express this the term “compatible situation awareness” is used. “Transactive situation awareness” is focused on team members’ exchange of situation awareness and thereby acknowledges differences due to individual roles and functions that contradict sharedness of situation awareness (Sorensen & Stanton, 2015).

In ATCO teams–consisting of a radar and a planner controller–the tactical and strategic monitoring tasks are distributed among team members. Despite splitting the main tasks and allocating responsibilities, both controllers are involved in both types of monitoring. For human-machine team situation awareness some information about the situation might be shared (e.g. conflict detection), while others might just be compatibility (e.g. ML estimates of traffic complexity vs. individual subjective workload).

Human-machine team situation awareness could offer effective support in developing human expertise. With increasing experience and level of skill, ATCOs can reduce the amount of information to be perceived and processed by recognition from memory. Research on experts in different domains (chess, different sports, aviation, medicine etc.) has found that experts–among other characteristics–spend much time and effort on the task of situation assessment (Feltovich et al., 1997). Machine situation awareness can prevent overreliance on experience by pointing out available information relevant for a differentiated situation awareness. In that way ATCOs could develop more differentiated mental models for situation awareness in combination with highly adapted solutions for the specific situation and task at hand.

What might be harder to convey in a human-machine team situation awareness are metacognitive reflections on the suitability of strategies: for instance, the reasoning about underlying commonalities and the uniqueness of situations and the adequacy of a performance strategy. Metacognition is the consciousness and control over one’s own thoughts (Flavell, 1979). Metacognition helps relate own abilities to situational requirements, it regulates planning and monitoring processes for goal achievement, it monitors performance and learning processes, adaptation, and necessary transfer

EUROPEAN PARTNERSHIP
Co-funded by the European Union

knowledge to solve new problems (Häusler, 2006). Metacognitive skills allow experts to have a highly effective, hierarchically organised knowledge structure about classes of tasks and situations connected to procedures for task accomplishment adjusted to the specific requirements and to make inferences about novel aspects of the task or situation.

2.1.2 Methods to Measure Situation Awareness

The most common assessment methods for situation awareness are: (1) subjective rating by the participant or a subject matter expert, (2) implicit performance measure, and (3) probe techniques (Durso et al., 1999).

Examples for subjective ratings are 3D SART (Situational Awareness Rating Technique; Taylor, 1990) or for ATC context SASHA_Q (Situation Awareness for SHAPE[2]_Questionnaire; Dehn 2008). Implicit performance measures disclose the level of awareness about important situational aspects. Probe techniques ask questions about the current situation and future development. The technique can be applied in two versions: In the freeze technique, simulation is stopped, and the sources of information (screen) are blanked. In the online technique, questions about aspects of situation awareness are asked while the simulation is continuing. Situation Awareness for SHAPE_Online (SASHA_L; Dehn, 2008) is an example of this type. While queries with frozen simulation and blanked display is used in the Situation Awareness Global Assessment Tool (SAGAT; Endsley, 1995a). Online techniques without freezing are considered less intrusive (Salmon et al., 2009), but involve extra workload (Jeannot 2000). A variety of examples for measurement tools is summarised by Jeannot et al. (2003).

The three methods for situation awareness assessment have different strengths and limitations. The use of the self-rating techniques is widely spread, as subjects have the most direct access to their own awareness. Methodologically however they suffer criterion deficiency, as it is difficult for subjects to recognise aspects they have missed, because they are not aware of them (Jeannot et al., 2003). This is lowering the validity of self-ratings for the assessment of the degree of situation awareness. They represent a measure of the certainty people feel about their situation awareness (Endsley, 1995a) (Endsley, 1995a). Implicit assessments of situation awareness from performance represent a non-intrusive method. However, it is challenging to define adequate performance indicators, because it is not entirely clear how situation awareness and performance are related (Salmon et al., 2009) and performance may be influenced by other factors that situation awareness, which results in criterion contamination. Probe techniques with freezing require subjects to freely recall from memory and therefore partially imply a memory test that might not fully be representative of subjects' situation awareness (criterion deficiency). Online techniques are less intrusive for subjects and do not alter the "fluency" in performance in simulations. It is a real-time assession of situation awareness with no interruption or blanking. But asking queries while a task is performed leads to additional task load and may lead to distraction. Considering the advantages and disadvantages of the different techniques, combining several techniques to compensate for shortcomings and deficiencies has been recommended and chosen in the AISA project.

[2] SHAPE=Solutions for Human Automation Partnerships in European ATM

EUROPEAN PARTNERSHIP
Co-funded by the European Union

2.1.3 Attention and Gaze Behaviour

Eye-Tracking is a popular method to analyse visual attention with gaze behaviour. Attention is the behavioural and cognitive process of selectively focusing on a particular aspect of information, whether it is considered subjective or objective, while ignoring other perceptual information (James et al., 1890). It corresponds to the ability to flexibly use computational resources (Lindsay, 2020). In a recent publication the usefulness of a unitary construct for attention and the neural system has been challenged. “Attention is one of the most misleading and misused terms in the cognitive sciences” (Hommel et al., 2019, p. 2288). Alternatively, subsets of processes and mechanisms that lead to task-specific performance should be investigated separately – acknowledging the interconnected and integrative nature of the human sensorimotor information processing systems. Instead of a single concept of attention, multiple underlying processes are claimed (Di Lollo, 2018). To single out attention from the whole complex of processes is seen as counterproductive to gain a “… comprehensive understanding of human behaviour because it ignores integrated, parallel, and reciprocal relationships among sensory, cognitive, and action processes” (Hommel et al., 2019, p. 2288).

For the investigation of situation awareness attention is considered as the capacity needed for important information (Endsley, 1988) and the ability to select from a multitude of sensory impressions and mental activities and prioritise those necessary for planning and execution of the tasks. Attention is needed in monitoring, in communication and in renewing mental models.

Two concepts are interrelated with attention: alertness and vigilance. Alertness is a precondition for attention and is defined as the state of being awake, aware, attentive, and prepared to act or react (‘Alertness’, n.d.; APA Dictionary of Psychology). Vigilance is the aspect of sustained attention. It “refers to the state in which attention must be maintained over time. Often this is to be found in some form of “watchkeeping” activity when an observer, or listener, must continuously monitor a situation in which significant, but usually infrequent and unpredictable, events may occur” (Vigilance; n.d.; Encyclopedia Britannica). ATCOs need vigilance when watching the radar screen to detect an aircraft as soon as possible.

“Gaze” is described in various disciplines (e.g. sociology, philosophy, psychoanalysis, etc.) as an awareness and perception of an object, a group or oneself (Wikipedia, 2022). Gaze behaviour is conceptualised as a fixation of gaze on a location in the targeting environment or as a shift in gaze from one environmental location to another (Zangmeister & Stark, 1982). Gaze shifts can be initiated by the eyes, the head, and the body.

Eye-trackers offer a method to measure gaze behaviour and visual attention. Gaze behaviour is conceptualised as an indicator for attentional prioritisation (Ernst et al., 2020). Indicators most often used for gaze analysis are: (relative) frequencies of gaze on Areas of Interest (AOIs), fixations and gaze points, heatmaps, time spent focusing (dwell time), average fixation duration, time to first fixation (TTFF), first fixation duration, ratios (proportion of participants e.g., missing an AOI), fixation sequences, revisits.

Gaze direction is normally controlled unconsciously. Visual processing and coordination of eye movements requires the use of many cortical and subcortical regions of the brain. The focus of attention is guided bottom-up by salient visual cues and top-down by the goals, intentions and re-activated knowledge and expectations. This interplay to search and select task-relevant information is highly automated and hence not accessible to consciousness. It may be considered a result of learning

EUROPEAN PARTNERSHIP
Co-funded by the European Union

history and as an expression of the level of expertise in terms of understanding the complex nature of the specific domain and tasks.

2.1.4 Effects of Workload and Stress on Situation Awareness

Active information processing requires cognitive resources which are limited for mental operations (Kahnemann, 1973). Knowing and applying task strategies that lower the need for mental resources are signs of expertise and allow for superior performance and resistance to workload and stress (Häusler, 2006).

Task load corresponds to the level of demand or stress originating from external factors–the task or the system (e.g., complexity or time pressure). The resulting strain from the impact of the external stressors is called mental workload. The level of workload experienced is subjective and depends on individual constitution, resources, abilities or level of experience, and processing style (Cain, 2007; International Organization for Standardization, 1991). A change in task load can therefore imply different levels of workload. In ATC research task load is often operationalised as the number of aircraft: the more aircraft to control, the higher the mental workload (Ahlstrom & Friedmann-Berg, 2006).

Increased workload can be compensated by investing more effort, prioritisation of most important aspects or/and simplification of the task by an adaptation of strategy. There is evidence that high workload leads to poor performance and a higher number of errors. If the demands start to exceed the capacity, skilled operators either adjust their strategy or performance starts to degrade (Young et al., 2015). Skilled ATCOs adapt their strategies with increasing task load to prevent exhaustion of mental resources: They prioritise more strongly and simplify the information processing involved in handling aircraft (more standardised) and thereby save mental resources (Sperandio, 1978).

Workload can be assessed using subjective rating scale (e.g., NASA Task Load Index (TLX); Hart & Staveland, 1988), with a secondary task approach or with psychophysiological methods. Latter offer an interesting access to contemplate workload as they assess physiological functioning and reactions to task load in a non-invasive way. They are used in engineering psychophysiology to address questions and problems of engineering psychology through scientific study of human interactions with technology. Psychophysiology is the study of the interrelationships between mind and body (Schell & Dawson, 2001). Typical psychophysiological measures are heart rate, skin conductance, and skeletal muscle activity. They capture states of arousal and emotion, represent cognitive processes, and are used to analyse behaviour.

Backs and Boucsein (2000) suggest a general framework that integrates physiological reactions and clarifies their relations to demands, processes, reactions, and outcomes. Demands may be task-related, emotional, or physical. They call for cognitive, affective, and energetic processes to react to the demands. The reactions that result from processes can be measured on the behavioural, subjective, and physiological level and lead to outcomes such as productivity, well-being, health, motivation, skill, and expertise.

From physiological reactions measured inferences are made about the state of the subject. The state is the result of many physiological and psychological reactions to task demands or environmental stressors. Those reactions regulate the brain and the body and serve the goal of enabling optimal accomplishment of the demands from the work environment. For instance, a discharge of adrenaline increases blood pressure, pulse, skin conductance and muscular activity; it puts the organism on alert

EUROPEAN PARTNERSHIP
Co-funded by the European Union

and raises the body's readiness to perform. On a generalised level, physiological indicators provide information on how difficult it is for a subject to fulfil a task compared to his/her relative baseline in a relaxed state (Friedrich et al., 2018).

The transition of psychophysiological methods from the laboratory to operational environments is still challenging. A multitude of variables may influence psychological reactions, and artefacts in measurements arise from interferences induced by body movements or changes in breathing rate during verbal expression with impact for psychophysiological parameters.

2.2 AI and Machine Situation Awareness

The capture of human intelligence is a complex endeavour that started 150 years ago and is still ongoing–differentiating new factors of intelligence and measuring the world of thought with high-tech methods. AI technology is providing intelligence to systems, mimicking human reasoning and inference for different cognitive tasks. The following chapters provide a brief overview on automation and machine learning.

2.2.1 Automation and Monitoring

Automation refers to a wide range of technologies that reduce human intervention in processes. They are reduced by defining decision criteria, sub-process relationships and associated actions in advance and having them executed by machines. For that control systems, machines and information technology is used to increase productivity (Groover, 2019).

Monitoring is the observation and check of a system, quality, or progress over a period of time to notice unexpected behaviour. Therefore, automation in a monitoring context means that the automation software/machine is allowed to react to inconsistencies, either by fixing them itself or by alerting an authority or system.

A lot of automation can be found in ATC. It should provide ATCOs with early and accurate information, help to increase visibility at airports and improve communication with pilots. It is widely used in navigation, communication, and surveillance tools, to support the ATCOs.

It is used in:

- Short-term conflict alert system: automated warnings are used to let the tower ATCO know when an aircraft is heading to a runway which is already occupied
- Remote digital towers: automation is used to transmit data to different control centres, and it also enables that data of an aircraft can be expanded with further information on the screen
- Crossing detection tool: automation detects and warns the ATCO of conflicts
- Routes: automation is used to check route clearance and to find the most efficient one
- Tracking: digital tracking methods are progressively replacing flight progress strips. Therefore, telephone conversation is now automated, which is reducing workload and increasing ATCOs' capacity
- Time-based separation procedures for aircraft: computer-generated indicators are projected on radar during approach to provide ATCO with better guidance on minimum separation limits. This also results in more efficiency of approach handling and thus less delays and cancellations occur

EUROPEAN PARTNERSHIP
Co-funded by the European Union

- NextGen program (automation system) (Darr et al., 2008): Standard Terminal Automation and Replacement System (STARS) & En-Route Automation Modernisation (ERAM): enhanced tracking of aircraft though automatic dependent surveillance-broadcast systems and data block feature (automatically lists the number of aircraft in airspace)

Automation brings a lot of benefits, but also some disadvantages. The biggest problem is the possibility of failure, which is why it is especially important that the operators get notified when it has malfunctions, and that ATCOs are aware that it can happen. Furthermore, ATCOs should be trained on new automation tools and there should be options if the systems fail. This plays a significant role, especially with remote towers (Durso & Manning, 2008).

2.2.2 Machine Learning

Machine learning (ML) is a subset of artificial intelligence, aimed at automated model building. Created models/algorithms solve tasks by establishing relationships between input and output data, which can then be applied to new data to generate predictions. The established relationships are not known to the user but are contained in the system, usually in a form unclear to humans. For this reason, ML systems are often regarded as “black boxes” when discussing the mechanism by which they generate predictions.

Since ML systems use data to both “learn” how to accomplish a task and to generate new predictions, a distinction must be made between the data groups:

- Training data, used by the system to learn how to solve a task
- Validation data, used to tune the parameters of the model
- Test data, used to get an evaluation of the final model (also called holdout data, if it has not been used for training the model)

Choice of data for the model creation process will influence the final model, so care must be taken to avoid common problems such as overfitting – creating an overly complex model which fits the training data perfectly – or bias – where the model notices unconscious biases present in the dataset. These problems must be accounted for during the development of the model. Other checks are available for the execution phase – for example, metadata of the testing dataset used for the conflict detection ML module was added to the knowledge graph to serve as a preliminary check for the operation of that module.

2.2.3 Machine Situation Awareness

It is challenging to capture human intelligence and integrate it into AI. It is also difficult to measure intelligence and SA. Nevertheless, there are techniques to assess the situation awareness of AI:

- Integrating a neural network in the AI system so that the system learns to estimate itself
- Evaluation of AI generated outputs by experts
- Enhancing the AI with other technologies so that it learns to estimate its own performance

For instance, AI systems can be expanded with local optimisations, approximate reasoning, and neural networks that simulate expectation based on previous training (*Can AI Systems Match Human-Level Situational Awareness? | Bench T*, n.d.). Another possibility is to estimate the situation awareness of

EUROPEAN PARTNERSHIP
Co-funded by the European Union

AI by having SMEs validate the generated output (Pullum, 2021). SME problem solving can also be used in comparison with novices and AI – the difference in approach and results can be analysed and, depending on how AI performs, its situation awareness can be estimated.

Jantsch and Tammemäe (2014a) developed a framework for assessing the awareness level of AI systems, positing that analysis of awareness and self-awareness and their implementation into AI systems is a way to improve those systems' robustness. They present seven conditions for awareness, with 5 being conditions of being aware of a certain property *P* and 2 conditions for a subject (system) to be aware of itself:

Table 3: Conditions for AI system awareness

	Condition Code and Name	Condition Description
CONDITIONS FOR AWARENESS OF A PROPERTY	(C.1) Meaning Condition	Subject makes physical measurements or observations that are used to derive the values of property *P* by means of a meaningful semantic interpretation.
	(C.2) Robustness Condition	The semantic interpretation is robust.
	(C.3) Attribution Condition	There is a semantic attribution which is meaningful.
	(C.4) Appropriateness Condition	The subject's reaction to its perception of *P* is appropriate.
	(C.5) History Condition	A history of the evolution of the property over time is maintained, in particular of the increasing or decreasing deviations over time.
CONDITIONS FOR AWARENESS OF SELF	(C.6) Goal Condition	The subject can assess how well it meets all is goals, thus having an understanding which goals should be achieved and to which extent they are achieved.
	(C.7) Goal History Condition	The subject can assess how well the goals are achieved over time and when its performance is improving or deteriorating.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

A brief description of the five awareness levels, with an emphasis on conditions necessary to reach each level, is as follows:

Table 4: AI system awareness level classification

Awareness Level	Necessary Conditions to reach Level
AWARENESS LEVEL 0 (FUNCTIONAL SYSTEM)	• System output is a mathematical function of inputs (always reacting in the same way to inputs) • System fulfils conditions (C.1) to (C.4)
AWARENESS LEVEL 1 (ADAPTIVE SYSTEM)	• System is adaptive, meaning that it tries to minimise the difference between input and reference values by use of a PID controller or similar algorithm • System fulfils conditions (C.1) to (C.4)
AWARENESS LEVEL 2 (SELF-AWARE SYSTEM)	• System is aware of at least one property and one environment property according to (C.1) to (C.4) + (C.6) • System contains an inspection engine which periodically derives one integrated attribution of the subject as a whole • System computes its actions based on (a) monitored and attributed properties of the system and of the environment
AWARENESS LEVEL 3 (HISTORY SENSITIVE SELF-AWARE SYSTEM)	• System fulfils all requirements of an Awareness Level 2 system • System fulfils the history conditions (C.5) and (C.7)
AWARENESS LEVEL 4 (PREDICTIVE SYSTEM)	• System fulfils all requirements of an Awareness Level 3 system • System's decision-making process involves a simulation engine which can predict the effects of actions on the environment and the system itself and, in case of an anomalous result, search through simulations for the best action
AWARENESS LEVEL 5 (GROUP-AWARE SYSTEM)	• In addition to being self-aware, the system distinguishes between itself, the environment, and the peer group (which is treated differently because of its own set of expectations and goals)

The original article does not offer a mechanism for applying this classification to existing AI systems. The conditions and awareness levels will therefore be used as guidelines, but preliminary analysis shows that an informed choice regarding system scope must be made since it will influence the classification results. The classification of the AI SA KG system will be performed in its own section of the "Results" chapter of this document.

To obtain better situation awareness and to improve the accuracy of the acquired situation awareness, intelligence from multiple sources is required to filter the discrepancies reported from a particular intelligence source (Munir et al., 2022). The intelligence itself depends on how the machine was developed and with what data it was trained or tested.

2.3 Human-Machine Team Situation Awareness

Human-machine collaboration is investigated for simple interactions with robots in production (Buxbaum, 2020). There, challenges consist in the costs of attentional resources to ensure coordination of human and robot actions in a flexible assembly production, as well as in the modality in which

EUROPEAN PARTNERSHIP
Co-funded by the European Union

information on the state of assembly process and the current task performed by the robot should be conveyed to the human operator. ATC is presenting a much more dynamic environment for human-machine collaboration than manufacturing. The chapters below outline the concept of human-machine team situation awareness and a systematic to describe the level and quality of awareness that can be reached by a human-machine situation awareness system.

2.3.1 Distributed Human-Machine Team Situation Awareness

As described in section 2.1.1.4 team situation awareness with technical systems focusses less on sharing information, but more on mutual and compatible situation awareness to reach a commonly understood mental image of what is happening and what is going to happen. In terms of the AISA concept, machine/system should be treated as a person because, in this scenario, the system is part of the team. Figure 2 depicts the ATCO-AI situation awareness system in collaboration presented in D2.2 (AISA Consortium, 2020b).

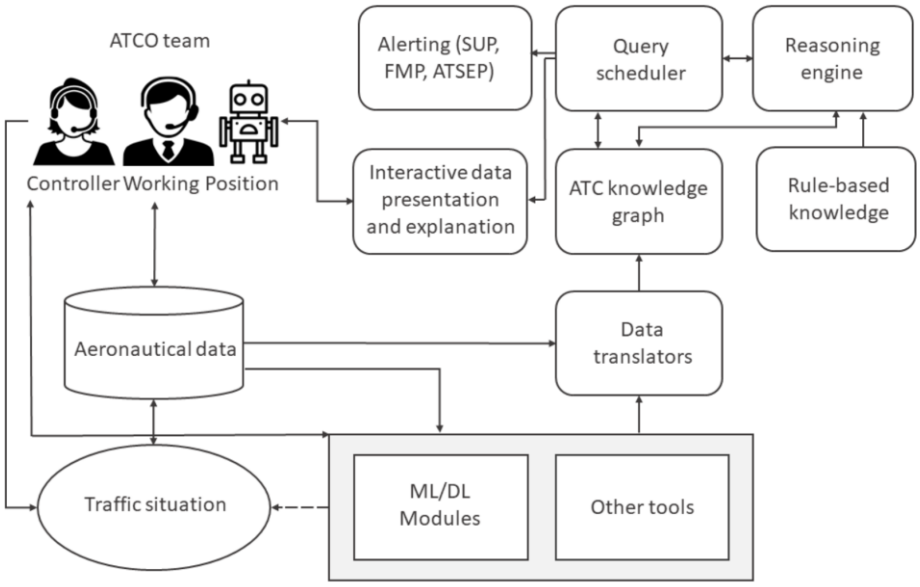


Figure 2: Conceptual diagram of the system including ATCOs and AI situation awareness system

Figure 2 depicts the sharing of information sources for human and machine situation awareness. This will be further described in the chapter below. If the goal for effective human-machine team situation awareness is to achieve mutual understanding, then ATCOs need to be able to understand machine situation awareness (comprehensibility and trust) and the machine situation needs to be aware of the system state (the state of its subsystems and of ATCOs).

2.3.2 Comparison of Human and Machine Situation Awareness

Sharedness and/or compatibility of information between ATCO and AI situation awareness system for human-machine team situation awareness is depicted in Table 5.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Table 5: Knowledge represented in the AI SA KG (from D2.1)

AI SA Knowledge and Information Sources	ATCO Knowledge and Information Sources
Static information (from AIXM, …)	The map on the radar screen with info about airports, navaids, etc.
Situation-specific situations (weather, aircraft positions)	Radar screen with current positions of aircraft and additional aircraft information
Predictions (from ML module)	Augmented radar screen with predicted trajectories + alerts
Logically derived information (from rule-based reasoning)	Implicit or explicit thoughts/judgements in the ATCO's mind/memory
ATCO's actions (activity log of human-machine interactions)	ATCO's self-observation and observation of colleagues
The provenance of 1.-5.	ATCO's knowledge about equipment + ATCO's reflective thinking

2.3.3 Aspects of Human-Machine Collaboration for Situation Awareness

Human situation awareness is built by combined bottom-up and top-down information processes (see 2.1.1). The environment and systems can support human situation awareness, if …

- the system can provide the relevant information (e.g., sensors, data transmission capabilities, networking),
- the interface design makes critical information available in an effective format for transmission of information (e.g. modality, display design),
- the system complexity and more specifically the number and interrelatedness of subcomponents in combination with the rate of information change allow to keep track of new inputs and changes,
- the level and design of automation allow the individual to stay "in-the-loop" and understand what is happening and what the system is doing,
- stress, fatigue, and workload as a function of the task environment and the system interface are minimised (Endsley, 2018).

Inadequate design of support systems can lead to mental load and time pressure which in turn may create problems for situation awareness such as attentional narrowing, where attention to peripheral sources of information is decreased (Broadbent, 1971). Other effects include premature closure and shortcuts to information processing that lead to decision without exploring all information available (Keinan, 1987) as well as poorly organised scan patterns, which may lead to level 1 situation awareness problems (Janis, 1982). A reduction in working memory capacity and deteriorated retrieval may negatively affect level 2 and 3 situation awareness.

These aspects will be relevant for human-machine interactions and the respective human-machine interface of the future AI SA system.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

3 Methods

This chapter summarises information on the experimental plan and the methods for data acquisition and analysis. Additional information on technical details are reported in the appendix section.

3.1 Experiments

To meet the requirements of task 5.1 Comparison of situation awareness between AI and ATCO and task 5.3 Human performance in distributed SA defined in the Grant Agreement No. 892618, two experiments were conducted with human-in-the-loop simulations. Experiment 1 took place from 8 to 12 November 2021, experiment 2 was held from 11 to 14 January 2022.

The impact of machine situation awareness on human performance was studies with a between-subjects design comparing human performance – the dependent variable – across the working conditions "without AI SA input" (experiment 1) and "with AI SA input" (experiment 2) – which represent two independent variable conditions. The experimental plan of the AISA project can be found in Appendix L.

Table 6 Between-Group Comparison for effect of AI SA support

Independent Variable conditions	Groups	Subjects	Dependent Variables
Without AI SA Support	Control Group	N= 20 ATCOs (Experiment 1)	Human Performance: Start of conflict solution, Conflict duration
With AI SA Support	Experimental Group	N= 16 ATCOs (Experiment 2)	

Experiment 1 is a human-in-the-loop simulation and provided recordings for situation awareness of ATCOs, for a baseline of human performance without "AI SA input" in an environment similar to ATCOs work conditions at Skyguide, and for computing AI SA outputs by the AI SA system. The exercises done by ATCOs were converted to RDF graphs and were fed into the AI situation awareness system to generate artificial situation awareness–AI SA outputs. Those outputs could later – in the second experiment - be compared with the situation awareness of ATCOs using query technique. In addition, the AI SA system's outputs were used as "AI SA inputs" to ATCOs in experiment 2 to explore human-machine team situation awareness.

Experiment 2 started with one human-in-the-loop scenario, followed by scenarios where ATCOs were in a "watch only" role and complied with the procedures that an ATCO from experiment 1 had selected. This was necessary to standardise the scenarios as the AI SA system at the current stage is not able to operate in real-time. For the ATCOs, the role in the "watch only" scenarios was similar to the job of a coach in the ACC. They were asked to observe the situation and provide answers about specific aspects of situation awareness.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

The experiments were conducted on two working stations in parallel. Figure 3 shows the timeline of experiment 2. Measurements were taken during the scenarios. At the end of each scenario, questionnaires were filled out. A debriefing questionnaire followed at the end of the experiment.

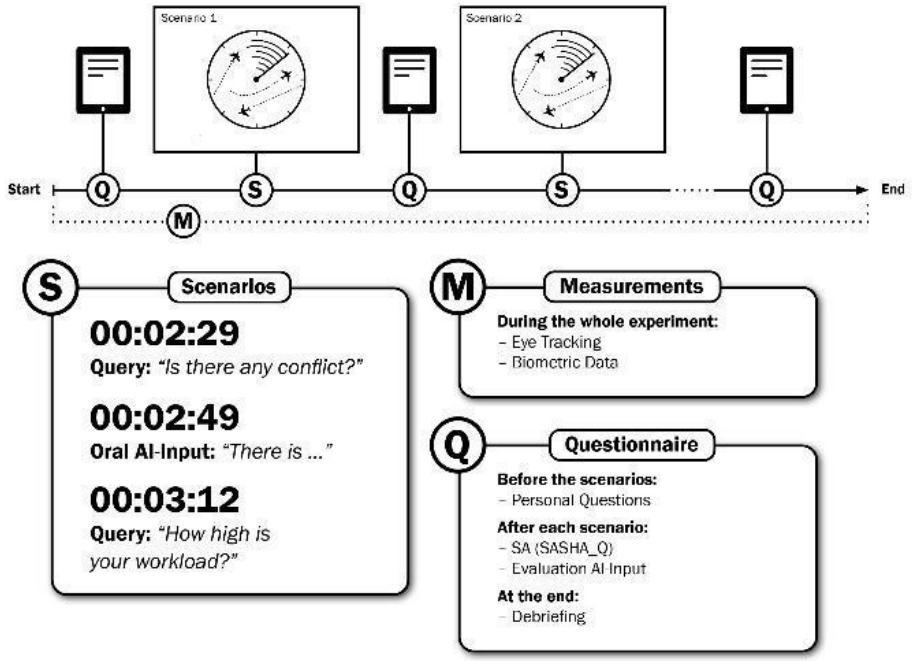


Figure 3: Overview of experimental setup regarding the approximate timing of measurements

Information about the variables measured (Section 3.8 and following), the socio-demographic analyses for the participants (Section 3.6) and the methods for data collection and processing (Section 3.11) are provided in the respective chapters.

Experiments were realised on site in the facilities of the air navigation service provider and AISA consortium partner Skyguide in Dübendorf (Switzerland). Apart from providing the facilities for the experiments, Skyguide strongly contributed to the successful execution of the experiments by means of consulting, accompaniment through subject matter experts (SMEs) and the provision of ATCOs as experiment participants.

As a further consortium partner, the Faculty of Transport and Traffic Sciences (FTTS) of the University of Zagreb contributed the setup of the simulation (Section 3.2). Members of FTTS created the scenarios, accompanied the experiments as pseudo-pilots in the simulations and accomplished many tasks to prepare for experiment 2.

The Zurich University of Applied Sciences (ZHAW) held the lead of work package 5 in the AISA project and was responsible for the planning and execution of the experiments and analysing results for some of the research questions.

During the experiments, different data were recorded including:

- Eye tracking recording for each scenario
- Screen recording for each scenario

EUROPEAN PARTNERSHIP
Co-funded by the European Union

- Frontal recording over the whole experiment for each ATCO
- Biometrical data over whole experiment for each ATCO

The post-processing methods for the collected data are described in subsequent chapters.

3.2 Simulation Tool

For the experiments, a computer-based simulation program was used. “EUROCONTROL simulation capabilities and platform for experimentation" (ESCAPE) is a scalable EUROCONTROL ATM real-time simulation platform supporting small- and large-scale simulations. ESCAPE is also available on a light platform, which was used in the experiments. The ESCAPE Light simulator is a lightweight yet high-performance version of the ESCAPE software (https://www.eurocontrol.int/simulator/escape). It was installed on 4 laptops, allowing the experiments to run on two positions simultaneously. Two laptops corresponded to the controller working positions, and the remaining two were used for starting the exercises and hosting the pseudo-pilot working positions.

Scientific associates from FTTS populated the simulation program with actual Swiss en-route traffic data from the 4 July 2019 and further enriched the scenarios by modifying flight paths and adding or removing aircraft in accordance with the Skyguide SME in order to generate situations of a large variety. The vast majority of the trajectories used in the simulations are actual flown trajectories, therefore the raw data did not contain conflicts or situations of interest since ATCOs have done their job to separate the flights. The modifications to the flight paths were minor, limited mostly to slightly changing the entry time of flights or adding more points to the planned routes. If the SME estimated the actual workload to be too high for the purposes of the experiment, some flights were deleted, or their trajectories changed to reduce the workload. The goal of the modifications was to achieve pre-planned conflicts and situations where specific ATCO input is expected. This flight data was previously excluded from training for the AI. That way ***generalizability*** of AI situation awareness to new data is investigated.

To match the operational context of the ATCOs, ESCAPE Light was adapted to include measurement tools ATCOs are used to working with (such as distance measuring). Nevertheless, some differences persisted to the system SkyVisu, which is developed by and implemented at Skyguide. The tools mentioned in the list below are not included in ESCAPE Light, only in SkyVisu. Furthermore, the operation of the systems is different, i.e. the inputs of the commands and handling of the flight label are different, and the mouse buttons in SkyVisu are programmed with more functions and buttons than the ones used in the experiments. The mouse function has been modified to be more similar to the ones used in the daily operations by the participants.

Conflict manager: Conflicts are detected by the Conflict Manager (CM) function, which is a 3D evolution of the Horizontal Scanning Tool (HST) and Dynamic Scanning Tool (DST). The CM permanently scans for conflicts and encounters based on aircraft current position and trajectories. When conflicts or encounters are detected and the conflict display criteria are met, the CM displays the conflicts.

E-coordination: Used for electronic coordination between two Area Control Centre (ACC) sectors. For example, a direct point or another level can be given electronically. This is not an important tool for the experiments since there were only one-sector exercises.

Direction finder: Every time an aircraft calls on the frequency, the antenna of Skyguide picks up the message and can determine from which direction the message came. A thin line is then displayed on

EUROPEAN PARTNERSHIP
Co-funded by the European Union

the radar showing the approximate direction of the aircraft. Therefore, it is possible to quickly detect the aircraft. In ESCAPE Light this function was not available.

Entry window: With the help of the entry windows, it is possible to see which planes are coming from which direction, at which altitude and at which time. Each entry sector has its own entry window. So, it is possible to plan earlier if a crossing occurs and how to solve it. In ESCAPE Light an entry window could have been shown, but only one for all sectors, which would not have supported the overview.

3.3 Scenario Descriptions

The subsequent chapters describe the scenarios used in the experiments. The order of the scenarios was randomized to avoid systematic influence of learning over time. Although all the participants were presented the same scenarios, not all ATCOs encountered the same traffic situations in the exercises. This is because the traffic situation changes with every ATCO clearance, leading to some crossings being unintentionally created by the ATCOs themselves, but also certain planned crossings becoming either more complicated or simplified by the controllers' actions. For the convenience of the understanding, each scenario was given a specific abbreviation which indicates if it was conducted as a part of experiment 1 or 2 (E1 or E2), and an additional number used to distinguish one scenario from another (STR for training, S1, S2, S3 and S4).

3.3.1 Experiment 1

Experiment 1 includes a total of six scenarios. The participants always started with the **Training scenario (E1ST)**, which is an introductory exercise where the ATCOs would have a chance to get used to the differences in the human-machine interface (HMI) compared to their system. In the E1ST scenario, two SMEs from Skyguide who were already familiar with the environment were there to coach the participants. The E1ST scenario lasts for approximately 20 minutes and includes 26 aircraft. The goal was for each participant to issue all the instructions they normally would when working on live traffic and get a grasp of the available tools.

After the Training scenario, each participant would do the **Light scenario (E1S5)** which lasts 7 minutes and includes 19 aircraft. There were no intentional non-compliances by the pseudo-pilots and the traffic complexity was low. The Light scenario was not used for the situational awareness analysis, but rather just for the ATCOs to get comfortable with the simulation platform.

After the E1ST and E1S5 scenario which were done in the same order for all participants, the rest of the scenarios explained below were randomised in order to avoid the measurement of a learning effect in the results.

Short (E1S1): This scenario lasts five minutes and includes 19 aircraft. There is one planned non-compliance regarding the change of flight level.

Crossing (E1S2): This scenario lasts for 15 minutes and includes 24 aircraft. The main focus of the scenario is a triple crossing that needs to be solved before the loss of safe separation occurs. In the same scenario, there is a non-compliance by the pilot that results in a flight level bust. The timing of this conflict depended on the pseudo-pilots.

High traffic (E1S3): This scenario lasts 15 minutes and includes 27 aircraft. There is a planned non-conformance concerning the speed control that is necessary for one flight pair where the following

EUROPEAN PARTNERSHIP
Co-funded by the European Union

aircraft is faster and catching up to the leader. However, the choice of solution for that pair lies on the ATCO, so the non-compliance is not present for each scenario run. There are additional four crossings present in the scenario design.

Military (E1S4): This scenario lasts 23 minutes and includes 32 aircraft. There are two TRAs (Temporary Reserved Areas) in the simulated sector – one in the central part of Switzerland (referred to as Mil Centre) and one in the eastern part (referred to as Mil East). Mil Centre is not available for commercial traffic in the beginning of the exercise. It becomes available two minutes after, and the aircraft can be cleared to fly direct routes through the previously restricted area. Mil East activates in the second part of the scenario and remains unavailable for commercial traffic until the end. The challenge is to re-route all aircraft whose planned trajectories are crossing the TRA. In addition to the activation and deactivation of the TRAs, there are two crossings included in the scenario design.

3.3.2 Experiment 2

A significant difference between the experiment 1 and 2 is the handling of the scenarios. In experiment 1, all scenarios were interactive, whereas in experiment 2, only one scenario (apart from the **training scenario (E2ST)**) was executed with the possibility of ATCO handling the traffic. All others were “watch-only", meaning that the ATCOs themselves had no control. They watched a replay of the exercises from experiment 1 and listened to audio recordings of the exchange between the controller and the pseudo-pilot on the frequency. They had to put in the ATCO instructions they heard into the radar label and had the possibility of using the measuring tool on the controller working position (CWP). That way, they were able to gain situational awareness. Additionally, they were given an audio AI input that was generated by analysing the data collected in experiment 1. The production of the audio AI inputs is below. The scenarios used in experiment 2 are explained below.

The **E2ST** scenario in experiment 2 was prolonged by the simulation developers by merging the E1ST and E1S5 scenarios to give the ATCOs more time to get used to the system since they would get less chance to control the traffic later on. The new E2ST scenario included 32 aircraft and lasted 30 minutes, but the ATCOs could end the scenario sooner if they felt they were ready. After the E2ST scenario, the rest of the scenarios were done in a random order for each ATCO. The only exception is that the E2S2 scenario is always done before the “watch-only” E2S2.2 scenario.

The **crossing interactive (E2S2.1)** scenario started at the same time as in the experiment 1 and lasted for 13 minutes. There were 23 aircraft in the scenario. It included oral AI inputs regarding the predicted conflicts. There were no pilot non-compliances.

The **“watch-only" crossing (E2S2.2)** lasted for 8 minutes and included 23 aircraft. It included the same traffic and had a temporal overlap with the human-in-the loop scenario, which is why it was important to let the participant in experiment 2 have control of the traffic before only observing the playback. The flight level bust was part of the watch-only scenario.

The **“watch-only” high traffic (E2S3)** scenario included the first ten minutes of the original scenario. There were 24 aircraft. It contained crossings and the speed bust, but not the non-conformance. Instead of the speed non-conformance, there was a wrong readback by the pilot that the ATCOs should have noticed.

The military scenario (E1S4) from experiment 1 was split into 2 watch-only scenarios in experiment 2. **Military 1 (E2S4. 1)** started at the same time as E1S4. The main observed situation in the scenario was

EUROPEAN PARTNERSHIP
Co-funded by the European Union

the deactivation of the Mil Centre TRA and following if the controllers will notice the flights that can use the previously restricted airspace, thus improving the quality of service. Military 1 lasted 6 minutes and included 23 aircraft.

Military 2 (E2S4. 2) watch-only scenario starts two minutes after the end of Mil 1. It included the activation of the Mil East area and the re-routing of traffic around it, as well as the exit crossing and the crossing in the sector. This scenario lasted 6 minutes and included 30 aircraft.

3.4 Materials

Audio inputs presented to the ATCOs in experiment 2 were used for two purposes. One was to help the ATCO achieve some degree of situational awareness of the traffic situation they are not in control of themselves, and the other to present the AI inputs so ATCOs could assess their accuracy and usability.

For achieving situational awareness, the frequency transmissions between the ATCO and pseudo-pilot from the selected exercises in experiment 1 were used. They were first written in the form of a transcript and then turned into audio clips using Shotcut (*Shotcut - Home*, n.d.), a free and open-source video editing app. Two different synthetic voices were used, one corresponding to the ATCO and one to the pseudo-pilot transmissions. They were timed to correspond to the video replay of that same exercise.

Following experiment 1, the data from the recorded exercises was used to populate the KG. AI SA tasks were then applied to the data to generate KG system outputs. The full list of AI SA tasks and their status can be seen in Table 2 in Section 1.2.1.4.

Since there was not an option for additional HMI to be developed to visually present the inputs to the ATCOs, and the used simulator does not have that capability, the number of outputs to be presented to the ATCOs depended greatly on the means of presentation. The frequency is a scarce resource, and the transmissions must be clear and concise to take up as little time as possible. Adding another audio element over the already recorded ATCO and pilot transmission means the frequency becomes even more used up. The number of KG outputs is therefore kept to a minimum needed to query the ATCO situational awareness. The KG system outputs that would be used in experiment 2 as inputs were hand-selected based on their relevance and in accordance with the planned queries (see Appendix A.2). The selection process started by including the predicted conflicts where the predicted minimum distance is less than 12 NM. This filter was the first added to reduce irrelevant KG outputs. Each conflict prediction was only presented once. The timing was discussed with the SME. The outputs regarding other situations where there was a recorded degradation of situation awareness in experiment 1, i.e., a flight level bust or speed non-conformance, were then added to the predicted conflicts. The outputs regarding aircraft whose planned trajectories are crossing an active TRA or which can be given a direct route for improved quality of service were also included.

Once the AI inputs were selected from the full list of KG system outputs, they were turned into audio clips in the same way as the controller-pilot transmissions. The voice used was different from the ones selected to represent the ATCO and pseudo-pilot. The AI inputs were incorporated into the same audio file as the frequency transmissions in the appropriate places right after the planned queries concerning the level of situational awareness. Figure 3 in Section 3.1 shows the full overview of the experimental setup, including the order of the queries and AI inputs in the bottom left corner.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

3.5 Experimental Manipulation

This chapter reports on the verification, if the conditions low vs. high for task load were successfully implemented with the design of the scenario. Workload measures with ISA (Instantaneous Self-Assessment, Section 3.10.1) were assessed several times during each interactive scenario in experiment 1 and 2. To test the experimental manipulation of task load, the mean ISA workload rating was compared with the task load ratings provided by subject matter experts (SMEs). In Table 7 it can be seen that scenarios E1S2, E1S3 and E1S4 have a similar average ISA value. These three scenarios were also rated the most complex by the SMEs. The ISA value of the E1S1 scenario is a bit lower, which is also in agreement with the SMEs' evaluation. The workload of the E1S5 scenario is the highest. This can be attributed to the fact that a lot of capacity was needed to get used to the system. It was the first scenario for all ATCOs. The SMEs estimated the complexity of the E1S5 scenario to be much lower.

Table 7: ISA results for experiment 1 and 2 (N= 20; 16)

	Experiment 1					Experiment 2	
	E1S5	**E1S2**	**E1S3**	**E1S4**	**E1S1**	**E2ST**	**E2S2.1**
Min	2	1	1	2	1	1	2
Max	4	5	5	4	3	5	5
Mean	3.55	2.75	2.72	2.77	2.30	2.69	3.25
SD	0.6	0.88	0.87	0.62	0.66	0.94	0.88
Complexity rated by SMEs	1	4.5	4.5	3	2		

In addition, SMEs were asked to assess the subjects' manipulation skills with ESCAPE Light on a scale from 1 to 5 (low to high skill): In experiment 1 they rated “ESCAPE Light handling skills” after the E1ST scenario and again after the E1S5 scenario. In experiment 2 they were asked to rate after scenario 1 (E2ST). While the SMEs were asked about the handling of the labels and the conflict detection tool in the first experiment, they were asked about the following assessment criteria in experiment 2:

the handling with ESCAPE Light concerning …

- … the speed vectors
- … the labels (in general)
- … the measuring tools (e.g. VERA)
- … the change of the current FL and exit level
- … the transfer
- … the display of the planned route
- … "direct to" inputs

Table 8 shows the results for the manipulation check for “handling skills” to make sure participants were ready for simulation. In experiment 1 skills increased from the training scenario to the first experimental scenario from 2.75 to 3.55 on average. The lowest rating (1) was used by SMEs in both scenarios indicating some ATCOs were still struggling with the simulation tool even when the experiment started. In the training scenario of experiment 2 the handling skills were rated 3.83 on

EUROPEAN PARTNERSHIP
Co-funded by the European Union

average. No ATCO was rated lower than 3 indicating that they had sufficient practice before the experiment started.

Table 8: Manipulation check for subjects' handling skills with ESCAPE Light rated by the SMEs during experiment 1 and 2 (N= 20; 16)

	Experiment 1		Experiment 2
	E1ST	**E1S5**	**E2ST**
Min	1	1	3
Max	4	5	5
Mean	2.76	3.55	3.83
SD	0.70	0.81	0.68

To see if participants with higher scores on handling skills actually felt less workload the SMEs' assessment was compared with the subjects' self-assessment (ISA). A comparison of the SMEs rating regarding the label handling and the subjects' ISA in the light scenario (E1S5, experiment 1) and the scenario 1 (E2ST, experiment 2) is seen in Figure 4.

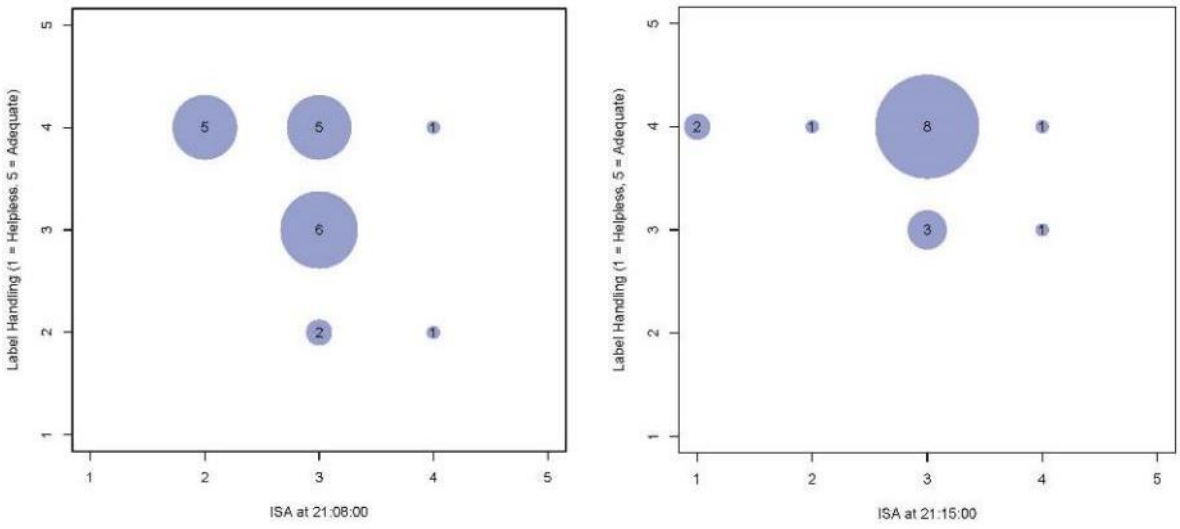


Figure 4: Comparison of subjective workload (ISA) and SME rating for ESCAPE Light handling skill "label handling" in experiment 1 and 2 (N= 20; 16)

Both graphs in Figure 4 compare the workload perception of the participants with their label handling skills assessed by the SMEs. The left plot shows the comparison in experiment 1 while the right plot compares the assessments in experiment 2. In experiment 1, the label handling capabilities of 6 ATCOs were rated medium by the SMEs. These 6 ATCOs also rated their workload assessment as medium. In experiment 2, 12 ATCOs were assessed with a label handling capability of 4 out of a maximum of 5 by the SMEs. The 12 ATCOs assessed their workload between 1 and 4 out of a maximum of 5. Out of these 12 ATCOs, 8 assessed their workload as medium (3). This suggests that in these cases there is a correlation between the ATCOs' workload assessment and the SMEs' handling assessments.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

In conclusion, it can be said that the participants became increasingly familiar with the ESCAPE Light simulation software over time and accordingly improved their handling skills. This can be seen because the SMEs assessed the handling skills after the E1S5 scenario in experiment 1 better than after the E1ST scenario. In experiment 2, the average rating of handling skills by SMEs was slightly higher than in experiment 1 (Table 8). One possible reason is that participants (N = 3) had already taken part in experiment 1 before experiment 2 and were therefore more experienced in the use of ESCAPE Light.

Nevertheless, it should not be neglected that the unfamiliar work method with the ESCAPE Light simulation software can have a major influence on the results. At least this is what the questionnaire results of the debriefing after the scenarios suggest. Figure 5 shows that the mental capacity needed to work with ESCAPE Light was considerable according to the subjective assessment of the subjects.

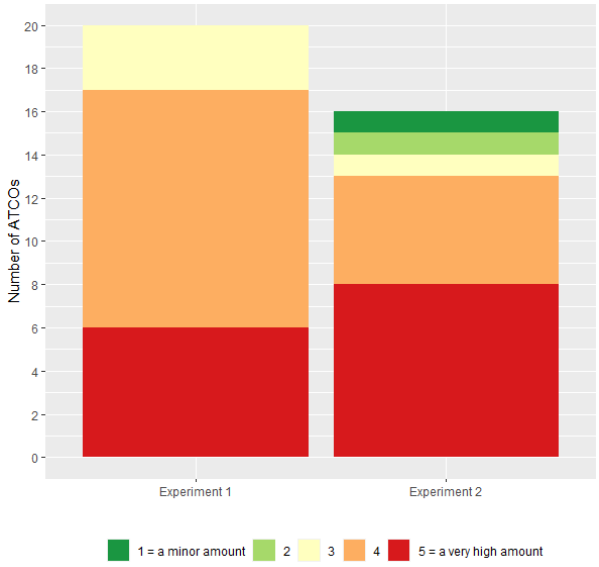


Figure 5: Mental capacity absorbed to handle new system (ESCAPE Light) (N= 20; 16)

And Figure 6 shows ATCOs' signs of discomfort with the simulation tool: A minority of the ATCOs felt comfortable with ESCAPE Light. Half of the ATCOs indicated discomfort, some even at a very high level.

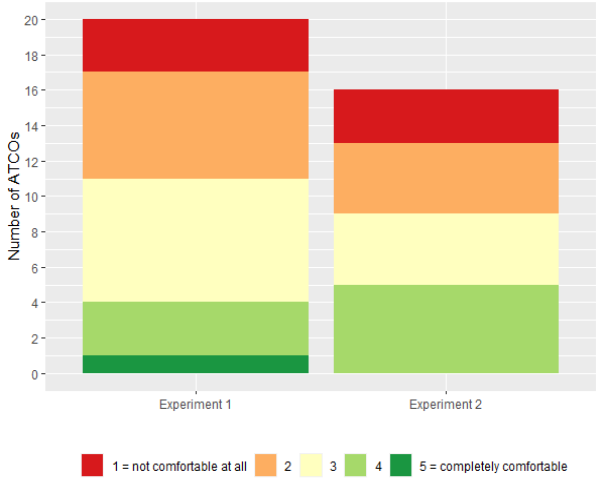


Figure 6: How comfortable did you feel with the ESCAPE Light simulation software? (N= 20; 16)

EUROPEAN PARTNERSHIP
Co-funded by the European Union

From the analysis of the handling skills, it may be concluded that most ATCOs were capable to handle ESCAPE Light but at considerable costs in terms of subjectively felt absorption of mental capacity. It can be assumed that the unfamiliarity with the tool came along with side effects such as a lack of automatised routine skills in using the tool's functionalities (e.g., measuring equipment) and an increased need to correct in case of operating errors.

To exclude a systematic influence of learning and fatigue in the experiment, the scenarios were presented in randomised order (Section 3.3).

3.6 Participants

ATCOs from en-route sector in Swiss airspace (LSAZM567) were asked for voluntary participation in the two experiments. The experiments were scheduled during ATCOs' work time. Participants were previously informed about the experiment via the internal information system by their supervisor.

ATCOs were independently sampled for experiments 1 and 2. By chance, three ATCOs participated in both experiments. Altogether, data was gathered from 33 different ATCOs.

Table 9: Descriptive analysis for participants in experiment 1

	Quantity	Av. Age [years]	Av. Working Experience [years]
Female	5 (25%)	41.40	16.00
Male	15 (75%)	43.20	18.30
Total	20	42.75	17.75
SD		8.11	7.50

In experiment 1, 20 ATCOs took part (5 women and 15 men) with an average age of 42.75 years (SD = 8.11), ranging from 22 to 55 years. The average work experience was 17.75 years (SD = 7.5), ranging from 0 to 28 years.

Since only 3 probands had less than 10 years of professional experience, the sample from experiment 1 consists mainly of experienced ATCOs. While 9 subjects work exclusively as ATCOs, 11 have additional functions: 6 supervisors, 3 training instructors and 2 experts (e.g., domain managers). Apart from one ATCO, the subjects had never worked with the ESCAPE Light simulation software before experiment 1.

Table 10: Descriptive analysis for participants in experiment 2

	Quantity	Av. Age [years]	Av. Working Experience [years]
Male & Female	15 (93.75%) & 1	40.38	17.31
SD		7.73	7.77

The group in experiment 2 is on average slightly younger than in experiment 1. As in experiment 1, most of the probands are experienced ATCOs. A third of the participants (N = 6) work exclusively as ATCOs, 4 also as supervisors and 6 have other additional functions within Skyguide. Three participants had already taken part in experiment 1 and have therefore gained previous experience with the

EUROPEAN PARTNERSHIP
Co-funded by the European Union

ESCAPE Light simulation tool. Most of the participants (13 ATCOs) had never worked with ESCAPE Light before experiment 2.

No monetary compensation for participation was provided to subjects. Instead, they received a symbolic present at the end of the experiment.

3.7 Measurement of Artificial Situational Awareness

In experiment 1, the ATCO participants were asked to complete a total of 6 exercises on an ATC simulator which were recorded and used as a base for experiment 2 that followed (as explained in Section 3.1). The data from experiment 1 was also used to compare human and machine situational awareness. With 20 different ATCOs controlling the air traffic, each in their own way, plenty of information was gathered so that human situational awareness could be analysed and objectively measured against the outputs generated by the AI SA KG system (later referred to as „the KG system“). It is important to emphasise that the Skyguide ATCOs were not working on a system they are used to. They had to adapt quickly to a simulator they never used before and work with only the basic controller tools, whereas in their day-to-day job they can rely on a wide range of different safety nets. Those circumstances greatly increased their workload. Any recorded loss of situational awareness will therefore **not be categorised as a controller's mistake** and **cannot be an indicator of ATCO performance**, since it is very unlikely it would have happened in real traffic. Henceforth, loss of situational awareness of the participants on the ESCAPE Light system in experiment 1 will be referred to as „degraded human situational awareness “.

There are three approaches to the measurement of artificial situation awareness presented below.

3.7.1 Knowledge Graph and Task Analysis

In this stage of the analysis, only AI SA tasks not dealing with future conflict predictions are taken into account for each defined timestamp and situational awareness is formed based solely on the currently observed situation and historic data. For example, all the outputs corresponding to timestamp 12:00:00 in the high scenario will describe the current state of the aircraft in the airspace. At that moment, their position or status at 12:00:15 are not predicted. A possible output regarding a climbing aircraft is „Aircraft is climbing towards its cleared flight level “, because in that moment, it can be seen that it has a positive rate of climb, its current flight level is higher than the previous and lower than the cleared flight level. While the system is generating outputs for future prediction related tasks, they are not analysed in this section.

To be able to objectively measure situational awareness in both human and KG system, a list of situational awareness indicators was made that are unambiguous and easy to assess. For certain situations, a time buffer was introduced for the ATCO, as it cannot be claimed that their situational awareness has degraded if they did not react instantaneously as a computer does. Those buffer values were negotiated with the SME. A wrong output by the KG system that does not correctly describe the traffic situation would also be considered as degraded situational awareness, however, all outputs regarding the recorded scenarios in experiment 1 have been tested and improved to avoid that situation. The objective situational awareness indicators are further explained below and summarised in Table 11.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Check if aircraft is transferred on time

The KG system has the information about the coordinates of the sector boundary, stored in the static data graph of each scenario. By calculating the distance between the current aircraft position and the sector boundary, it is possible to use the aircraft's current speed to calculate the time needed to reach the boundary. The KG system will send an output 2 minutes before the aircraft reaches the sector boundary, reminding the ATCO that the aircraft needs to be transferred. Additionally, the KG system checks whether the aircraft is cleared to its crossing flight level by the time it has been transferred.

For the machine, degraded situational awareness is, therefore, failure to notify the ATCO 2 minutes prior to being transferred and/or never checking whether the aircraft is reaching or maintaining its crossing flight level. For the human, it is considered that their situational awareness has been degraded if the aircraft crosses the boundary without being transferred, and/or not clearing aircraft to its crossing flight level before transferring it. In case the ATCO transfers the aircraft earlier than 2 minutes before the boundary and the flight is cleared to or maintaining the crossing flight level, it is considered that neither human nor machine have suffered degradation of situational awareness. A case where the aircraft is transferred on time but not at the appropriate flight level is degradation of human situational awareness. If AI SA output indicated the necessary flight level change, machine situational awareness is not degraded. Table 11 below summarises the situations where situational awareness is considered to be degraded regarding this indicator.

Check for reaction to non-compliance

After issuing a clearance to the pilot, the ATCO is responsible for monitoring whether the pilot has reacted appropriately. Usually, a readback by the pilot is expected, but it is still possible that an error occurs even if the readback is correct. For example, the pilot can put in a wrong value for the flight level, resulting in a flight level bust. In experiment 1, two non-compliances by the pilot have been intentionally included in the simulation exercises. More precisely, in E1S2, there is a flight level bust. In the E1S3, there is a speed non-conformance. Subsequent analysis showed that there have been unintentional non-conformances regarding the cleared rate of climb/rate of descent (ROC/ROD) as well.

For the KG system, it is expected to generate an output indicating the non-conformance as soon as it happens, i.e., in the first timestamp after the clearance was issued and the pilot failed to respond or responded the wrong way. For the controllers, a time buffer of 30 seconds from the moment when the non-compliance is noticeable on the radar screen was introduced.

Non-conformances fall into the following categories:

- Heading/route non-conformance
- Flight level non-conformance
- Speed non-conformance
- ROC/ROD non-conformance

There have been no cases of heading/route non-compliances in the analysed scenarios. After consultation with the SME from Skyguide, it was decided that introducing that kind of non-compliances might cause unwanted conflicts and further increase ATCO workload. However, the KG system does include tasks which can notice deviation from the cleared route, which were tested aside from the human-in-the-loop experiments to see if they give expected results. It considers the known current

EUROPEAN PARTNERSHIP
Co-funded by the European Union

track of the aircraft and the expected track resulting from either cleared point, heading instruction or the flight plan. A comparison between those values shows whether the aircraft is following the cleared or the planned route in the first timestamp after the clearance is issued. For the KG system, if the expected output was missing or wrong, it would imply degradation of situational awareness. The ATCO would have a buffer of 30 second to notice the deviation before they are considered to have degradation of situational awareness

The planned flight level non-conformance was always initiated by the pseudo-pilots when they asked for a level change and then put in the wrong value when the level change was approved. The cleared flight level is FL390 in that case, and the pseudo-pilots would climb the aircraft to FL395. The KG system is expected to give the output „Flight level bust “as soon as the aircraft passes its cleared flight level and continues to climb/descend. The human situational awareness is considered to be degraded if there is no corrective action by the ATCO within 30 seconds from the moment the aircraft levels off at the wrong flight level.

The speed non-compliance initiated in the E1S3 depends on the ATCO input. To clarify, there are two aircraft flying on the same flight level and the same route, entering the airspace with 8 NM between them. The following aircraft is maintaining speed M.80, while the leading aircraft's speed is M.78, resulting in their distance gradually decreasing and them being in conflict by the time they exit the sector if no action by the ATCO is initiated. In some scenarios, the controllers notice the speed difference and issue speed clearances to one or both aircraft to ensure their distance does not continue to decrease. In that case, the pilot does not comply with the clearance. If the ATCO did not notice the non-compliance within 30 seconds, it is considered that the human situational awareness has been degraded. As for machine situational awareness, it is expected to receive an input indicating that the aircraft is not changing its speed in accordance with the clearance in the first timestamp after the clearance has been issued.

The ROC/ROD non-conformances were not planned as an intentional part of the experiments that was meant to be analysed, but a few cases did happen as a result of pseudo-pilot error. The information the KG system uses are cleared ROC/ROD, current ROC/ROD, and previous ROC/ROD. A simple comparison shows whether the aircraft is maintaining the cleared rate, accelerating/decelerating towards it or not maintaining the cleared rate. Machine situational awareness is degraded when the KG system fails to provide an output or gives a wrong result. Human situational awareness is considered to be degraded when the ATCO does not notice the non-compliance within 30 seconds of the pilot's error.

Check if flight is assumed on the label

Prior to entering the sector, pilots usually make an initial call on the frequency of that sector, and the ATCO can identify them both on the frequency and on the radar label, therefore assuming the responsibility for that aircraft. If the flight enters the sector without making the initial call, controllers will try to reach them to establish control. In this analysis, one of the indicators of situation awareness is whether an aircraft has been properly assumed, i.e., whether the controller has assumed the flight on the label after having received the initial call.

The KG system gathers information about when the flight makes the initial call and when it is assumed on the label. If the flight has sent the initial call but has not been assumed, the KG system should have the appropriate output in the first timestamp after the initial call, otherwise the machine situational awareness is considered to be degraded. The degradation of human situational awareness happens

EUROPEAN PARTNERSHIP
Co-funded by the European Union

only if the ATCO did not assume the flight on the label within 30 seconds after having received the initial call and responded to it on the frequency. The error might occur due to high workload at that moment or confusion due to similar callsigns, when the ATCO assumes an aircraft that did not yet make the initial call.

Check if aircraft will fly through restricted airspace and if they can use previously restricted airspace

There are TRAs used for military purposes in the Zürich upper airspace. In the military scenario, the central military area is active in the beginning and then deactivates in a few minutes. Later, the previously available eastern military area becomes active and therefore unavailable for commercial flights. The coordinates of the military areas as well as the activation times are written in the KG and with information on current position and planned trajectory, the KG system is able to compute whether the planned trajectory crosses the restricted area and when it is safe to use the previously restricted area.

In the beginning of the scenario, the KG system identifies aircraft that can fly direct routes through the previously unavailable military airspace. If it fails to list all aircraft that can use the airspace, the machine situational awareness is considered to be partially lost. The ATCO, however, is not expected to issue direct-to clearances for all aircraft that can use the previously restricted airspace since it may not always be in line with their strategy for managing the traffic, i.e., ATCO situational awareness will not be considered degraded should they simply let the aircraft fly on their planned routes.

In the second part of the scenario, the eastern military becomes active. Since the KG holds information about the activation time in advance, it can indicate which planned trajectories are crossing the military airspace as soon as those flights first appear in the KG. Degradation of machine situational awareness is a situation where one of the flight whose trajectory passes the restricted airspace is not recognised. Degradation of human situational awareness is when the ATCO does not vector the flight and it ends up entering the restricted area.

Table 11: Summary of objective situation awareness indicators and the conditions for comparison between the machine and human

Objective Situation Awareness Indicators	Degraded Machine SA	Degraded Human SA
CHECK IF AIRCRAFT IS TRANSFERRED ON TIME	There is no „Transfer aircraft“ output 2 minutes before boundary.	The aircraft is not transferred before the sector boundary.
	There is no output that the aircraft is not cleared to its XFL before the boundary.	The aircraft is not cleared to its XFL before being transferred (if XFL is different from actual flight level at the time of sector entry).
CHECK FOR REACTION TO NON-COMPLIANCE	There is no output indicating the deviation from the cleared or planned route in the timestamp after the clearance has been issued or immediately, in case of deviation from the planned route.	There is no corrective action by the ATCO within 30 seconds from the non-compliance.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

	There is no „Flight level bust“ output in the timestamp after the cleared flight level has been passed.	
	There is no output indicating that the aircraft is not at cleared speed or accelerating/decelerating towards cleared speed in the timestamp after the clearance.	
	There is no output indicating the wrong ROC/ROD in the timestamp after the clearance.	
CHECK IF FLIGHT IS ASSUMED ON THE LABEL	There is no output indicating that the aircraft has made the initial call but was not assumed on the label in the timestamp after the initial call.	The flight is not assumed on the label within 30 seconds of the initial call or the wrong flight has been assumed.
CHECK IF AIRCRAFT WILL FLY THROUGH RESTRICTED AIRSPACE AND IF THEY CAN USE PREVIOUSLY RESTRICTED AIRSPACE	There is no output indicating that the aircraft can now use the previously restricted airspace.	The flight has entered the restricted airspace.

Differences of KG tasks at stage I and II

During experiment 2, only seven KG system tasks had been presented (task 1.1, 1.8, 4.2, 5.1, 6.2, 8.1, 8.2 of Table 2). At the time of writing, the KG were adapted to 46 tasks, which have been tested and fully implemented (compare Table 2). For instance, false recognition of the aircraft non-compliances (Flight level bust) occurred when aircraft would still have a non-zero vertical rate just after reaching cleared flight level. The marginal situations followed by some changes in the state of the aircraft were often the reason for faulty KG system outputs. KG system task which checks if aircraft will enter restricted (military) airspace had also been improved. Times of military sector activation and a list of the exit points that are placed inside the military sector had been added. This change made it possible to accurately predict whether the aircraft would be in the military sector even though the sector of interest is not currently active. Also, calculating whether the planned trajectory of the aircraft cleared on a heading crosses the military sector has also been improved. To reduce the number of false non-compliances, a buffer of 4 kts for cleared speed and 2.5 NM for deviation from the cleared route was added. Newly added KG system tasks and presented improvements reduced the number of errors, increased the situational awareness of the system, and thus increased the number of situations monitored.

The KG system will need more work before being used in an actual ATC department. For this reason, not all functions are available yet, as in the conflict detection system that Skyguide currently uses.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

3.7.2 Analysis of Conflict Detection ML Module Predictions Regarding Situations of Interest

As mentioned in 1.2.1.3, three ML modules have been produced from which the conflict detection module is the most frequently used. The output data was analysed to check the accuracy and applicability of the conflict detection module output data. Not every conflict detection ML module prediction was analysed at this stage, but rather the initial and the final prediction for each aircraft pair which had a predicted minimum distance of less than 25 NM.

The initial prediction is made at the first timestamp when the conflict detection ML module provides an output about a certain aircraft pair, provided the ATCO did not previously issue any instructions to either of the aircraft. The information taken into account is the predicted minimum distance and predicted time to minimum distance from the moment the prediction is made. For those same aircraft pairs, actual minimum distance and time to minimum distance are measured in the simulator to be able to precisely compare the predicted and actual values.

Similarly, the final prediction is made for each aircraft immediately after the last ATCO instruction for both aircraft in the observed aircraft pair. In case there were no ATCO instructions for the aircraft in that aircraft pair, the initial and final prediction will match. The actual minimum distance and time is measured and compared to the predicted values.

In addition to the initial and final predictions being compared to the actual distances and time in the scenario, the statistical data described in 1.2.1.3 is added for each aircraft in analysed aircraft pairs. This information is then used to try to find the correlation between the data used for training the machine learning module to the accuracy of the predicted minimum distances.

Although the conflict detection module recognises all aircraft pairs and provides a prediction for them if their minimum distance will be less than 25 NM, it is important to distinguish what is for ATCO considered as a situation of interest and what is a conflict.

According to Metzger et al. (2001), every situation where the minimum distance between aircraft is less than double the lateral norm (5 NM) should require ATCO's attention. Therefore, while analysing initial and final minimum distance predictions 10 NM is set as a limit for a predicted crossing to be a situation of interest.

Four different outcomes of the minimum distance analysis are possible:

- Initial prediction and actual prediction less than 10 NM,
- Initial prediction and actual prediction more than 10 NM,
- Initial prediction less than 10 NM and actual prediction more than 10 NM,
- Initial prediction more than 10 NM and actual prediction less than 10 NM.

Based on these result outcomes, it is possible to identify and count Type I Error and Type II Error as well as the rest of the result distribution. Later in Section 4.2.3 these results are presented and elaborated.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

3.7.3 Analysis of Conflict Detection ML Module Predictions Regarding Conflicts

In the previous section, situations of interest are analysed when comparing conflict detection module predictions. It is also possible to analyse the prediction of the module for aircraft pairs that would cause a violation of the declared separation minima without any ATCO action. The analysis of the conflict detection module prediction for aircraft pairs in conflict explains how the module predictions behave for these aircraft pairs i.e., whether the prediction corresponds to the actual minimum distance and whether the predicted distance increases after the ATCO resolves the conflict. Aircraft pairs that would cause a violation of the declared separation minima without any ATCO action cause a conflict. Inside each scenario, conflicts were implemented deliberately. Additionally, some conflicts have arisen through the action of ATCO. In the human-in-the-loop experiments, if the ATCO resolved a conflict by separating aircraft vertically, that situation could not be used for this analysis because the CD module cannot provide an output for vertically separated aircraft. By this analysis, a comparison of the module performance for aircraft pairs that would have minimum distance less than separation minima with the performance of the ATCO is enabled.

3.8 Measurement of ATCO Situation Awareness

In experiment 1 three different techniques were used to measure the ATCOs' situation awareness: subjective rating, gaze analysis with eye-tracking to retrace ATCOs' visual attention and implicit performance measures from behavioural coding for radio calls and actions when interacting with pilots. A fourth technique was added in experiment 2: probe technique. They are described below.

3.8.1 Subjective Rating for Situation Awareness

After each scenario, subjects were asked to complete the SASHA_Q questionnaire (Dehn, 2008) for subjective situation awareness. This consists of six questions with behavioural descriptions for situation awareness aspects that were rated on a dimension from 0 (never) to 6 (always). E.g., "I was ahead of the aircraft." (See A.1 for the full questionnaire.)

3.8.2 Gaze-Based Analysis of Situation Awareness

Eye tracking data were recorded with Tobii Pro Glasses 3 (100 Hz sampling rate) and processed using Tobii Pro Lab and a Computer Vision Tool. The two programmes and their benefits are described in the next chapters.

3.8.2.1 ET Data Processing

The raw ET data is processed for later analysis with Tobii Pro Lab software to export ET data such as ET coordinates, pupil dilation, AOI hits and gaze duration. The software displays the gaze in the ET video (see Figure 7), where the green circle represents the actual gaze, and the green line represents the gaze history. This video was used for behavioural coding in Mangold Interact (see 3.8.3).

EUROPEAN PARTNERSHIP
Co-funded by the European Union

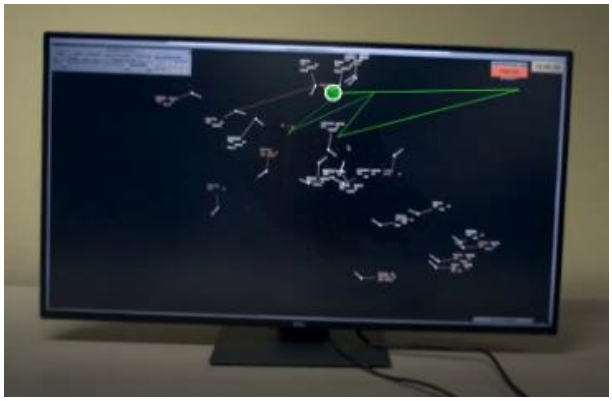


Figure 7: ET Gaze with Gaze History

The gaze coordinates are reproduced on a static image using algorithm for automatic mapping. In the simulation the aircraft are constantly moving and consequently the Areas of Interest (AOIs) on the screen change location. Procedures to improve the accuracy of the mapping are described in Vetter (2022).

The static image was used to define static AOIs (the dynamic AOIs functionality in the software could not be used due to necessary manual corrections). For this purpose, rectangular AOIs of the same size were defined over the entire screen. Only the freeze button and the radar toolbox were explicitly marked (see Figure 8).

Figure 8: Static AOI Creation of Screens

With the software data frames for the pupil diameter of both eyes, the mapped gaze points in x and y direction and pixel coordinates, the gaze event duration, and AOI hits were generated for further analysis such as gaze plots, AOI plots and as input (raw gaze points in x and y direction in pixel coordinates stored every 10 milliseconds) for the CVT.

3.8.2.2 Computer Vision Tool for Dynamic AOIs

For several reasons (head movements etc.), automated dynamic AOI mapping from Tobii Pro Lab could not be used. A tailored solution using computer vision was designed by an employee of the Microsoft Robotic Lab. A description of the processing mechanisms for radar screen detection and for the localization of the ATCOS' gaze is described in Appendix B.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

By use of computer vision and other technologies the CVT detects aircraft and identifies when ATCOs were looking at aircraft. Inputs needed are raw gaze coordinates from ET records, log data from the ESCAPE Light simulation to define the exact positions of aircraft and screen recordings.

The tool first searches for the screen in the ET recording using computer vision algorithms. Once the screen has been identified, the gaze in the ET coordinate system is transformed into the coordinate system of the screen. The positions of the aircraft on the screen are then identified with ESCAPE Light simulation logs. Image to text algorithms identify the labels corresponding to aircraft. With this the CVT tool recognizes when aircraft and aircraft labels are looked at by ATCOs. The exact process is described in the Appendix B.

CVT is an efficient tool but has limits: The screen is not always correctly identified. ATCO head movements require the tool to constantly search for the screen and adjust the coordinate system. This leads to errors and inaccuracies. Background noise in the video impairs the algorithm from identifying the screen. For example, in some cases, another screen of similar size was in the background. If the screen is not recognised correctly, the gaze in the screen recording does not match the real gaze. For this reason, a confidence value was included. The tool estimates its confidence in a range between 0 and 1, where 1 is best. Thus, the correctness of the tool can be estimated, and detections can be checked. Confidence was used to exclude data from analysis.

3.8.3 Implicit Performance Measurement for Situation Awareness

For implicit assessments of air traffic controllers' situation awareness behavioural codes for radio communication were generated in Mangold Interact. Based on these codings different data frames were created to capture the way how ATCOs work (e.g., how many transfer calls they transmitted or at what time specific events were performed). These are described in the sections from Appendix D to Appendix I.

3.8.4 Probe Technique for Situation Awareness

To measure situation awareness aspects during the scenarios queries were asked in accordance to SASHA_L (Dehn, 2008) in experiment 2. The content of the queries was created by subject matter experts (SMEs). On average five queries were asked per scenario. Most of the queries were directed to conflicts and verified which conflicts were detected by ATCOs. For this purpose, questions were allocated at predefined sections of the scenarios. "Trick questions" were included to prevent priming ATCOs' expectations. ATCOs were previously informed about all types of questions used. The exact queries are listed in Appendix A.2.

AI SA was prompted with SPARQL queries of identical content at the exact same time using data from experiment 1. This allowed comparison of machine and human situation awareness and to provided AI SA inputs for human-machine team situation awareness in experiment 2.

After each query, simulation was stopped for 20 seconds, giving the ATCO enough time to review the screen and respond. The pause was extended to 30 seconds in the military scenario (E2S4.1 and E2S4.2) because more information had to be communicated. ATCO's responses to queries were recorded and analysed. Several answers (conflict pairs) were possible for each conflict query.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

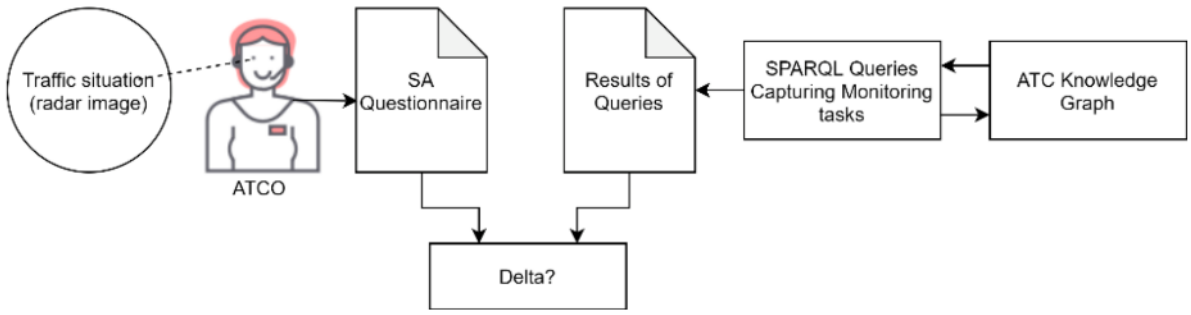


Figure 9: Comparison of human and system SA

How AI SA system proceeds to generate output is described in Section 1.2.1.5. More information about the algorithm is described in the concept of operations (see D2.1), the proof of concept for KG system (see D4.1 (AISA Consortium, 2021g)) and the final report.

3.8.5 Scale Scores for ATCO Situation Awareness

To interrelate different methods for measuring situation awareness and analyse consistency Pearson correlations (Benesty et al., 2009) were calculated. Scale scores were calculated for each measurement method.

SASHA_Q (described in Section 3.8.1): An overall score is calculated for ratings of the six questions according to the procedure in Figure 10. A high score corresponds to a positive self-assessment of situation awareness.

Item	Item score		
1) … I was ahead of the traffic.	y =	x	
2)* … I started to focus on a single problem or a specific area of the sector.	y =	6 - x	
3)* … there was a risk of forgetting something important (like transferring an a/c on time or communicating a change to an adjacent sector).	y =	6 - x	
4) … I was able to plan and organise my work as I wanted.	y =	x	
5)* … I was surprised by an event I did not expect (like an a/c call).	y =	6 - x	
6)* … I had to search for an item of information.	y =	6 - x	
OVERALL SCORE (Σ)			
MEAN (Σ/6)			

Figure 10: Scoring key for SASHA_Q

SASHA_L (described in Section 2.1.2): The numbers of queries varies across scenarios and the possible number of answers to them, too. To avoid meaningless answers, only those queries were taken into account to which a clear answer can be given. Therefore, vague queries such as "What do you need to pay attention to?" were not evaluated and trick questions were excluded, too. In the E2S2.1 scenario, the first query is not scored because the question was asked too early. Thus, a total of 82 answers to the queries were included in the calculation of an overall score. The percentage of correct answers is calculated for each ATCO. A high score corresponds to many correct SASHA_L answers.

Eye tracking (described in Section 3.8.2): Accumulated time (minutes) to detect a conflict was used. A low score indicates a good situation awareness.

Implicit performance measurements (described in Section 3.8.3): The overall score combines three aspects: Whether the ATCO recognised the conflict at all (aspect 1), when the resolution of the conflict was initiated (aspect 2), and the duration of the conflict (aspect 3). How the scores of aspects 1 and 2 were calculated is shown in Equation 1 and

Equation 2. The mean value of the conflict was subtracted from the individual ATCO's value. The differences were summed up across all conflicts and divided by the average mean value of all conflicts. Aspect 3 is evaluated as follows: the ATCO gets one plus point for each conflict detected, one minus point if the conflict was not detected and no point if the conflict did not occur. The values of these three aspects are summed up in an overall score that expresses good conflict solution, if the overall score is low.

Equation 1: Score calculation of implicit performance measurement aspect 1

$$\frac{\left(\sum_{n=1}^{\#\,conflicts} x_{duration,conflict} - \bar{x}_{duration,conflict}\right)}{\bar{\bar{x}}_{duration,all\ conflicts}}$$

Equation 2: Score calculation of implicit performance measurement aspect 2

$$\frac{\left(\sum_{n=1}^{\#\,conflicts} x_{startTime,conflict} - \bar{x}_{startTime,conflict}\right)}{\bar{\bar{x}}_{startTime,all\ conflicts}}$$

An example data frames containing the scores per ATCO for experiment 1 is outlined in Appendix J.

3.9 Measurement of Performance and Control Strategies

Performance can be interpreted in terms of safety, efficiency, and orderliness (Griffin et al. 2000). Behavioural coding was used to observe ATCOs performance and control strategies.

3.9.1 Behavioural Analysis of the ATCO

To analyse the performance of an ATCO more objectively, discrete behaviours were observed combining simultaneously data from ET recording, screen recording, audio on communication between pseudo pilot and ATCO and inputs to the simulation to avoid misinterpreted or loss of information during the analysis. For this purpose, the Mangold Interact software was used.

Clear signals that are visible in all recordings were used to synchronise the recordings: Eye tracking video was tuned to frontal video recording based on light signals or hand movements. The screen recording was matched to the eye tracking recording, as the screen can be seen in both recordings. Biometrical data were added with a sampling rate of 256 Hz.

In Mangold Interact a sampling rate of 32 Hz was chosen as a trade-off to accuracy and computational requirements for the software. This will provide a rough impression on psychophysiological reactions. If higher accuracy is desired, the sampling rate can be adjusted accordingly for later analyses.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

After the recordings have all been synchronised, the individual events performed by the ATCO and the conflicts that occur were coded. This allows to recognize preferences in working style of individual ATCOs and comparison of conflict detection of ATCOs and AI SA system.

The events include all communication with the pseudo pilots (initial and assume calls, directs, transfers, flight level, heading and speed changes), measurements (usage of VERA tool and speed vectors) and other events (activation and deactivation of military areas, STCA warning, adapting label settings). Problems occurring during the simulation were recorded, too: e.g., failure of the screen, when the ATCO had problems with labels for a longer period of time, or when he discussed issues with the subject matter expert.

The occurrence and duration of the events were coded in real time. Only the initial call was timed differently. This event started with the call from the pseudo-pilot and lasted until the ATCO had identified the aircraft. This became apparent by the ET recording and the mouse movement.

Interrater reliability training included sessions with all raters and a subject matter expert with videos from each scenario to define the start and end of conflicts in a uniform way. The coding of the conflicts always started as soon as the conflicts were visible on the radar or as soon as they took place and ended when the ATCO recognised the conflict. In some cases, the ATCOs did not detect all conflicts. Then, the conflict is terminated in the coding at the time when the aircraft was transferred. It must be considered that the ATCO may have detected the conflict earlier than it was coded. However, detection of conflicts often required the use of assistance tools (e.g., VERA to measure the distance). So, most of the time, when a conflict was detected a first reaction was to measure the distance. Therefore, the use of VERA was taken as an indication for conflict detection.

For each conflict, subject matter experts name the most obvious solutions. Thus, for each conflict in each scenario, the corresponding solution of the conflict could be coded for each ATCO. For the solutions, the duration of the event was not of great importance. It was relevant when the solution was applied. There are four different approaches for solving the conflict:

- Level change: solving the conflict by changing the FL
- Speed change: solving the conflict by adapting the speed
- Direct: solving the conflict by changing the HDG or applying a direct
- Confirm: solving a non-conformance by contacting the pilot again
- No solution needed: when the conflict was solved, but it was not possible to identify the crucial event
- No solution: when the ATCO did not solve the conflict

3.9.2 Processing Behavioural Observation Data in R

For implicit assessments of ATCOs' situation awareness, behavioural codes for radio communication were used in observation. Based on the codings, different data frames were created that capture the ATCO's work styles. These are described below. All these data frames take into account different events and solution approaches. The generated data frames are called:

EUROPEAN PARTNERSHIP
Co-funded by the European Union

- The *counter data frame* counts how often certain events (e.g., transfer, assume, speed vector change, etc.) occurred. Based on this data frame, it is possible to roughly estimate the mental workload of the ATCOs (see Appendix D)
- The *conflict comparison data frame* determines the duration of the conflicts and the solution. It is used to determine if ATCOs had identified conflicts, how fast they were (quick/slow) and which solutions did not meet the standard (see Appendix E)
- The *reaction times data frame* assesses how fast ATCOs react to initial calls. When they have detected the aircraft upon the pseudo-pilots' first calls to Swiss radar (see Appendix F).
- The *checkbox data frame* assessed ATCOs' commands for each aircraft on the basis of code words (e.g., assume, transfer, speed, HDG, direct, climb, and descent). Whenever one of these terms was used in communication with an aircraft, it was coded accordingly. To compare ATCOs only aspects were included that needed to have occurred (see Appendix G).
- The *times comparison data frame* extends the checkbox data frame. Each checkmark was replaced by the time stamp when the event occurred. This provides indication on how soon or late ATCOs deal with the events. Especially the time when the transfer call was made is interesting because it can be deduced how long the aircraft was on the frequency (see Appendix H).
- The *number of events per call data frame* counts how many commands were given to the pseudo pilot in one single radio call. ATCOs with preserved situation awareness are expected to convey many commands in a single call for efficiency reasons - e.g., a direct command in the initial call (see Appendix I).
- The *number of conflict solutions data frame* counts how many solutions were applied until a conflict was resolved. ATCOs with preserved situation awareness are expected to apply few steps to solve conflicts (see Appendix I).

Based on these data frames, commonalities and differences in ATCOs actions could be analysed.

3.10 Workload Measurement

Workload was measures by subjective ratings and using psychophysiological parameters.

3.10.1 Subjective Rating

The Instantaneous Self-Assessment (ISA) measures the subjective workload using a single question that needs to be rated. ISA was developed to assess mental workload in ATCOs (Shahid et al., 2012). It provides an instantaneous grading of the workload by participants on a scale from 1 to 5 (see Figure 11).

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Level	Workload	Spare Capacity	Description
1	Under-utilised	Very much	Little or nothing to do. Rather boring
2	Relaxed	Ample	More time than necessary to complete the tasks. Time passes slowly.
3	Comfortable	Some	The controller has enough work to keep him/her stimulated. All tasks are under control.
4	High	Very little	Certain non-essential tasks are postponed. Could not work at this level very long. Controller is working 'at the limit'. Time passes quickly.
5	Excessive	None	Some tasks are not completed. The controller is overloaded and does not feel in control.

Figure 11: ISA workload rating

Participants were familiarised with the workload question before the experiment started. A leaflet explaining the different workload levels laid next to them during the experiment. The workload questions were asked verbally during phases without radio communication between the ATCOs and the pseudo-pilot. The questions were asked several times during different phases of each interactive scenario. These phases were by subject matter experts in the planning of the experiments.

3.10.2 Biometrical Analysis

For event-based analysis of workload with biometric parameters, a total of eight specific events were selected with the help of subject matter experts. Those events represent a subset of all events that were analysed for situation awareness. Table 12 describes these events:

Table 12: Definition of events in the scenarios of experiment 1

Event Denotation	Scenario	Description
Non-conformance	E1S1	Non-conformance from SWR516 by not starting to descent after ATCO issues the clearance
Quality of Service	E1S4	Military training airspace Centre is available and thus, aircraft can get a direct-to through previously restricted airspace
Deactivation MIL EAST	E1S4	Military training airspace gets deactivated for civil aircraft and followingly they need to be redirected.
Exit Crossing	E1S4	Exit crossing between the IBK36FS and the EJU67NL.
Crossing	E1S4	Crossing between BTI8EP and FPO85J.
Crossing	E1S2	Consists of three crossings. First between FPO85J and BTI8EP and later between TVF4740, IBK1CH and the AFL2548.
Speed Bust	E1S3	Speed bust of the TOM4BA which is behind the VPBVV.
Crossing	E1S3	Crossing between the THY4CL and the IBK5VZ.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Except for the non-conformance event which did not occur in the simulation of five participants due to the interactive nature of experiment 1 all other event occurred in all simulations for all ATCOs.

Measurement of biometric parameters were taken before, during and after finishing the experiment. This allowed to calculate an individual baseline across all "off-task" phases. With the baseline an individual's absolute and relative change in parameter as a reaction to an event can be determined. All measurement phases which did not belong to a scenario (see Section 3.3) were aggregated in the baseline. The baseline is therefore an averaged value of measurements from a variety of "off-task" situations including the period prior to the simulation, during breaks between the scenarios and the period after the simulation until the debriefing questionnaire was accomplished.

3.10.3 Blood Volume Pulse

Blood volume pulse was measured on a finger with a blood volume pulse sensor from MediTECH using a photo-optical lens to measure pulse signal. The sensor was attached to a finger of the weaker hand. Signals were recorded with ProComp INFINITI (8-channel system) from MediTECH. BioGraph Infiniti software solution was used to calculate heart rate in beats per minute.

3.10.4 Skin Conductance

Electrodermal activity was measured with two skin conductance sensors from MediTECH each attached to one finger of the weaker hand to avoid impairment of mouse handling needed in the simulation. Signals were recorded with ProComp INFINITI (8-channel system) from MediTECH and processed with BioGraph Infiniti software solution. The unit of measurement is Siemens and is often used with the SI-prefix μ because the order of magnitude is a millionth. Since the skin conductance is the reverse value of resistance, an increase in skin conductance is equal to an increase in moisture (i.e., sweat) on the fingers, thus leading to a reduction of the resistance.

3.10.5 Transformation of Biometrical Data for Analysis

The output of the biometric measurements (e.g., beats per minute for blood volume pulse) were processed and analysed with the statistics software R and the console R Studio. Beside graphics for the distribution of raw values, the median (md) for parameters were calculated per ATCO and scenario. The median is less concerned by outliers than the mean and thus is expected to be free of artefacts from body motions. The corresponding measure of dispersion is called interquartile range (IQR) covering values from the 25 percent to the 75 percent quartile. To captures ATCOs' trends per specific scenario the difference of the calculated median for the respective scenario and the off-task baseline was determined. This is necessary because of interindividual differences of body functions. Between subject comparisons can only be made on the basis of relative change to the baseline as a reaction to a specific scenario (see Equation 3 as an example for blood volume pulse).

Equation 3: Relative change of blood volume pulse as a reaction to the specific scenario

$$\Delta_{median_{BVP}}\,[\%] = \frac{md_{BVP_{scenario}} - md_{BVP_{offtask}}}{md_{BVP_{offtask}}} * 100$$

A positive value of change indicates a higher heart rate in the scenario compared to "off-task" baseline.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

3.11 Statistical Analysis

Calculation of parameter and scores, and statistical analyses were performed with R-Studio (version 2022.02.0) and Excel (version 16.0.14326.20900). Data was saved in csv format and read into R for further processing. No special libraries were used. ggplot (*Create a New Ggplot — Ggplot*, n.d.) was imported to create the graphs.

R studio was used for the following tasks: To sort the biometrical data and calculate the median, standard deviation and other parameters (Section 3.10.5), for boxplots (e.g., for biometrical parameters in Section 4.1.4), to extract eye tracking data (Section 3.8.2), to create gaze plots, AOI plots and dwell time plots (Section 4.1.3), for the analysis of behavioral coding to create the data frames described in Section 3.9.2 and evaluated with Excel, and for Pearson correlations (Section 3.8.5) and to compare performance and the workload (Section 4.1.2 and 4.1.4).

EUROPEAN PARTNERSHIP
Co-funded by the European Union

4 Results

The result section is structured according to the research questions (1.3) with some additional chapters with further analyses on the AI situation awareness system. It starts with the topical section on human situation awareness, followed by a comparison of human-machine situation awareness (Section 4.2), and an exploration of the effect of human-machine team situation awareness on ATCO situation awareness and performance (Section 4.3). Results are described for the accuracy of the AI SA system (Section 4.4), for the AI SA system performance (Section 4.5) and the robustness of the AI SA system (Section 4.6). The Experimental Plan can be found in the Appendix L.

4.1 Results on Human Situational Awareness

Experiment 1 provided data to investigate how ATCOs proceed in gaining situational awareness and to compare situation awareness across participants. Data from simulation exercises were used as input for the AI SAS KG system to calculate estimations and make predictions.

In experiment 1 three different techniques were used to measure the ATCOs' situation awareness – a fourth technique was added in experiment 2: subjective rating, behavioural coding for radio calls and actions from interactions with pilots, gaze analysis with eye-tracking to retrace ATCOs' visual attention and probe technique (only in experiment 2). Results on ATCO situation awareness collected with these methods and the consistency across methods are described in the following sections.

4.1.1 Descriptive Results on Situation Awareness Measures

This chapter characterises ATCOs' situation awareness by descriptive analyses for different measures (subjective rating-scales, implicit measurement, gaze analysis). This is followed by a short discussion.

4.1.1.1 Subjective Rating of Situational Awareness in Different Scenarios

At the end of each scenario in experiment 1 and after the interactive scenario in experiment 2 ATCOs subjectively rated their personal situation awareness. Mean scale values were calculated for SASHA_Q (Dehn, 2008) across the six subscales (see 3.8.1). Figure 12 depicts boxplots with mean values per scenario. On average ATCO self-rating scores for situation awareness vary between 20 and 33.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

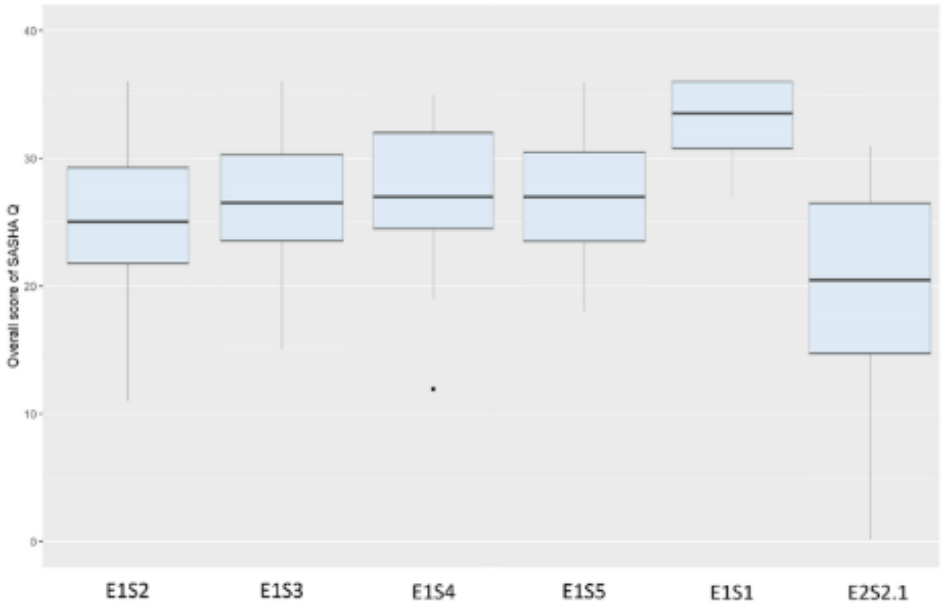


Figure 12: Boxplots for mean scale scores for SASHA_Q for all interactive scenarios in experiment 1 (N=20) and 2 (N=16)

Subjects estimated their situation awareness highest in the scenario E1S1 of experiment 1, a short scenario with 19 aircraft. It was rated lowest in the interactive scenario of experiment 2 (E2S2.1) with 23 aircraft. The mean score for subjective situation awareness varies visibly between the crossing scenario in experiment 1 (E2S2) and the identical scenario in experiment 2 (E2S2.1). This might be due to some differences in the overall setting: ATCOs participating in experiment 2 were asked queries about specific aspects of the situation and received AI SA inputs. This all might have led to a more explicit awareness of aspects that were missed by one's own situation awareness. As it is visible from Figure 12 some ATCOs rated their situation awareness as very low in E2S2.1 (even zero).

Discussion:

On average the subjectively rated situation awareness was similar across the different scenarios. Under conditions that make ATCOs explore and reflect their situation awareness more specifically and with feedback on the correctness and comprehensibility provided by AI SA inputs, ATCOs are more critical in the judgment of their own situation awareness. Provided accuracy of machine situation awareness can be improved, this would offer ATCOs opportunities to validate their situation awareness, to learn about new aspects or aspects they have miss and develop more distinguished mental models. Machine situation awareness might provide a second opinion.

4.1.1.2 Implicit Measurement of Situation Awareness in Different Scenarios

ATCOs' situation awareness is measured implicitly based on behavioural codes for their radio communication (see 3.8.2). This step included counting how often certain events occurred, how conflicts were solved, when aircraft were transferred and other aspects.

Implicit measurement was analysed by the number of conflicts each ATCO had solved in total and how quickly conflicts were recognised. Figure 13 depicts ratio of conflicts solved to unsolved for the most challenging events in scenarios of experiment 1. Of all events analysed, the proportion of the events Quality of Service (QoS) solved by ATCOs was lowest (7 to 10 ATCOs out of 18 ATCOs).

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Non-conformance in radio communication was detected by half of the participants (9 of 18 ATCOs). The exit crossing in the E1S4 scenario was also recognised and solved by only 10 ATCOs. The exit crossing was the only crossing event in the entire experiment 1 that was partially not solved. ATCOs had no problems solving all other crossings. Speed bust was the event solved by most ATCOs (15 out of 18 ATCOs).

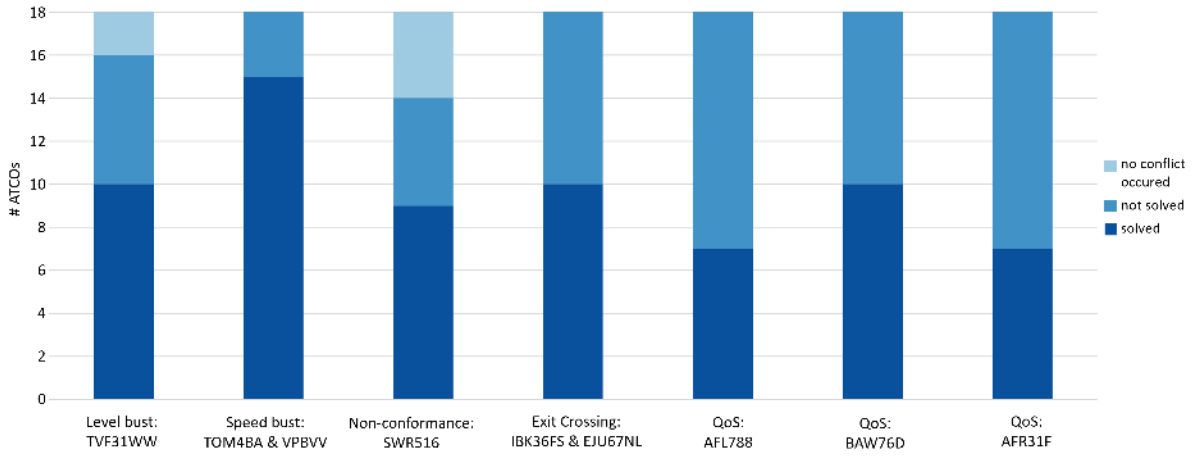


Figure 13: Most challenging events and conflicts from all scenarios from experiment 1 (N=18)*
(* Reduction in N due to missing eye-tracking data for 2 participants; QoS: Quality of Service)

Whereas Quality of Service issues are to some degree at an ATCOs discretion–how much optimisation in routing should be offered for individual aircraft- solving conflicts and detecting non-conformances are more vital for safety.

Discussion:

ATCOs performance on monitoring tasks such as conformance management, quality of service and conflict detection was measured in a simulation experiment with suboptimal work tools for ATCOs. Results show that ATCOs might profit from inputs from a system capable of machine situation awareness. It would need to be able to direct attention to important time-critical aspects, whereas other information might be provided in a more passive manner, leaving ATCOs focused on their current tasks. The misses - especially in the case of non-compliances - could be a side effect of relying on automation tools in daily operation. Future machine situation awareness might support training of situation awareness in regard to aspects most often missed by an individual ATCO or ATCOs in general and provide feedback on these aspects in specific exercises. This can counteract the complacency effect that extinguished redundancy in human-machine team situation awareness.

4.1.1.3 Gaze-Based Analysis for Situation Awareness

How do characteristics of the gaze behaviour influence the performance in detection of conflict? Figure 14 depicts a scatterplot for duration for mean conflict detection and mean dwell time in scenario E1S2. Most of the 9 ATCOs analysed with eye-tracking had a mean dwell time around 4 minutes per aircraft aggregated from multiple gazes on that aircraft throughout the scenario.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

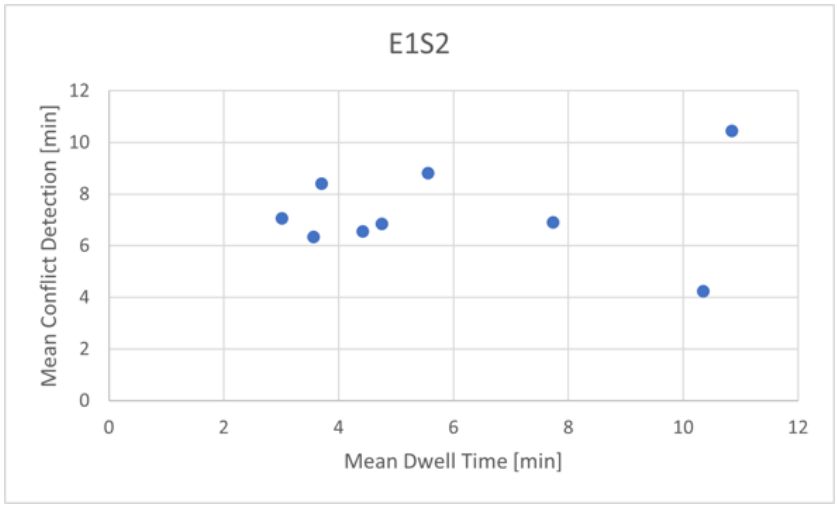


Figure 14: Dwell time compared to conflict detection for E1S2 scenario (N= 9)

Mean conflict detection time for most of the 9 ATCOs analysed was around 7 minutes. The two ATCOs with the longest mean dwell time detected conflicts on average either faster or much slower than the rest of the ATCOs. The remaining 11 ATCOs could not be included in this analysis, as they either had no eye-tracking data at all (N= 2) or the CVT tool (see 3.8.2.2) was not able to automatically map their gazes with sufficient confidence.

Discussion:

Intuitively it may be expected that ATCOs with average dwell-time would detect conflicts fastest. However, in a population of ATCOs that is strictly selected, such differences might be minimal. It is necessary to make more fine-grained analysis of dwell-time, but also the order of aircrafts scanned to get relevant insights about the relation of ATCO gaze behaviour and situation awareness. Long fixations (mean dwell time on aircraft) may indicate concentrated evaluation of information and enable fast conflict detection. In contrary, it may also be a sign of lack of understanding and go along with slow conflict detection. In combination with other features of gaze behaviour the findings on mean dwell time could possibly be interpret in a more distinct manner (e.g., number of revisits, average number of saccades or saccade velocity). Alternatively biometric parameters regarding the level of arousal and mental workload could be interesting to consider together with eye-tracking. It might provide information about mental states of high load and reduced capability for conscious information processing.

4.1.2 Correlational Results on Situation Awareness Measurement Methods

Research question Q1.2 investigates the consistency of different methods for situation awareness measurement. For this, overall scores were calculated for each situation awareness measurement method used - SASHA_Q, SASHA_L only in experiment 2, eye tracking for gaze analysis, implicit performance measurements (Section 3.8.5). The results of the correlations between the different methods are presented in this chapter. The Pearson correlation can take a value from -1 to 1, where 1 corresponds to a strong positive relationship (both parameters increase or decrease in the same direction), whereas -1 expresses a perfect negative relationship (if on parameter increases, the other decreases vice versa).

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Table 13 represents the correlation between the situation awareness measurement methods for the E1S2 scenario in experiment 1. Self-rating (SAHSA_Q) correlates moderately negative with conflict detection by eye-tracking (ET) (r_p= -0.31; p= 0.4) and implicit situation awareness measurement (r_p= -0.29; p= 0.2). The higher the subjective rating for situation awareness, the faster conflicts were detected (ET) and the shorter the conflict duration (implicit situation awareness measurement). And conflict detection (ET) correlated significantly positive with implicit situation awareness (conflict duration) (r_p= 0.82; p< 0.001). The earlier conflicts were detected, the faster they got solved vice versa.

Table 13: Pearson correlations between different categories of E1S2 scenario in experiment 1

E1S2	SASHA_Q [score] N=20	ET Conflict Detection [min] N=18	Implicit SA : Conflict Duration [min] N=18
SASHA_Q [score]	1	-0.3133	-0.2902
ET Conflict Detection [min]		1	0.8152***
Implicit SA : Conflict Duration [min]			1

* p < 0.05; ** p < 0.01; *** p < 0.001

Table 14 shows the Pearson correlation of the E2S2.1 scenario of experiment 2, where all four measurement methods for situation awareness were used. The percentage of correct SASHA_L query answers showed a significant high negative correlation with the detection time for conflicts (r_p= -0.79; p< 0.001) measured by eye-tracking (ET) and also with the score for implicit situation awareness measurements (r_p= -0.71; p= 0.004). There is a significant high positive correlation between the conflict detection time (by ET) and the score for implicit situation awareness measurements (r_p= 0.85; p=0.0002). So, if ATCOs detect the conflict earlier, the duration of the conflict is shorter, as they can start solving the conflict earlier. All correlations for questionnaire self-ratings (SASHA_Q) are non-significant and low (r_p= 0.15; p= 0.6 for SASHA_L; r_p= -0.06; p= 0.8 for conflict detection (ET) and r_p= -0.20; p= 0.5 for implicit situation awareness measurement).

Table 14: Pearson correlations between different categories of E2S2.1 scenario in experiment 2

E2S2.1	SASHA_Q [score] N=16	SASHA_L Correct [%] N=16	ET Conflict Detection [min] N=14	Implicit SA: conflict duration [min] N=14
SASHA_Q [score]	1	0.1508	-0.0621	-0.1985
SASHA_L Correct [%]		1	-0.7932***	-0.7095**
ET Conflict Detection [min]			1	0.8538***
Implicit SA : Conflict Duration [min]				1

* p < 0.05; ** p < 0.01; *** p < 0.001

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Discussion:

Overall, the intercorrelations in Table 13 and Table 14 reflect generally moderate to high consistency for the different situation awareness measures with the exception of subjective situation awareness measurement (SASHA_Q) in experiment 2. Despite low or no correlation with other measurement methods subjective situation awareness ratings might be especially valuable when combined with objective feedback on perception of air traffic monitoring details (see 4.1.1.1). Self-monitoring capability is an important skill for accurate situation awareness, adaptation and developing expertise in a domain. Used as a single indicator self-rated situation awareness might lack reliability, as it rather measures a stable self-concept ("how good am I at monitoring") than performance in monitoring tasks in a specific situation. But it seems to be receptive to specific feedback and critique in the sense of acknowledging the cognitive dissonance between expectation and real outcome.

4.1.3 Group-Level Analysis for ATCOs with Preserved and Degraded SA

For further analysis ATCOs who recognised conflicts fastest and resolved most conflicts were grouped as "preserved SA" (N=4). Vice versa ATCOs that detected least conflicts constituted the group "degraded SA" (N=3). 11 ATCOs with average performance on conflict resolution were assigned to neither group. They represent standard performance on detecting conflicts.

4.1.3.1 Comparisons of ATCO Groups for Implicit Measurement of SA

Results for all ATCOs in respect to implicit performance measures for situation awareness have been reported in section 4.1.1.2. This chapter focuses on the groups of ATCOs with preserved and degraded situation awareness. They are compared in regard to their interactions with pilots and (re)actions to events. (For a review of the relevant events see in Figure 13 in Section 4.1.1.2).

ATCOs with preserved situation awareness:

The best ATCO with preserved situation awareness missed one Quality of Service (QoS). The speed bust did not occur due to an error from the pseudo-pilot. The other ATCOs in this subgroup failed to solve two of the challenging events concerning a Quality of Service and an exit crossing.

General:
Common characteristics of these ATCOs are that they have made many transfer calls and have not duplicated any assume-calls, i.e., they only assumed aircraft once. Furthermore, the ATCOs' behaviour was rather inconspicuous: all necessary commands were executed in time. They neither had an excessive number of calls nor very few. ATCOs accomplished many steps in interactions with pilots with one single radio call. These ATCOs needed only a few steps to solve a conflict.

Evaluation of timing of radio calls:
None of the ATCOs performed above or below average in their execution time for transfer calls and flight level changes. It could be expected that these calls were made early, which would indicate good situation awareness. However, ATCOs behave rather standard. Two of the ATCOs made rather quick calls. These results could indicate that the ATCOs did not immediately make the required transfers or flight level changes but waited a short time to see if all conditions were fulfilled or if they discovered an opportunity to find a better solution. However, an early transfer can also indicate that the ATCOs needed free capacity and therefore sent the aircraft away as early as possible. If transfers are carried out too late, the ATCOs could have forgotten about it which can be due to distraction of high workload.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Some of these ATCOs used CPDLC messages, and some did not. No pattern emerged in this regard. It shows that the use of CPDLCs is a matter of habit and cannot be attributed to good or bad situation awareness.

Evaluation of measuring tool usage:
A similar pattern was seen with the use of speed vectors. Speed vectors can be used to estimate distances and plan the next steps. ATCOs used the tool with different frequency. This suggests that speed vectors are a matter of habit. Compared to the VERA tool ATCOs use in their work, distances can be estimated more precisely. ATCOs used the tool less often compared to all other ATCOs. However, this might not indicate that they have good situation awareness, but rather that they might feel confident in having assessed the situation correctly.

ATCOs with degraded situation awareness:

ATCOs with degraded situation awareness had the lowest situation awareness regarding conflict resolution. They had problems performing the QoS and detecting the exit conflict in the military scenario (E1S4). They also had problems with the level bust. It seems that unexpected or not directly obvious conflicts caused problems to them.

General:
It was found that these ATCOs tended to make fewer transfer calls and more duplicate assume-calls. Two of the three ATCOs did not perform all necessary actions required by events: actions regarding each aircraft were missing which all other ATCOs had performed. No similarities occurred regarding the number of actions per call: One ATCO's calls included very few actions whereas the others' encompassed many actions. The ATCOs also differed in the number of steps per conflict solutions: one ATCO needed many steps while the others performed standard.

Evaluation of timing of radio calls:
Two ATCOs performed the transfer and flight level change calls very late, whereas one ATCO performed the transfer calls exceptionally early. As a result, the ATCOs with degraded situation awareness stood out with either fast or slow calls.

CPDLC use is similar to the preserved ATCO group. Two ATCOs used it frequently, whereas one ATCO did not use it at all.

Evaluation of measuring tool usage:
The ATCOs did not often use the speed vectors and the VERA tool. It appeared that either one or the other tool was used preferentially. One-sidedness could be a risk for degraded situation awareness. These tools support assessing the situation and planning for the future. If they are not used appropriately, the situation might be misjudged.

Discussion

Comparison of ATCOs with a preserved situation awareness to those with degraded situation awareness showed differences in the timing of actions: ATCOs with preserved situation awareness executed actions neither extremely fast nor slow. The assumption that ATCOs with preserved situation awareness would perform transfer calls fastest can thus be refuted. It seems to be a sign of good situation awareness practice, if ATCOs neither rush nor hesitate. This is also confirmed when comparing all 18 ATCOs. The ATCOs who's timing was either very early or late had more problems to solve conflicts. In contrast, ATCOs performing well on conflict solution tended to execute actions

average to fast. Best performing ATCOs executed more additional actions (e.g., transfer calls) compared to other ATCOs. ATCOs with degraded situation awareness performed less actions than the rest of the ATCOs.

It is difficult to indicate generalizable signs of preserved situation awareness because ATCOs work consists of handling many interrelated events that need to be considered jointly. Difficulties in the degraded situation awareness group might be caused by general problems to adapt to the simulation environment and inconsistencies in the tools available compared to those at work. ATCOs who adapted well seemed to invest time in gaining an overview that allowed for a structured and efficiently organized approach. Speeding up might not be a good strategy because important information might be overlooked.

4.1.3.2 Comparison of ATCO Groups for Gaze-Based Analysis of Situation Awareness

To determine how scanning influences situation awareness, the ATCO groups with preserved and degraded situation awareness are compared regarding dwell time and frequency of gaze behaviour. Partially low confidence for automated mapping of areas of interests with the CVT reduced the number of ATCOs analysed to two for each group. Table 15 compares mean dwell time and mean dwell count of ATCOs with preserved situation awareness and degraded situation awareness for each scenario.

Table 15: Dwell time and frequency comparison of ATCOs with preserved and degraded situation awareness

Scenario	Situation Awareness	Mean Dwell Time [Sec]	Mean Dwell Count
E1S1	Preserved:	1.5	8.13
		1.28	16.8
	Degraded:	3.5	19.47
		1.94	13.47
E1S2	Preserved:	4.42	27.33
		3.56	35.38
	Degraded:	5.56	41
		10.35	52.25
E1S4	Preserved:	6.04	37.32
		6.22	75.77
	Degraded:	10.08	59.81
		10.3	52.35
E1S5	Preserved:	2.13	18.84
		2.13	25.89
	Degraded:	8	28.74
		5.08	42.21

EUROPEAN PARTNERSHIP
Co-funded by the European Union

All scenarios of experiment 1 are analysed except E1S3 scenario, because from CVT are not available for this scenario. As a general pattern ATCOs with preserved situation awareness on average scanned aircraft for shorter time and looked at aircraft less often - their revisit rate was lower. There were exceptions to this: in the E1S1 scenario the mean dwell time was short for both groups and in the E1S4 scenario one ATCOs with preserved situation awareness had the highest number of fixations on aircraft (mean dwell count).

Comparison of the dwell time and frequency:

Figure 15 shows dwell duration and dwell count per aircraft for one ATCO with preserved situation awareness (green dots) with one ATCO with degraded situation awareness (red dots). The size of the dots indicates how often an aircraft was looked at in the whole scenario. ATCOs with preserved situation awareness focused their attention less often on aircraft which do not require any ATCO clearances in all scenarios. For example, the difference in mean duration to the ATCO with degraded situation awareness for the HOP413F in the E1S4 scenario is 17.12 seconds. This aircraft only needs a flight level change, which is a routine task for ATCOs. ATCO with degraded situation awareness focused longer and more often on this aircraft. The same holds for BCS8531 the E1S5 scenario.

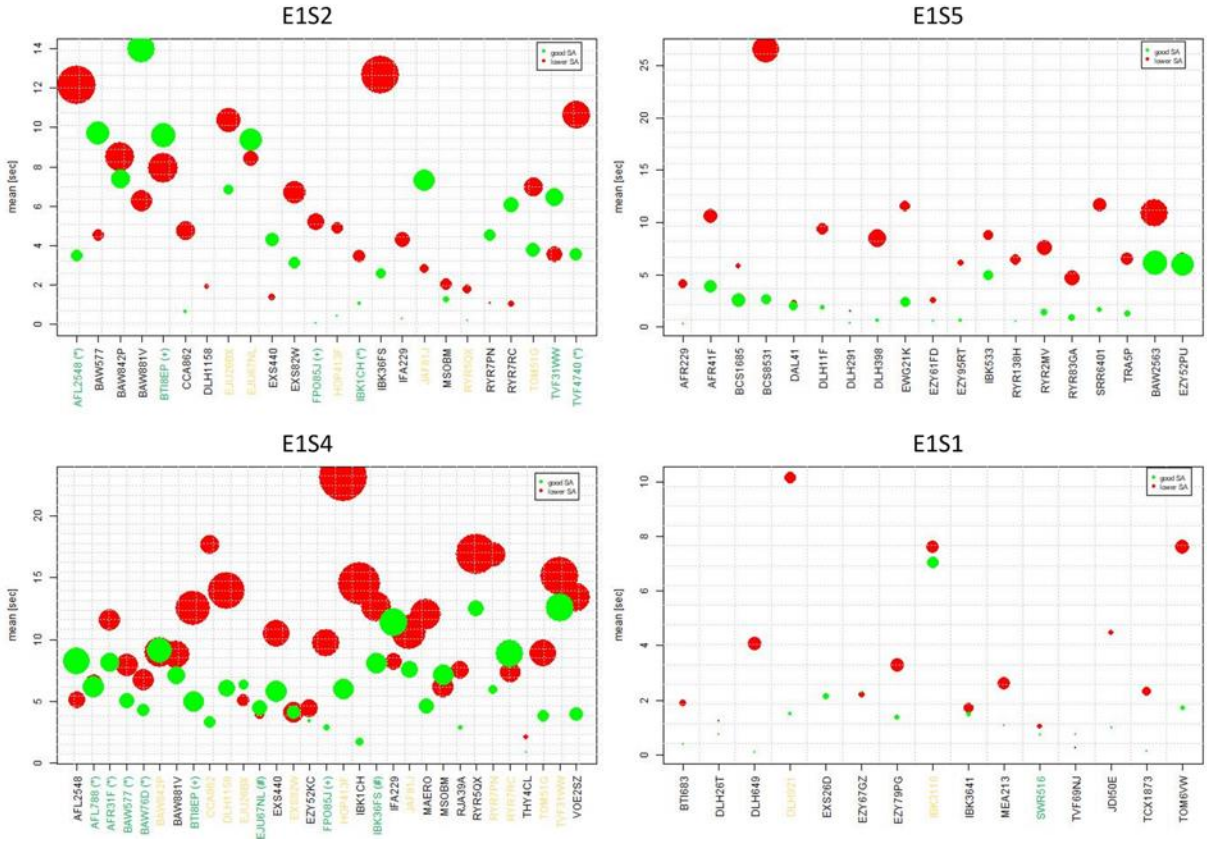


Figure 15: Gaze duration and gaze count for aircraft per scenario of ATCO with preserved (green circles) and degraded situation awareness (red circles). Green aircraft labels indicate aircraft participating in a conflict (symbol marking on aircraft label for respective conflict); yellow aircraft needed a flight level change and black aircraft did not require much attention.

In the E1S1 scenario the ATCO with degraded situation awareness looks at most aircraft more often (most red circles are bigger than the green circles) and for a longer time. Most red circles are in the

EUROPEAN PARTNERSHIP
Co-funded by the European Union

range between 2 and 10 seconds of mean dwell duration, whereas most green circles are in the range between 0 and 2 seconds. The ATCO with preserved situation awareness scanned more regularly and avoided fixations in E1S4, E1S5, and E1S1 scenarios. This pattern is not visible in the E1S2 scenario, where the ATCO with preserved situation awareness (green) also scanned less important aircraft. For example, the BAW881V is scanned for a long time although this aircraft is not important and requires few events. The complexity of this scenario is high, as can be seen by the number of aircrafts including many events. This could complicate ATCOs' scanning.

Comparison of scanning based on gaze plot:
ATCOs were compared regarding their scanning pattern shortly before the conflicts occurred. For each conflict, two ATCOs were considered who solved the conflict well or quickly and two ATCOs who solved the conflict slowly or not. For this purpose, five different conflicts were selected: the multiple crossing and the level bust in the E1S2 scenario; the exit crossing and QoS conflict in the E1S4 military scenario and the non-conformance in radio communication in the E1S1 scenario.

ATCOs who solved the conflicts scanned quickly and regularly and were not fixing on one point. The scanning followed a pattern.

Triple Crossing: The triple crossing in the north was scanned similarly by all ATCOs. Once the aircraft were on the radar, they were scanned, and the conflicts were measured. However, since the aircraft were not yet on frequency, the ATCOs could not immediately intervene. Most ATCOs left the VERA tool active until they could solve the conflicts. In the meantime, they scanned the entire radar area again. As soon as the aircraft were on the frequency, the ATCOs gave the necessary instructions to solve the conflict. The ATCOs who turned off the VERA tool and used it again later scanned the conflicting aircraft more frequently before they could start solving the conflict. They did react to the conflict when the pilots were calling. Instead, they scanned the aircraft and solved the conflicts at a later point. The reason for this could have be that sometimes other aircraft report on the frequency, and therefore the ATCO's attention got absorbed. This does not mean that these ATCOs per se had a degraded situation awareness.

Level bust: ATCOs who did not solve this conflict were scanning faster and more frequently. Their focus was not on all aircraft but absorbed by the triple crossing. Even if ATCOs scanned the level busting aircraft they did not realize and react to the non-compliance. Their attention was absorbed by other aspects so they could not perceive the relevant information during scanning of the non-compliant aircraft. ATCOs who recognised the level bust realized it early. They regularly scanned all aircraft to be aware of other sorts of conflicts.

Exit crossing: The exit crossing was evident from the start, but few ATCOs immediately responded to it. The ATCO who solved this conflict quickly scanned regularly and had few fixes. ATCOs who did not detect the conflict scanned the relevant aircraft but did not recognize the conflict between the two aircraft involved in the exit crossing. One of these ATCOs scanned all aircraft that were on the same flight level several times but did not detect the crossing. Another ATCO was focused on a potential crossing from the southeast. When the aircraft further approached the location with the exit crossing the focus was absorbed by different aircraft.

QoS: In Quality of Service, all ATCOs scanned the aircraft around the Centre as it became available. There was concurrent requirement for attention by other aircraft shortly after the change in military activity. ATCOs who solved the QoS well scanned the aircraft around the Centre repeatedly. The ATCOs who detected fewer aircraft lost focus of the relevant aircraft because of other aircraft that needed

EUROPEAN PARTNERSHIP
Co-funded by the European Union

input at the same time. It seemed they had forgotten the aircraft that could profit from QoS. All ATCOs scanned all aircraft involved, yet no ATCO performed all of the possible QoS.

Non-conformance: Although all ATCOs scanned the relevant aircraft while the non-conformance occurred or after the conflict was not always detected in the E1S1 scenario. Again, it is noticeable that when in the E1S1 scenario. In cases where the non-conformance was detected, this was done so shortly after it occurred. Those ATCOs specifically checked if the aircraft performed the required flight level changes, and thus the conflict was detected. The other ATCOs assumed that the aircraft complied with the flight level change and therefore overlooked the non-conformance even though they scanned the aircraft.

Discussion:

Although all ATCOs scanned the aircraft that could profit from QoS, the reaction was differently. ATCOs either did not perceive the opportunity for a QoS when they scanned the aircraft concerned or they judged this situation differently - did not consider a QoS necessary or useful. It is crucial to scan regularly and not to get fixed on single aircraft or conflict because more conflicts could arise at the same time. Free capacity – if available – can be used to check if aircraft comply with instructed changes. What caused some ATCOs to miss conflicts while others recognized them readily is difficult to judge. Even if aircraft were scanned regularly, not all conflict were detected. Scanning needs to be fast and regular but to be effective it might need to be slow enough to detect and extract relevant information. A requirement that is probably hard to keep up with when many aircraft are on the radar and conflicts start to pop up. To resist the urge to hurry up might be important.

More obvious conflicts were generally solved faster, for example the triple crossing in the E1S2 scenario where ATCOs needed to zoom out to see the aircraft calling in. However, the exit crossing in the military scenario E1S4 was visible from the beginning but attention was not directed to the conflict. Distance between aircraft was still at the beginning and other aspects seemed were more urgent at that time.

4.1.4 Performance and Workload

4.1.4.1 Intercorrelation of Performance Aspects

The relationship between different aspects of performance is investigated with Pearson correlations. Radio communication duration and radio communication frequency with aircraft involved in conflicts represent implicit measures of workload.

Table 16 shows intercorrelations among these aspects with the duration of conflicts and mean reaction time to initial calls.

Correlations between the four performance aspects are low to moderate: r_p ranges between -0.05 and 0.31. Correlation of duration of radio communication with the frequency of radio communication ($r_p$is low and positive (r_p= 0.25; p= 0.3) and a moderately with the duration of conflicts (r_p= 0.31; p= 0.2). The longer the radio communication duration, the more often ATCOs communicate. Correlation of duration of radio communication with duration of conflicts is moderate and positive (r_p= 0.31; p= 0.2). The longer the radio communication duration was the longer it took ATCOs to resolve the conflict. Correlation of duration of radio communication with mean reaction time of ATCOs ($r_p$to initial calls is very low and negative (r_p= -0.05, p= 0.8).

EUROPEAN PARTNERSHIP
Co-funded by the European Union

ATCOs who were more busy communicating - indicating more workload with handling traffic - solved conflict slower when communication took longer (r_p =0.25) and faster if they communicated more frequently (r_p = -0.18).

Correlations of the frequency of radio communication are low and negative with the duration of conflict (r_p= -0.18, p= 0.5) and with the mean reaction time (r_p to initial calls (r_p= -0.17, p= 0.5). If ATCOs communicated more or less often did not matter for how fast conflicts were solved and initial calls were reacted upon. I.e., long conflict duration was present with many and with few radio communication calls to solve them.

Table 16: Pearson correlation of performance measures

E1S2 (N=18) Pearson Correlation	Radio Communication (duration) [sec]	Radio Communication (frequency) [count]	Conflict Duration [min]	Mean Reaction Times to Initial Call [sec]
Radio Communication (duration) [sec]	1	0.2462	0.3139	-0.05401
Radio Communication (frequency) [count]		1	-0.1775	-0.1737
Conflict Duration [min]			1	0.1085
Mean Reaction Times to Initial Call [sec]				1

Correlation of ATCO reaction time to initial calls and duration of conflict was low (r_p= 0.11, p= 0.7). ATCOs who reacted faster to initial calls did not automatically recognize and solve conflicts faster

Discussion:

Correlations among performance aspects related to radio communication and to reaction time (time to respond to initial call in) and outcome (duration of conflict) are generally low indicating low consistency. These aspects of ATCO performance are independent. Longer duration of communication, but not the frequency of communication goes together with longer duration of conflict. Concerning implicit workload aspects of performance ATCOs who were more busy communicating - indicating more workload with handling traffic - solved conflict slower when communication took longer (r_p =0.25), but faster if they communicated more often (r_p = -0.18). However, the strength of relationship is rather weak and statistically non-significant.

4.1.4.2 Psychophysiological Reactions to Events

Workload was measured objectively with skin conductance and heart rate for phases where selected events occurred (see Section 3.10.2). This was done to investigate which events were most demanding

EUROPEAN PARTNERSHIP
Co-funded by the European Union

for ATCOs. For interindividual comparisons psychophysiological measures need to be related to an individual baseline to calculate relative changes in reactions (see Section 3.10.3 and 3.10.4).

Figure 16 and Figure 17 describe the relative changes of the median heart rate and skin conductance. The straight black line at the value 0 represents the baseline against which the event-specific psychophysiological activation is compared (for description of baseline see). Both parameters show a large range of variance which indicates that the magnitude of ATCOs' reaction to events varies strongly for all events.

The line within each boxplot in Figure 16 shows the median heart rate for the respective event. All event-related medians are lower than the baseline for heart rate calculated from all the phases that were not related to scenarios. This indicates that ATCOs' heart rate activity and hence the workload level were lower during simulation when ATCOs handled air traffic than during all other phases (before and after simulation, during breaks between scenarios) that were taken as a baseline.

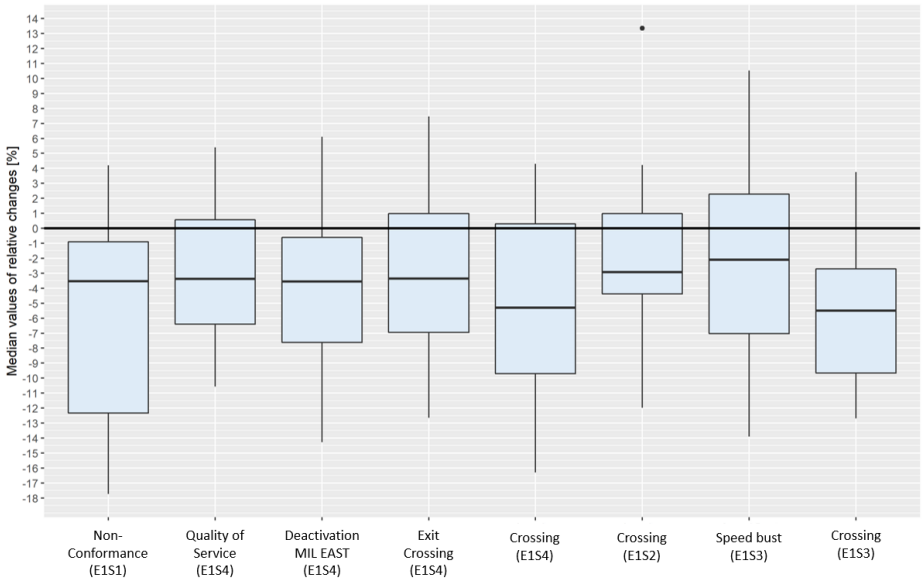


Figure 16: Cumulated distribution of relative changes of the median heart rate

Median heart rate level is similar for all events in Figure 16, the variance differs. Median heart rate was lowest for crossing in E1S4 scenario and highest for speed bust in E1S3 scenario.

Figure 17 shows median values of electrodermal activity during specific events measured as skin conduction. For all events except the speed bust in E1S3 scenario the median skin conduction is below the baseline value and varies across different events. Variance is generally high, especially for crossing in E1S3 scenario. Skin conductance is highest in the speed bust event in scenario E1S3 and lowest in the deactivation MIL EAST event in the E1S4 scenario. Higher skin conductance indicates a reduction in skin conduction resistance associated that is associated with higher mental activation. ATCOs were most activated when the speed bust occurred and least activated when MIL EAST was deactivated. Redirecting aircraft to avoid an active military zone did not create any additional mental workload compared to the baseline – it was just the opposite.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

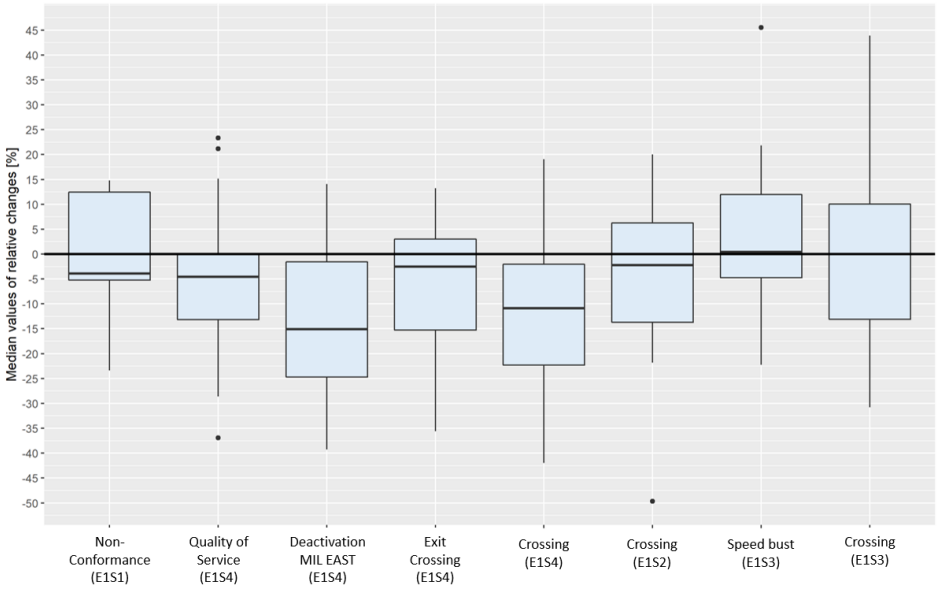


Figure 17: Cumulated distribution of relative changes of the median skin conductance

Discussion:

Event-related analyses of psychophysiological parameters did not show significant elevation of mental workload compared to baseline. In contrary, ATCOs level of activation and mental workload was lower – they seemed more relaxed when working in the simulation than in off-task phases including phases before and after the simulation and during breaks between the scenarios. This can be seen as an indication of "flow" during work performance. Flow (Csikszentmihalyi, M., 1987, 2000) describes a mental state of total absorption (full concentration), that is reached, when someone experiences an optimal level of demand (neither too high, nor too low). Additionally, due to time constraints the baseline was not determined with a standardized method as suggested by the provider. Therefore, the baseline might include phases where ATCOs were nervous before the simulation experiment or frustrated after a scenario, when ATCOs were not satisfied with their performance or struggled with the unfamiliar simulation tool. As a second aspect it might be necessary to shorten the time span considered for accumulation of psychophysiological parameters. MIL EAST deactivation in scenario E1S4 for example lasted 10 minutes. Single peaks of psychophysiological activation might have been eliminated by averaging. Further analysis is needed for shorter periods of event-related parameters and a more homogeneous baseline should be used – possibly from a scenario with low complexity and high familiarity to ATCOs.

4.2 Results on Comparison of Human and Machine Situation Awareness

Research question 2.1 asked if artificial and human situation awareness are comparable. This question was investigated by comparing ATCOs' and AI SA's answers to queries about events related to en-route air traffic monitoring tasks. In experiment 2 ATCOs and AI SA system were compared in respect to their answers to queries (SASHA_L) on specific aspects of the situation. For this comparison FTTS of the University of Zagreb created outputs from raw data of the AI SA system applying filters, as the system implementation was still in progress. The filter sorted the AI SA system's raw data to ignore irrelevant outputs. The results are presented in Table 17. The queries were labelled like in the Appendix A.2.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

The fourth column of Table 17 shows the number of ATCOs who had named the respective aircraft in their query answers. Further to the right it is indicated if AI SA system detected the respective aircraft. (Note that the AI SA output was generated with 8 of 57 en-route monitoring tasks implemented.) The last column on the right indicates the correctness of the notion judged by two subject matter experts.

Table 17: Comparison of query answers between ATCOs and AI SA system (N= 16 ATCOs)

Scenario	Query	Answers	# Selected by ATCO	Mentioned by AISAs	Valid
E2S2.1	1.1	BAW842P & IBK1CH	0	X	Yes
		IBK1CH & TVF4740	0	X	Yes
		BAW577 & IBK36FS	5	X	Yes
		None	11		no
	1.2	IBK1CH & TVF4740	8	X	Yes
		TVF4740 & AFL2548	4	X	Yes
		None	7		no
	1.3	BTI8EP & FPO85J	4	X	Yes
		TVF4740 & AFL2548	3	X	Yes
		TVF4740 & IBK1CH	2	X	Yes
		JAF81J & EXS82W	1	X	Yes
		None	8		no
	1.4	IBK1CH & AFL2548	0	X	Yes
		JAF81J & EXS82W	3	X	No
		AFL2548 & RYR5QX	2	X	Yes
		BAW842P & TVF4740	1	X	No
		TOM51G & EXS440	1		No
		None	10		no
	1.5	None	16		Yes
E2S2.2	2.1	TVF4740 & IBK1CH	13	X	Yes
		TVF4740 & AFL2548	6	X	Yes
		BTI8EP & FPO85J	11	X	Yes
		RYR5QX & AFL2548	0	X	Yes
		IBK1CH & BAW842P	1	X	Yes
		RYR5QX & TVF4740	0	X	No
		None	1		No
	2.2	TVF4740 & IBK1CH	13	X	Yes
		TVF4740 & AFL2548	3	X	Yes
		RYR5QX & AFL2548	0	X	Yes
		EXS82W & JAF81J	1	X	Yes
		TVF31WW level bust	1	X	Yes
	2.3	TVF31WW level bust	10	X	Yes
		No	3		No

EUROPEAN PARTNERSHIP
Co-funded by the European Union

	2.4	Nothing	6		No
		EXS82W & JAF81J	2		No
		IBK1CH & AFL2548	1	X	Yes
		AFL2548 & RYR5QX	1	X	Yes
	2.5	No	14		Yes
E2S3	3.1	TOM4BA & VPBVV	12	X	yes
		GAC818X & VPBVV	3		no
		GAC818X & TOM4BA	1		no
		GAC818X & TUI618P	1		no
		TUI618P & TOM4BA	0	X	no
		TUI618P & VPBVV	0	X	no
		None	2		no
	3.2	TOM4BA & VPBVV	5	X	yes
		ECLCX & DLH319	0		yes
		DLH319 & TOM33H	0	X	yes
		TUI618P & TOM4BA	6	X	yes
		GAC818X & DAL467	7	X	yes
		No	4		no
	3.3	TOM4BA & VPBVV	2	X	yes
		ECLCX & DLH319	0		yes
		DLH319 & TOM33H	2	X	yes
		TUI618P & TOM4BA	3	X	yes
		GAC818X & DAL467	3	X	yes
		No	8		no
	3.4	Wrong readback	5		Yes
		Nothing	6		No
	3.5	TOM51G speed bust	3	X	Yes
		No	11		No
	3.6	Nothing	5		no
		EZY68HL transfer w/o climbing	5		yes
			2		yes
		MAC236 needs to climb	2	X	yes
			1		yes
		TOM4BA & VPBVV	3	X	yes
		THY4CL & IBK5VZ	1		yes
		ECLCX & DLH391	1		no
		GAC818X & TOM51G	1		no
		GAC818X & DAL467			

EUROPEAN PARTNERSHIP
Co-funded by the European Union

		DLH1300 needs to climb			
	3.7	TOM51G MAC236 GAC818X	12 7 8	X	Yes No Yes
E2S4.4	4.1	HOP423F	15	X	Yes
	4.2	AFL788 AFR31F BAW577 BAW76D BAW881V MSOBM MAERO CCA862 TVF31WW TOM51G	5 4 12 12 11 12 6 1 1 13	X X X X	yes yes yes yes yes yes yes yes no yes
	4.3	TOM51G & EJU67NL IBK36FS & BAW577 No	9 5 6	X	Yes Yes No
	4.4	No	15		Yes
E2S4.2	5.1	IBK36FS & EJU67NL No	10 6		Yes No
	5.2	TVF31WW EXS440 BAW842P EXS82W JAF81J RYR7RC TOM51G	11 11 3 6 6 6 1	X X X X X X	yes yes yes yes yes yes no
	5.3	BTI8EP & FPO85J IBK1CH & BAW842P IBK1CH & AFL2548 None EJU67NL & IBK36FS FPO85J & VOE2SZ	2 2 0 8 0 1	 X X	yes yes yes no yes no
	5.4	No	15		Yes

The ATCO must zoom out very far in the **E2S2.1 scenario** to reveal the aircraft from the north (BAW842P, IBK1CH and TVF4740). Therefore, no ATCO has answered query 1.1 thoroughly and comprehensively. Since these aircraft are also not yet on the frequency, it is not severe that the

EUROPEAN PARTNERSHIP
Co-funded by the European Union

conflicts are not detected. The AI already recognises the two northern conflicts and informs the ATCO about this. Therefore, the ATCO could recognise them in later queries about conflicts. Indeed, more ATCOs recognise the conflicts in the north in query 1.2 (8 ATCOs compared to 0 ATCOs in query 1.1). After query 1.3 and 1.4 there are again fewer ATCOs that name the crossings. The non-conformance query (1.5) is answered correctly by all ATCOs.

In general, the crossings in the north are not detected by many ATCOs. Compared to the AI SA system, humans perform worse in this scenario. Moreover, the AI SA system detects conflicts much earlier. However, it fails to recognise a crossing at the beginning of the scenario and points out a crossing at query 1.4, which was declared non-relevant by the subject matter experts.

E2S2.2 scenario was always performed directly after the E2S2.1 scenario. Since the two scenarios overlap, it was assumed that ATCOs could be aware of previous conflicts and recognise them more easily. This was confirmed by the data: More ATCOs named the crossings from the north, but still not all ATCOs. The level bust (it did not occur in the E2S2.1 scenario before) was mentioned by 10 out of 20 ATCOs in query 2.3. The non-conformance query 2.5 was answered correctly by almost all ATCOs (14 of 16 ATCOs). Again, the AI SA system performs better for the same reasons as in scenario E2S2.1.

In the **E2S3 scenario**, seven queries have been asked. Twelve out of 16 ATCOs recognised the speed bust between TOM4BA and VPBVV at the first question (3.1). However, the crossing between TUI618P and TOM4BA was recognised only by 6 out of 16 ATCOs–after the ATCOs received the AI SA input. The false readback in query 3.4 was only recognised by 5 ATCOs and the speed bust (query 3.5) only by 3 ATCOs. For query 3.7, many ATCOs named the correct aircraft. In this scenario, too, ATCOs perform less effectively than the AI SA system, especially in non-conformance queries. The AI SA system could not be asked about non-conformance "wrong readback", as it cannot process radio communication data. It could not provide AI SA input to ATCOs on that aspect.

In the **E2S4.1 scenario** query 4.1 was answered correctly by 15 ATCOs. For query 4.2, the majority of the ATCOs named most of the aircraft that could profit from a direct to. Only the AFL788 and AFR31F were not mentioned often by the ATCOs. ATCOs recognise more different aircraft that required a direct route than the AI SA system. Also, in the following query 4.3, five ATCOs mention a crossing that the AI SA system did not recognise. 15 ATCOs answer query 4.4 correctly, whereas AI SA wrongly indicated a non-conformance. In this scenario, it is no longer clear who has a better situation awareness. However, since the ATCOs identified more aircraft for a direct call, their situation awareness is considered better than artificial situation awareness in this scenario.

In the **E2S4.2 scenario** 10 out of 16 ATCOs noticed the crossing between IBK36FS and EJU67NL from the beginning, so did the AI SA system. For query 5.2 about aircraft flying through military zone, two aircraft are named by 11 ATCOs, other aircrafts are mentioned only by 6 or 3 ATCOs. The AI SA system, in comparison, recognises all aircraft correctly. Also, in the query 5.3 about conflicts, relevant crossings were named by few ATCOs (only 2) or no ATCO. 8 ATCOs reported a crossing that was not considered relevant by subject matter experts. In the last query 5.4 about non-conformances 15 ATCOs stated correctly that there was none–the AI SA system also reached this conclusion. In this scenario the AI SA system again had a better situation awareness–detected crossings and aircraft flying through the military reliably.

Discussion:

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Research question 2.1 asked if artificial and human situation awareness are comparable. This question was investigated by comparing ATCOs' and AI SA's answers to queries. From the results in Table 17 it looks – at first glance – as if the AI SA system had a better situation awareness than ATCOs. It correctly pointed out conflicts that in many times only a minority of ATCOs detected and sometimes even no ATCO named. However, it also indicated conflicts that were not present – as ATCOs also did. And some conflicts or events were missed by both. To interpret the results the circumstances of the experiment must be considered. These will be discussed in more detail in chapter 5.

4.3 Evaluation of AI SA's Contribution to Human-Machine Team Situation Awareness and Performance

This chapter describes how AI SA inputs contribute to team situational awareness. For this purpose, behavioural coding is used to compare the ATCOs from both experiments and the SASHA_L responses are used to compare the ATCOs from experiment 2 with the AI SA.

4.3.1 Evaluation of ATCO's Performance Based on Behavioural Coding

Research question 3.1 investigates if AI SA inputs enhance ATCOs' performance. For that purpose, performance parameters for implicit measurement of situation awareness were compared across the condition "without AI SA input" in experiment 1 and the condition "with AI SA input" in experiment 2, using the identical scenario in both experimental runs.

It is expected that AI SA contributes effectively to human-machine team situation awareness, e.g., by providing early warning of conflicts as well as alerts about non-conformances. Since only one scenario was interactive (E2S2.1) in experiment 2, only that scenario could be used for comparison. This allowed comparison of ATCOs' traffic handling in regard to three crossings for which ATCOs without AI SA support can be compared with ATCOs that were supported by AI SA inputs.

In both experiments all ATCOs solved the conflicts in the scenario. Table 18 shows on the left side how long the three conflicts lasted on average in the two experiments. In the second and third conflict ATCOs solved the conflicts faster. This makes sense since the AI SA system pointed out the conflicts early on and in most cases the ATCOs reacted to the message and detected the conflicts. However, ATCOs react slower on average to the first conflict in experiment 2 (with AI SA input) than in experiment 1.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Table 18: Comparison of conflict solution by ATCOs from experiment 1 (N=18) and 2 (N= 14)

	Conflict Duration [Sec]		Start of Solution [Sec]	
Conflict	**Without AI SA Input (Exp.1)**	**With AI SA Input (Exp. 2)**	**Without AI SA Input (Exp.1)**	**With AI SA Input (Exp. 2)**
1. TVF4740 & IBK1CH	53.4	59.4	412.13	461.41
2. TVF4740 & AFL2548	186.6	96.6	547.81	533.62
3. FPO85J & BTI8EP	156.6	119.4	470.5	548.01

Were the ACTOs able to detect conflicts faster? Start of solution was only slightly earlier with AI SA inputs in experiment 2 and only in regard to 1 out of 3 conflicts. It was expected that ATCOs react to conflicts faster, when they received respective AI SA input. However, that was not confirmed in 2 of 3 conflicts.

Discussion:

Results partially confirm the expectation from research question 3.1 – ATCOs solved some of the conflicts faster. Results showed that in 2 out of 3 conflicts, AI SA inputs allowed for faster conflict solutions. Except for the first conflict, where the AI SA input pointed to a conflict that most ATCOs had not zoomed out far enough to be able to see the aircraft involved in that conflict. So, the AI SA input came too early: Only few ATCOs reacted on the AI SA input by zooming out to localise the reported aircraft. That way the AI SA input was simply lost as a situation awareness prompt and the ATCOs did not react on it. The solution of that conflict on average hence also started late.

On the other hand, ATCOs with AI SA input did not seem to profit much in terms of an earlier onset of the start of solution. This might also be due to their judgment and personal preference about when and how to address a conflict.

In summary, some AI SA inputs enhanced ATCO performance, but support by AI SA inputs for early anticipation of conflicts was limited in its beneficial impact for conflict solving if it was delivered too early - when conflicting aircraft were noy yet in the ATCOs' sector and were not yet on the frequency at the time the AI SA input was given. Hence, the ATCOs could not solve the conflicts immediately nor quickly. Therefore, an early start of conflict solution depends not only on how well ATCOs can process the AI SA input given in an oral format.

Further analyses of performance in experiment 2 showed that ATCOs solve the conflicts more efficiently. That is, they need fewer clearances to solve the conflict. The mean number of clearance-related actions in experiment 1 was 1.71 attempts compared to 1.33 in experiment 2. This might have been enhanced by early indication of the conflicts by the AI SA inputs that enabled the ATCOs to better assess the situation and organise their reaction to it.

Execution-time to events was on average faster in experiment 1 than in experiment 2 (mean difference: 0.36 min). This might indicate that ATCOs from experiment 2 were absorbed by answering queries and listening to oral AI SA inputs and hence could not execute actions in this time span.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Therefore, conclusions about the usefulness of machine situation awareness based on measured time to reaction and conflict solution should not be taken at that stage.

In summary, there is evidence that AI SA input helped to perceive and solve conflicts earlier. However, it is important that the inputs are not given too early, otherwise the information may be lost – for reasons of working memory capacity–if it cannot be applied immediately and it creates memory load. Furthermore, it plays an important role how the inputs are transmitted to the ATCOs. If the inputs were displayed visually instead of oral inputs, ATCOs would have been less distracted, and the information were at disposition for later checks. In conclusion, the AI SA system contributed to conflict detection but still needs to be improved for conflict solving to be faster.

4.3.2 Evaluation of Artificial Situation Awareness Based on Questionnaire Answers

The inputs of the AI SA system were also analysed in a subjective manner to answer research question 3.2 (usefulness of AI SA inputs) and 3.3 (use of AI SA inputs for situation awareness and decision making): ATCOs were asked for their judgement on each AI SA inputs in terms of relevance, accuracy, and trust.

Table 19 shows results of ATCO judgements for relevance of AI SA inputs in the 5 scenarios investigated in experiment 2.

Table 19: Evaluation of the AI SA inputs by the ATCOs regarding relevance (N=16=

Scenario	No of AI SA inputs	Relevance of Input
E2S2.1	6	On average across all inputs, 40% of the ATCOs identified the crossings inputs as relevant. 60% of the ATCOs felt that these inputs were somewhat irrelevant.
E2S2.2	7	26% of ATCOs judged the crossing inputs as relevant. 74% of ATCOs felt that these inputs were rather irrelevant. However, the reference to the level bust was perceived as relevant by 88% of the ATCOs.
E2S3	8	On average, 47% of the ATCOs described the inputs on crossings as relevant. 53% of the ATCOs found these inputs somewhat irrelevant. The reference to the speed bust was considered relevant by only 44% of the ATCOs. The input that the TOM51G must climb to reach its exit level was only considered relevant by 2 ATCOs.
E2S4.1	3	The input for aircraft that can get a direct was only considered relevant by 6.25% of the ATCOs. The fact that the HOP413F must descend was only considered relevant by 12.5%. However, the input on the crossing was considered relevant by 18.8% ATCOs.
E2S4.2	5	Input on the crossings was considered relevant by 58% of the ATCOs. The input on which aircraft fly through military was found relevant by 38% ATCOs.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

With the results from experiment 2 that used an AI SA system with stage I of implementation (see 1.2.4) it is not possible to confirm research question 3.2. Only few ATCOs acknowledge relevance to AI SA inputs at that time and in that manner (see 3.7).

In summary, the AI SA inputs that the ATCOs considered relevant were those that pointed out conflicts that the ATCOs themselves did not recognise. This is exemplified by the input that the TOM51G must climb to the exit flight level. Most ATCOs themselves recognised this need in the high scenario (E2S3) and thus marked the input as irrelevant. The AI SA input about the crossing between IBK36FS and EJU67NL in the military 2 scenario (E2S4.2) on the other hand, was not recognised by any ATCO and thus marked as relevant. What is interesting, however, is that only 3 ATCOs recognised the speed bust in the high scenario, but this input was perceived as relevant by only 7 ATCOs. It seems that such a conflict is not considered to be a significant problem. Interestingly crossing inputs were generally perceived as less relevant, although some of the crossings were not recognised by the ATCOs.

The results for accuracy and comprehensibility are summarised across all AI SA inputs. The **accuracy** of the AI SA inputs was acknowledged by more than 50% of all ATCOs for all AI SA inputs across all scenarios (for each scenario: E2S2.1: 52%, E2S2.2: 57%, E2S3: 64%, E2S4.1: 79%, E2S4.2: 65%). Highest agreement was found for the level bust in the high traffic scenario: 14 out of 16 ATCOs agreed that this AI SA input was accurate.

The **comprehensibility** of the AI SA inputs was generally affirmed by more than 50% of the ATCO for all AI SA inputs across all scenarios (for each scenario: E2S2.1: 53 %, E2S2.2: 67%, E2S3: 67%, E2S2.1: 77%, E2S2.2: 77%).

How much ATCOs **trusted** the AI SA inputs was assessed for each scenario. Trust varied between the scenarios. The E2S2.2 scenario scored lowest: most ATCOs did not trust the AI SA inputs and– compared to the four other scenarios – the fewest ATCOs trusted the AI SA inputs. Overall, 51% of ATCOs rated the AI SA inputs of all scenarios as trustworthy, 23% tended not to trust the inputs, and 26% were neutral. In the final debriefing, the ATCOs were again asked how much they would trust an AI SA system in the future. Again, 50% of ATCOs selected a medium or high level of trust, 31% ATCOs were neutral, and 19% ATCOs had low trust in AI SA system.

Figure 18 compares the level of trust ATCOs have in current automation at ATCO working position with trust they have in AI SA system. 11 out of 16 ATCOs clearly trusted Skyguide's system Skyvisu compared to AI. However, ATCOs willingness to trust a future AI SA system is somewhat lower (8 out of 16 ATCOs). 3 ATCOs distrusted AI to some extent. The strongest level of distrust was not indicated neither for current system automation nor for AI-based system. So, ATCOs indicated that they are rather open to the potential of AI-based support. Compared to Skyvisu, which is an advanced system with many supporting tools, the AI SA system did not seem helpful at the first moment, as some information and tools are still missing. ATC is a safety-critical area, and therefore an AI SA system must be functional at 100% of the times so that no errors can occur.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

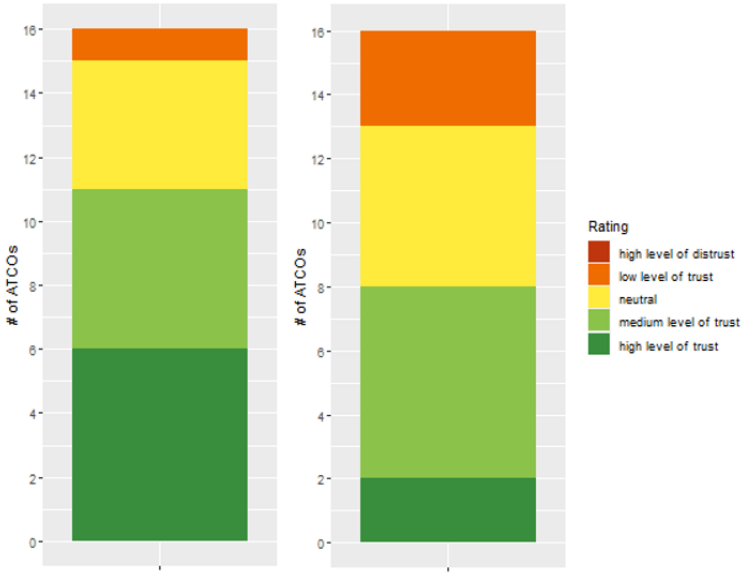


Figure 18: How much do you trust 1) the automation implemented in Skyvisu and 2) AI SA inputs at work in future?

ATCOs' openness to a new functional system was emphasised also by the fact, that 11 out of 16 ATCOs testified in the debriefing questionnaire that they could imagine working with an AI-based tool in the future, whereas 5 ATCOs indicated they were not sure.

To answer research question Q3.3 ATCOs were asked at the end of each scenario, if they had used AI SA inputs for their situation awareness and decision making. Figure 19 shows the results for use of AI SA inputs for situation awareness in experiment 2. Green indicates agreement that AI SA inputs were supportive for situation awareness. Only a quarter of the ATCOs or less did indicate they used AI SA inputs for their own situation awareness. Overall, the results indicate insufficient support for the expected subjective usefulness of AI SA inputs for situation awareness.

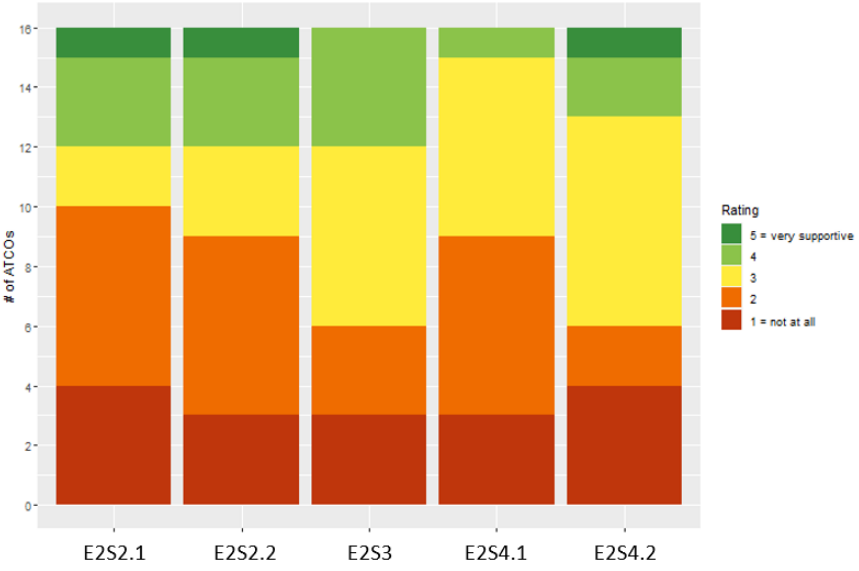


Figure 19: Did AI SA inputs support your situation awareness overall? (N= 16)

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Figure 20 shows the ATCOs responses to the question whether they could use AI SA inputs for their decision making. Again, the green colour indicates that the ATCOs used AI SA inputs in their decision making. Supportiveness of AI SA inputs for decision making was judged very low (only 2 to 3 out of 16 ATCOs). Therefore, these results did not support research question 3.3 neither.

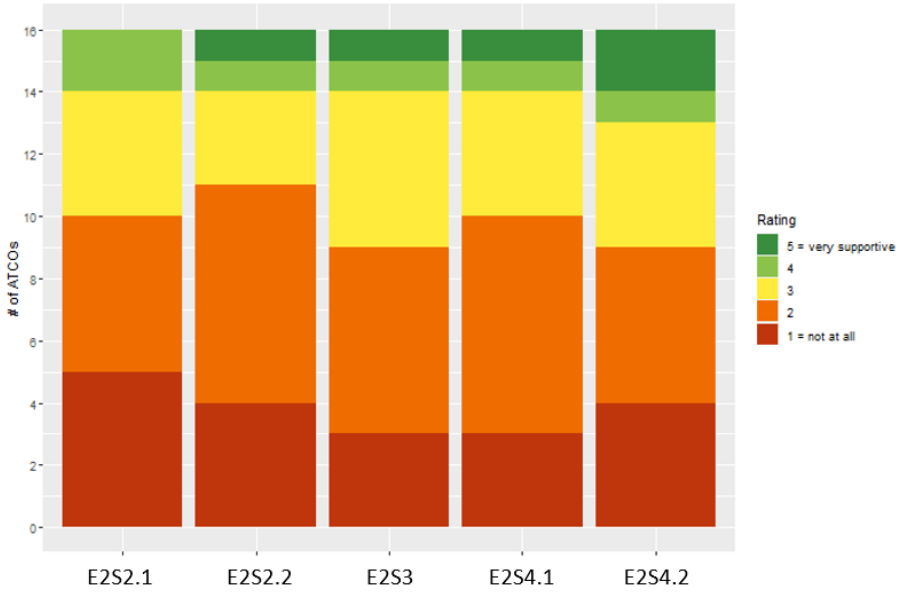


Figure 20: Did AI SA inputs support your decision making? (N= 16)

Discussion: Although some ATCOs failed to detect conflicts on their own, AI SA input was rated poorly. A possible reason for this may be that ATCOs sometimes did not have the capacity to listen to and implement the AI SA inputs. This was seen when ATCOs did not mention conflicts in their answer to queries, although the AI SA input made a message about it before. If the ATCOs cannot implement the AI SA inputs, the artificial situation awareness cannot support the situation awareness and decision-making. Additionally, many of the AI SA inputs were mentioned early, where ATCOs might have felt distracted as there was nothing urgent about them. By the time this information would have been relevant for ATCOs, they might have forgotten them already.

This is affirmed by some of the ATCOs' critique that AI SA inputs were given at inappropriate times. As a result, the significance of the inputs was often lost because the ATCOs could often not directly work on the conflicts, as the involved aircraft were not yet on frequency, for example. In those cases, AI SA inputs represented more a load for memory than an immediate help. Another criticism was that the inputs were transmitted orally and not visually. Visual inputs would have allowed ATCOs to access and incorporate the input when needed and check them later as often as needed. This point of criticism was expected but could not be prevented due to time constraints. Some ATCOs have also mentioned that the information was missing where the aircraft from the inputs are located. When ATCOs searched for the respective aircraft, they might have lost the capacity to listen carefully to the inputs. Furthermore, the AI SA system should be able to point out unresolved conflicts several times and emphasise the urgency. The inputs should also include when the crossing occurs in minutes and miles. This way, the situation could be assessed more thoroughly.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

4.4 Results on the Accomplishment of Artificial Situational Awareness

Experiment 1 has provided flight data that was used as input for the KG system, to analyse machine situational awareness.

Following analyses were performed:

- Knowledge graph and task analysis
- Analysis of conflict detection ML module predictions regarding situations of interest
- Analysis of conflict detection ML module predictions regarding conflicts
- Situational awareness level analysis

The results are presented in the following subsections.

4.4.1 Results of Knowledge Graph and Task Analysis

After the list of the situational awareness indicators explained in Section 3.7.1 had been defined, a comparison of situational awareness of the KG system and ATCO was performed. There are a total of 20 different exercises selected at random, one per every participant in experiment 1. There are four different scenarios from which those exercises had been selected. The comparison of the KG system and human situation awareness was done for each aircraft in a scenario, for the duration of the whole exercise. Indicators for degraded situational awareness presented in the previous section studied through 7 different categories listed below:

- Heading/direct: every situation when an aircraft was cleared on a heading or a direct route
- Level change: every situation when an aircraft was cleared to climb or descend
- Rate: every situation when an aircraft was cleared to change its flight level at a rate
- Speed change: every situation when an aircraft was cleared to fly at cleared Mach number
- Transferred: every situation when an aircraft was transferred to the next sector within the duration of the exercise
- Assumed: every situation when an aircraft was assumed on the label after the initial call
- Military: every situation when an aircraft had a route that will cross a military sector when active

For every aircraft, each of the above-mentioned situations was counted and added to the list of cases of degraded situational awareness if any of the objective indicators explained in Table 11 showed a loss of situation awareness. Figure 21 presents all the situations when the KG system correctly assessed the traffic situation, or the ATCO took action that can prove there was not any degradation of situation awareness. The results from all 20 exercises were summarised. There is a significant discrepancy in the number of situations between some categories. That is due to the fact that some ATCO instructions, such as vertical rate change or speed control, occur only a couple of times in the scenarios, while there are significantly more aircraft that entered the sector and were assumed during the scenarios. As seen in Figure 21, there are many situations when neither ATCO nor the KG System suffered any loss of situational awareness.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Figure 22 shows all the situations where there is proof of degraded situational awareness for the human or the KG System. Note that the scale is different in Figure 21 and Figure 22. As seen in Figure 21, there are not any cases of KG System situation awareness degradations. All the tasks that were used to monitor traffic successfully kept track of all changes in the flight data. Therefore, what is represented in Figure 22 are cases of human situational awareness degradation. There is also some discrepancy in the number of situations in Figure 22. There are not any cases of loss of situation awareness in the "Heading" category, but 12 cases in the "Level change" category because there were intentional non-compliances by the pilot regarding the flight level change, but none regarding the change of heading. The overall number of situations indicating degraded situational awareness in every scenario is shown in Figure 3. E1S1 scenario has the lowest count of situation awareness degradations because it is the shortest scenario with the fewest samples being analysed. In other scenarios, the number of situations indicating reduced situational awareness is approximately equal. E1S3 scenario has a lower number of those situations compared to E1S2 and E1S4 because the planned non-compliance related to Mach number did not occur in every exercise. It was dependant on the ATCO clearance and was therefore not always issued.

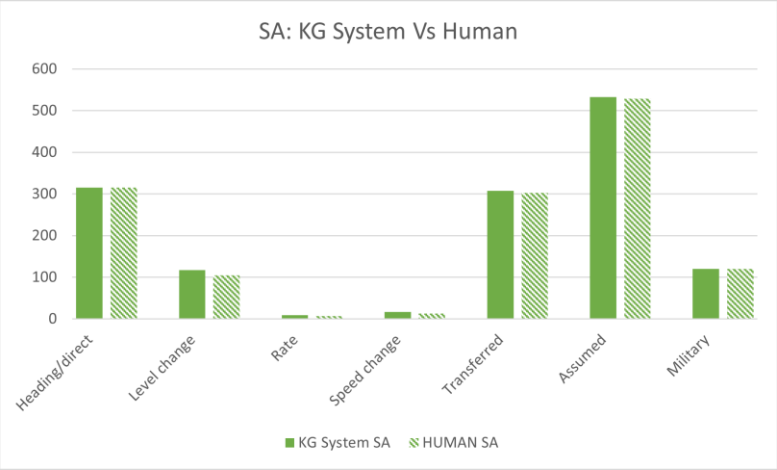


Figure 21: Objective count of occurrences of preserved SA

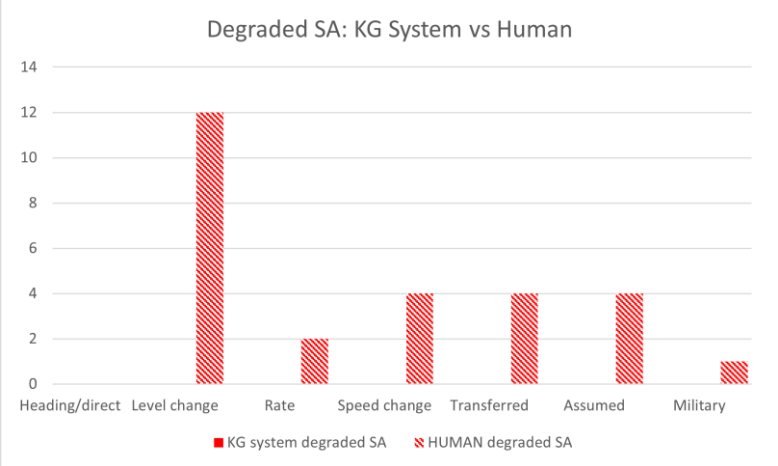


Figure 22: Objective count of occurrences of degraded SA

Co-funded by the European Union

The conflict detection ML module prediction accuracy analysis was performed for one exercise per participant in experiment 1, leading to a number of 20 different exercises. The scenarios are not equally distributed among the 20 chosen exercises; there are 2 exercises based on the E1S1 scenario and the other 18 exercises are divided between E1S2, E1S3 and E1S4, each scenario being used 6 times. In Section 4.4.1, the total distribution of the degraded situation awareness occurrences classified per type is presented. Figure 23 shows the overall number of occurrences of degraded human situation awareness grouped by scenario.

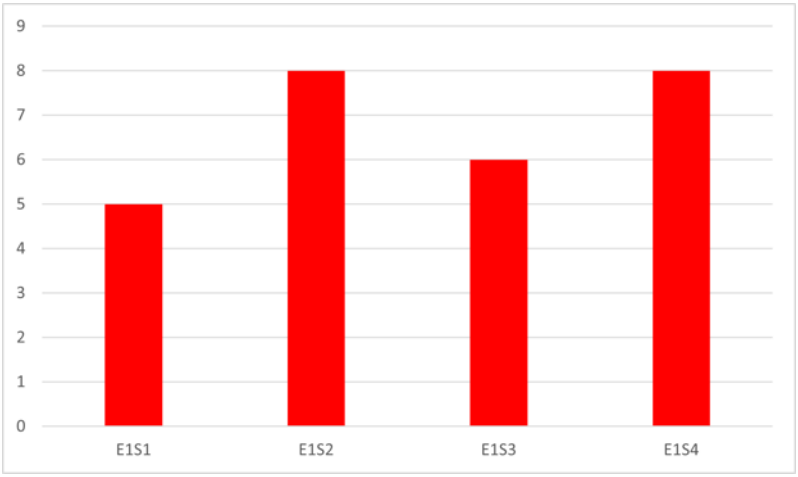


Figure 23: The overall number of occurrences of degraded human SA

4.4.2 Results on Conflict Detection ML Module Predictions Analysis Regarding Situations of Interest

By using two approaches of analysis described in Section 3.7.2, it is possible to recognise those module outputs in which the conflict detection ML module predicts false-negative results (aircraft pairs with actual minimum distance less than 10 NM, for which the conflict detection module predicted a value higher than 10 NM) and false-positive results (aircraft with actual minimum distance more than 10 NM for which the conflict detection module predicted a value lower than 10 NM). False-positive results or Type I error (false alert), and false-negative results or Type II error (missed alert), are presented in Table 20. Type I errors are more common than Type II errors – there are fewer results in which the conflict detection module predicts that the aircraft pair is not a situation of interest than there are results in which the module predicts that the aircraft pair is a situation of interest. True-positive and True-negative contain results where predicted and actual values are equal. Both values are greater than 10 NM for True-negative or both values are less than 10 NM for True-positive results.

Table 20: Type I Error and Type II Error results

		Actual	
		Positive	**Negative**
Predicted	**Positive**	131	110 (Type I error)
	Negative	92 (Type II error)	449

Co-funded by the European Union

The prediction for an aircraft pair can therefore be within Type I Error, Type II Error, or True Positive/Negative. The exceptions that are rarely present are shown in Figure 24. *"NaN"* results are aircraft pairs that the conflict detection module predicted as a situation of interest but at an incorrect time, e.g., when the conflict module prediction is made at a time when the aircraft have already passed the point of the actual minimum distance between them. The presence of the "*NaN*" results at the time when the aircraft pair is at the minimum distance is a result of the "black-box" effect, an effect which is a product of ML algorithm that is impenetrable and cannot be straightforwardly defined. Diverging traffic output – when the conflict detection module predicts a situation of interest when the distance between aircraft is increasing – is the least represented in the results.

From the graph in Figure 24 it can be recognised that the percentage of aircraft pairs that are initially or finally predicted as a situation of interest (True-positive) and aircraft pairs predicted as a situation of no interest (True-negative) by both ATCO and module are the most represented. Type I Error and Type II Error account for less than one third of the analysis results.

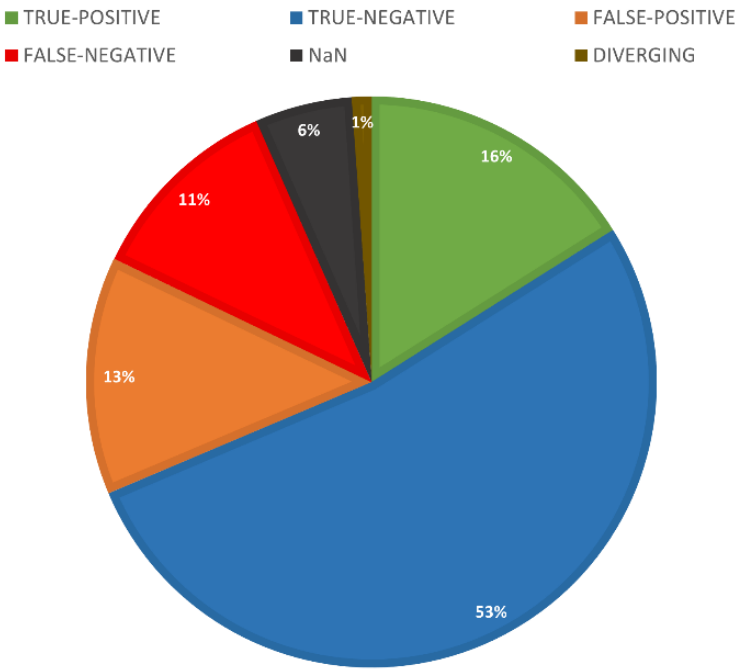


Figure 24: Distribution of the initial and final prediction analysis results

Multiple correlation analysis used to prove a correlation between the ML module training data and the predicted minimum distance accuracy showed that these variables are not statistically related. The results are presented in Appendix K.

4.4.3 Results on Conflict Detection ML Module Predictions Analysis Regarding Conflicts

For all conflicts that are initially implemented in scenarios or those caused by the action of ATCO, it is possible to assess the performance of human and conflict detection module. In Section 3.7.3 Conflict

EUROPEAN PARTNERSHIP
Co-funded by the European Union

detection ML module analysis the methodology of the analysis is introduced as well as the difference in defining the situation of interest and conflict.

For each exercise, aircraft pairs which are in conflict were selected and the following was checked:

- did the ATCO recognise the conflict and acted promptly on it (human accurate),
- did the ATCO omit conflict that led to the loss of separation (human inaccurate).

After evaluating what the ATCO actions were, it was checked and compared what the conflict detection module outputs were for the same conflicts. The following was checked:

- did the conflict detection ML module accurately predict conflict (CD module accurate),
- did the conflict detection ML module predict conflict but with inconsistent values (CD module inconsistent),
- did the conflict detection ML module omit conflict (CD module inaccurate).

The comparison of ATCO and CD module performance per scenario is depicted in Figure 25. Accurate CD module predictions recognised the conflict and changed the predicted values as ATCO resolved the conflict. Inaccurate CD module predictions recognised the conflict, but values did not change as ATCO resolved the conflict. Inaccurate CD module predictions did not recognise conflict or predicted values did not match at all.

With the E1S3 scenario having the most conflict, the most human and CD module accurate predictions are present in that scenario. Vice versa, E1S1 scenario having the least conflicts also has the least human and CD module accurate predictions. For the CD performance analysis, it is important to emphasise that column height indicates only the total number of occurrences in the observed scenario. To check what the ratio of accurate, inaccurate, and inconsistent occurrences is, it is necessary to look at the colour distribution. For example, the number of inaccurate and inconsistent predictions in E1S4 scenario is equal to the accurate predictions. E1S4 scenario has the highest number of ATCO omissions which is explained by it being the longest scenario and with the TRAs being activated and deactivated.

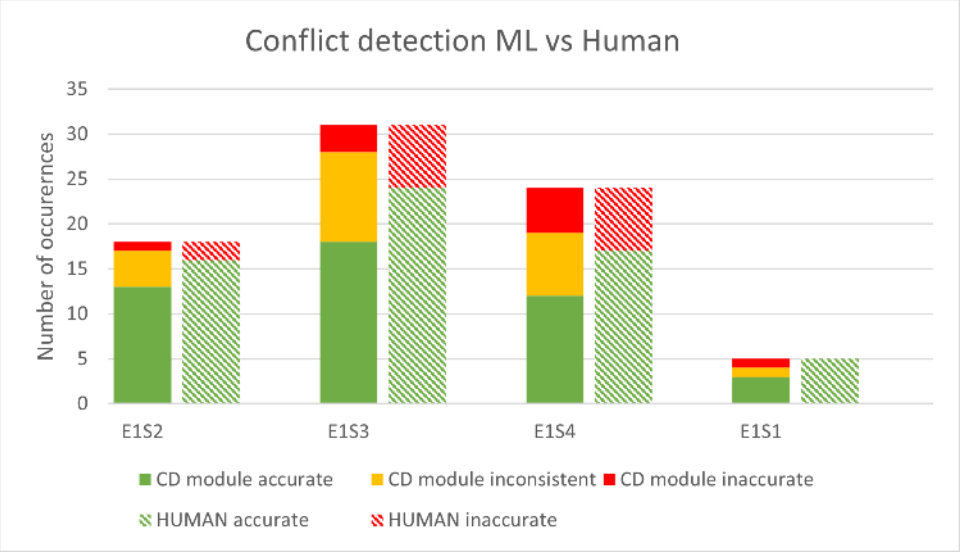


Figure 25: Comparison of conflict detection ML and human performance

EUROPEAN PARTNERSHIP
Co-funded by the European Union

4.4.4 Results on Levels of Situational Awareness

AISA ConOps proposed the use of an existing framework for AI system awareness level assessment. The framework, described in the "Theory" chapter of this document (see section 2.2.2), defines a set of 7 conditions whose fulfilment places the AI system into one of 6 awareness level categories (with Awareness Level 0 denoting a system with low awareness and Awareness Level 5 a system with highest awareness). The conditions are divided into conditions for awareness of a certain property and for awareness of the system itself (or self-awareness), while no subdivisions are defined for the awareness levels.

For the purposes of classification of the AI SA KG system, a border is set between the AI SA KG system (consisting of a KG, AI SA tasks and ML modules) and the simulator on the input side and the ATCO on the output side. Conditions defined in the framework and listed earlier in this document will be analysed from the perspective of the system and exemplified. The fulfilment of these conditions will then be used to classify the AI SA KG system on the awareness level scale.

For each condition in Table 21 and level requirement in Table 22, only the best performing sub-system was chosen to represent the system. This aims to show that, while not all parts of each awareness level are fulfilled to the same degree and by the whole system, most conditions and requirements presented in these tables can be fulfilled by the AI SA KG system.

Table 21: AI SA condition fulfilment estimate

Condition Code	Condition Description	AI SA KG System Function
(C.1)	Subject makes physical measurements or observations that are used to derive the values of property *P* by means of a meaningful semantic interpretation.	Measurements and observations from which values of a property can be derived are gathered from the various data sources the AI SA system has at its disposal. For example, the simulator delivers the values of flight information such as flight level, speed, position and others. A meaningful semantic interpretation consists of mapping gathered measurements to values of a property and choosing the appropriate interpretation if mapping results in more than one interpretation. Since the conversion of values (from data files to RDF graphs) maps values to properties directly, only a single interpretation is possible. The properties (flight level, speed, position, flight etc.) are, of course, meaningful in the context of ATC, so all parts of the (C.1) condition are fulfilled.
(C.2)	The semantic interpretation is robust.	The robustness of semantic interpretation is the task of SHACL rules. In case of faulty inputs (originating, in this case, from either human error or the program tasked with converting data to RDF format), the system returns an error and points to where the error occurred and why. SHACL rules may only detect some, instead of all, erroneous inputs. Therefore, robust semantic interpretation is accomplished, but not fully guaranteed.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

(C.3)	There is a semantic attribution which is meaningful.	Semantic attribution, the process of mapping values of a property to a desirability scale, is performed by AI SA tasks by comparing the values to those defined by the goals (e.g., cleared values). Values are then implicitly graded as desirable/equal to cleared or not desirable. An example of a system property being checked for desirability is the system's inspection of the conflict prediction ML module – it is checked for desirability of input data by comparing it to the statistics of the module's training set. Not all values are mapped to a desirability scale – we deem this to be acceptable because desirability (beyond the base test that are the SHACL rules, which are already applied to all properties) cannot be established for properties such as callsigns or statistical values of conflict ML module training data.
(C.4)	The subject's reaction to its perception of *P* is appropriate.	The AI SA KG system achieves appropriate reaction to the perception of properties by • analysing and storing property values • using property values for creating other properties and computing their values • creating appropriate outputs for property values
(C.5)	A history of the evolution of the property over time is maintained, in particular of the increasing or decreasing deviations over time.	History of evolution of each property is easily accessible since each situation graph is stored in the KG, along with output graphs created by each task. Increasing or decreasing deviations can be tracked through task outputs – they are implicit in outputs such as "Aircraft is not at CLFL." and "Aircraft is descending towards CLFL."
(C.6)	The subject can assess how well it meets all is goals, thus having an understanding which goals should be achieved and to which extent they are achieved.	Goals of the AI SA KG system are represented in the KG via cleared values. Coupled with the AI SA tasks, the KG can state which goals are achieved (e.g. "Aircraft is at cleared speed.") or are currently being achieved (e.g. "Aircraft is climbing towards CLFL.").
(C.7)	The subject can assess how well the goals are achieved over time and when its performance is improving or deteriorating.	The AI SA KG system runs all tasks and can, by analysing the outputs, check the status of each goal and its changes through the scenario. The storing of task outputs ensures goal completion can be assessed over time. Tasks related to the operation of the conflict detection ML module monitor both the status of each conflict (which are some of the goals of the system) and the performance of the module itself (the correctness of each prediction).

As shown in Table 4, the framework recognised 6 levels of AI system awareness. The table is repeated here, modified to show how the extended AI SA KG system fulfils the requirements of each level.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Table 22: AI SA awareness level estimate

Awareness Level	Necessary Conditions to reach Level	AI SA KG System Function
Awareness Level 0	System output is a mathematical function of inputs (always reacting in the same way to inputs) System fulfils conditions (C.1) to (C.4)	The AI SA KG system consists of computer code which, for identical inputs, always produces the same output. Conditions (C.1), (C.3), (C.4) are shown to be fulfilled by the AI SA KG system in Table 20. Condition (C.2) is partially fulfilled (since it's not guaranteed) so Awareness Level 0 requirements can be thought of as partially fulfilled as well.
Awareness Level 1	System is adaptive, meaning that it tries to minimize the difference between input and reference values by use of a PID controller or similar algorithm System fulfils conditions (C.1) to (C.4)	The AI SA KG system fulfils the adaptiveness condition through the outputs of the KG system – by having the outputs point toward the difference between actual and goal values, the system affects the actions of the ATCO, thus ensuring the minimisation of the differences. Conditions (C.1), (C.3), (C.4) are shown to be fulfilled by the AI SA KG system in Table 20. Condition (C.2) is partially fulfilled (since it's not guaranteed) so Awareness Level 1 requirements can be thought of as partially fulfilled as well.
Awareness Level 2	System is aware of at least one (system) property and one environment property according to (C.1) to (C.4) + (C.6) System contains an inspection engine which periodically derives one integrated attribution of the system as a whole System computes its actions based on (a) monitored and attributed properties of the system and of the environment, (b) attributed expectations on the system and on the environment, and (c) sets of goals on system and environment properties	The AI SA KG system is aware of both environment properties (such as aircraft trajectories) and system properties (such as conflict detection module performance) in ways prescribed by conditions necessary for this level. Apart from condition (C.2), which is fulfilled partially as described in the previous table, the conditions are fulfilled. The inspection engine condition is fulfilled by the conflict detection module – tasks which check the desirability of module inputs (against training data statistics) and outputs (by way of the "sanity check" and basic comparison calculations) are a way for the system to analyse itself. The AI SA KG system does compute necessary actions according to the values of properties defined in the KG (such as the already mentioned aircraft FLs or conflict detection module performance), the expectations on itself and the environment (which are defined by the SHACL rules and completeness of the KG) and goals (which are contained in the KG). Since the expectations on the system are contingent on the functioning of SHACL rules, this condition and awareness level cannot be guaranteed to be fulfilled. For this reason, Awareness Level 2 is reached partially.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Awareness Level 3	System fulfils all requirements of an Awareness Level 2 system System fulfils the history conditions (C.5) and (C.7)	The history conditions are fulfilled as demonstrated in the "Conditions" table – each timestamp's traffic data and task output graphs are stored in the KG and easily accessible. Combined, they form a history of each property and property value where values are direct proof of deviations. The improvement and deterioration are demonstrated only for appropriate properties – e.g. conflict detection module performance. The fulfilment of Awareness Level 2 system requirements is shown in the cell above. Since the previous level is reached only partially (because fulfilment of some conditions cannot be guaranteed), Awareness Level 3 is also reached partially.
Awareness Level 4	System fulfils all requirements of an Awareness Level 3 system System's decision-making process involves a simulation engine which can predict the effects of actions on the environment and the system itself and, in case of an anomalous result, can search through simulations for the best action	Simulation engine requirement is completed by the machine learning modules. They can use each traffic data graph as input and calculate how modifications of certain property values can lead to different traffic outcomes. A voluntary number of repetitions (with unique value modifications) can be performed, and the results parsed for the optimal action (or actions). The fulfilment of Awareness Level 3 requirements is shown in the cell above. In the same manner, Awareness Level 4 is deemed to be reached partially.
Awareness Level 5	In addition to being self-aware, the system distinguishes between itself, the environment, and the peer group (which is treated differently because of its own set of expectations and goals)	The AI SA KG system contains tasks dealing with environment properties, dealing with system properties, but also with properties formed by third parties (such as sector exit flight levels, dictated by agreements with neighbouring air navigation service providers). Those providers can thus be seen as a peer group with specific goals, whose existence is recognised by the KG. As with previous levels, Awareness Level 5 is conditional on the functioning of SHACL rules and (C.2), so it can be considered partially completed.

According to the framework defined by Jantsch and Tammemäe (2014b), the AI SA KG system is conditionally an Awareness Level 5 system. This conclusion hinges on the current method of checking if the system inputs – the SHACL rules. If their functioning is bolstered by the implementation of another layer of checks, the estimation provided in this chapter could be confirmed. Future system architecture could also improve upon or replace sub-systems which only fulfil the awareness requirements partially, so the future system can be more accurately assigned a higher level of awareness.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

4.4.5 Discussion on Results of Accomplishment of Artificial Situational Awareness

The artificial situational awareness resulting from the monitoring tasks applied to the traffic data was compared to the ATCO situational awareness on the basis of accuracy. For the purpose of comparison, a group of four objective indicators was assessed. The results show that the KG system is able to make error-free assumptions about the traffic situation, regardless of the type of task at hand. Furthermore, machine situational awareness is formed instantly upon receiving input data, whereas the human participants have a time buffer of 30 seconds in which they are expected to notice any changes on the radar screen that requires their attention.

The conflict detection machine learning module presents minimum distances for all aircraft pairs for which the minimum distances are predicted to be below 25 NM. The situations of interest are considered to be those predictions where the minimum distance is below 12 NM. A comparison of accuracy for conflict detection ML module outputs has been made in two stages: initial prediction before any ATCO inputs and the final prediction after all ATCO inputs were both compared to actual measured minimum distances. The results were classified regarding their accuracy, with the 12NM being the limit. There are 16% true-positive, 54% true-negative, 11% of false-negative (Type I error) and 13% of false-positive (Type II error) predictions.

To analyse if the conflict detection module predictions change for the aircraft pairs whose minimum distance would violate separation minima, the comparison between CD module predictions and human actions was performed. Those aircraft pairs were observed and categorised based on the module performance. There is a significant number of aircraft pairs that are recognised late. These inconsistent predictions also do not follow ATCO conflict resolution actions. Inaccurate predictions are those unrecognised or recognised with false minimum distances. In each exercise in the same scenario (different ATCOs working on the same scenario), this occurred for the same aircraft pair. Therefore, conflict detection module inaccurately predicts aircraft pairs regardless of the ATCO changes on those flights.

The framework chosen in the AISA Concept of Operations presents 7 conditions for (self-)awareness of a system. Combined with specific system functions or sub-systems, they form the requirements for 6 awareness levels. AI SA has been shown to fulfil all 7 basic conditions and additional requirements, which means the AI SA KG system is an Awareness Level 5 or a Group-aware AI system. Most of the results are not PoC-system-specific, so they will be applicable to the future AI SA system as well.

4.5 Results on AI SA System Performance

A short analysis of the AI SA KG system performance was done to determine real-time application feasibility. As was already posited, many improvements are possible and needed for the proof-of-concept – this analysis only establishes the current status so the influence of improvements can be better measured. The chosen analysis parameter, AI SA KG system runtimes, were measured using the full set of completed tasks.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Since the chosen exercises from experiment 1 were converted into a different number of RDF graphs (ranging from 14 to 58), the full runtime of exercises could not be directly compared, which is why a mean runtime was calculated for each exercise. Runtime per graph in relation to number of graphs in the exercise is demonstrated in Figure 26.

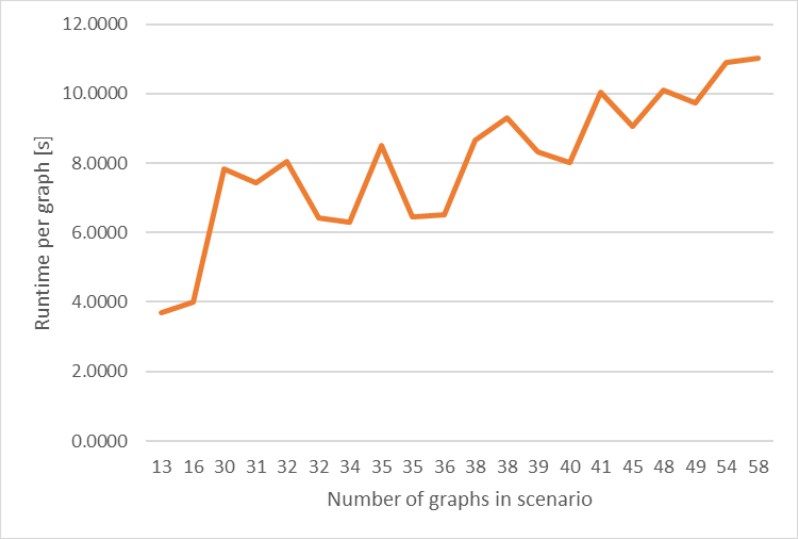


Figure 26: Runtime per graph vs Number of graphs in scenario

Since exercises were chosen from one of four scenarios of differing complexity, the analysis was also done to see how the maximum number of aircraft in the exercise might have affected the number of queries and thus the runtime. Since most exercises stemming from the same initial scenario share the same maximum number of aircraft, “median runtime per graph vs number of aircraft” (shown in Figure 27) and “median runtime per graph vs scenario set” (shown in Figure 28) are remarkably similar.

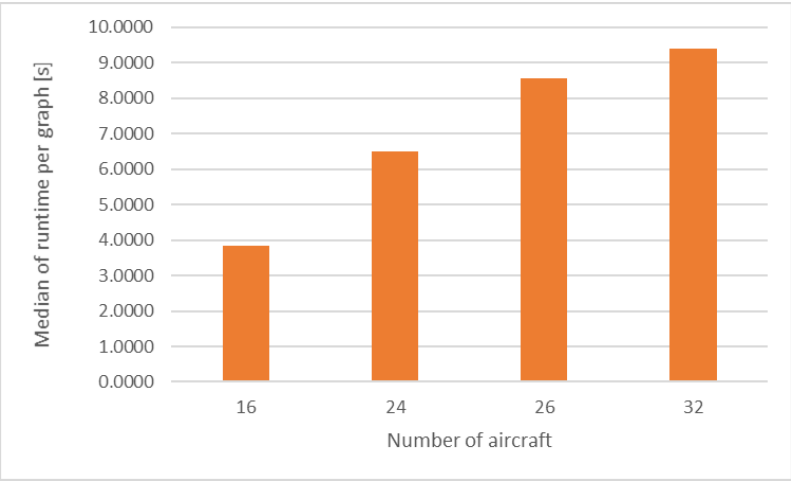


Figure 27: Median of runtime per graph for number of aircraft

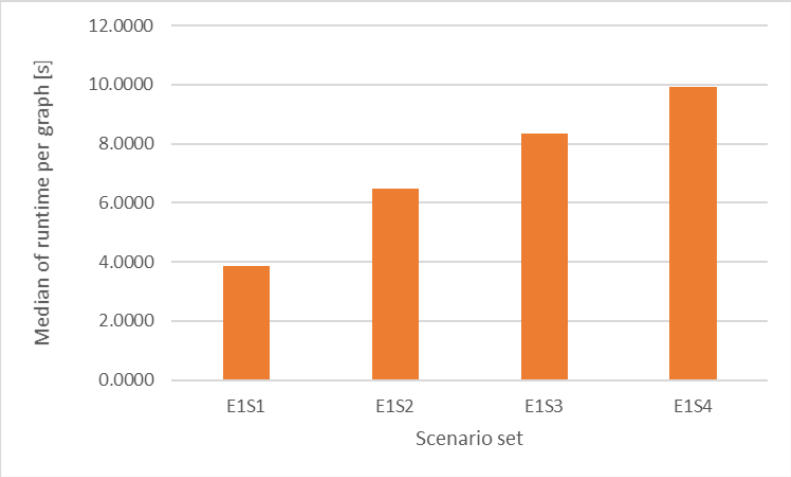


Figure 28: Median of runtime per graph for scenarios

Discussion: The PoC system is not able, but also not meant to operate in real-time. Analysis provided in this sub-chapter shows how quickly the system can process a single graph, depending on various parameters–these times range from under 4 s to over 10 s. These results can be interpreted in several ways, depending on which actual ATC systems are used for comparison:

- Aircraft ADS-B data is sent every 1 s – while it's not guaranteed to be received every second, it would be ideal if the future system could process the data as quickly as it is sent. For this (idealised) purpose, the PoC system is too slow.
- ATCO station refresh rate is every 5 s – the location and label data are updated together. Since this is the rate at which the ATCOs receive information, and since team members use the same data to achieve situational awareness, this is also the realistic rate with which to compare the AI SA system processing times. Depending on the number of aircraft (or, more accurately, the amount of data stored in each RDF graph, which is mostly dependent on the number of aircraft), the system processing times are in the appropriate range for real-time application.

The processing times described here are dependent on multiple factors, the influence of which must be further analysed if the system is developed for real-time use. One such factor is the architecture of the computer on which the system is running – ATM service providers have access to far more processing power than a single laptop. Other factors might be traffic, sector configuration, historic data storage strategies and others.

4.6 Robustness and Generalisability of the AI SA System

Presented below are two separate analysis approaches to measure the robustness and generalisability of the AI SA system.

4.6.1 Independence of the Conflict Detection ML Module Predictions Regarding Situations of Interest

In Section 3.7.2, it is described how aircraft pairs that were considered as a situation of interest were analysed based on the initial and final conflict detection ML module prediction. The results of the analysis are presented in Section 4.2.4. To measure if the conflict detection ML module prediction

EUROPEAN PARTNERSHIP
Co-funded by the European Union

accuracy is independent, the limit by which the predicted conflict is considered as a situation of interest is changed. If reducing the limit from 10 NM to 7.5 NM, the number of False-positive and True-positive results is reduced in the favour of True-negative results. If the limit is increased to 12.5 NM, the number of True-negative results is reduced in the favour of True-positive and False-positive results. Table 23 and Table 24 present how many results of each error group occurred.

Table 23: Type I Error and Type II Error results (7.5NM)

		Actual	
Predicted		**Positive**	**Negative**
	Positive	65	45 (Type I error)
	Negative	91 (Type II error)	561

Table 24: Type I Error and Type II Error results (12.5NM)

		Actual	
Predicted		**Positive**	**Negative**
	Positive	201	168 (Type I error)
	Negative	102 (Type II error)	291

The distribution of the results is shown in Figure 29 and Figure 30 . The overall number of the "True results" (True-positive + True-negative) reduces for the 12.5 NM limit and increases for the 7.5 NM limit with the respect to 10 NM situation of interest limit. False-negative results diverge the least when changing the limit for the situation of interest. Therefore, changing the limit for the situation of interest does not affect the Type I Error results.

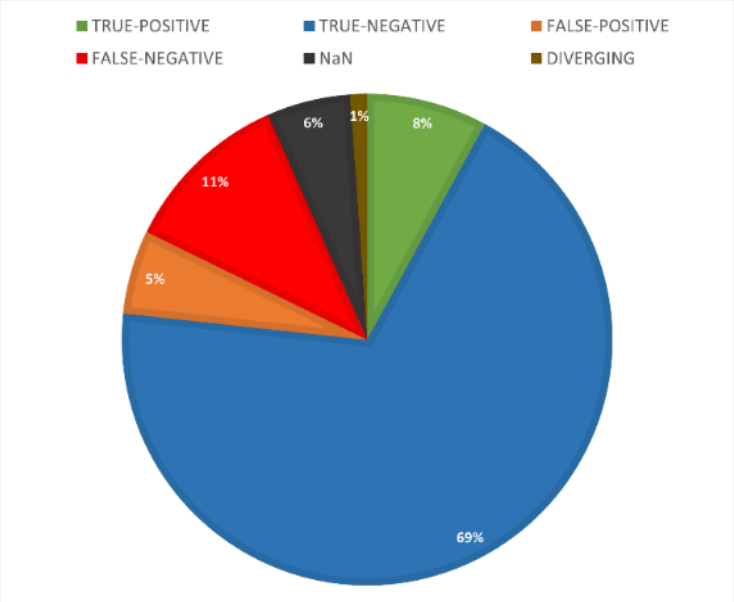


Figure 29: Distribution of the initial and final prediction analysis results (7.5NM)

EUROPEAN PARTNERSHIP
Co-funded by the European Union

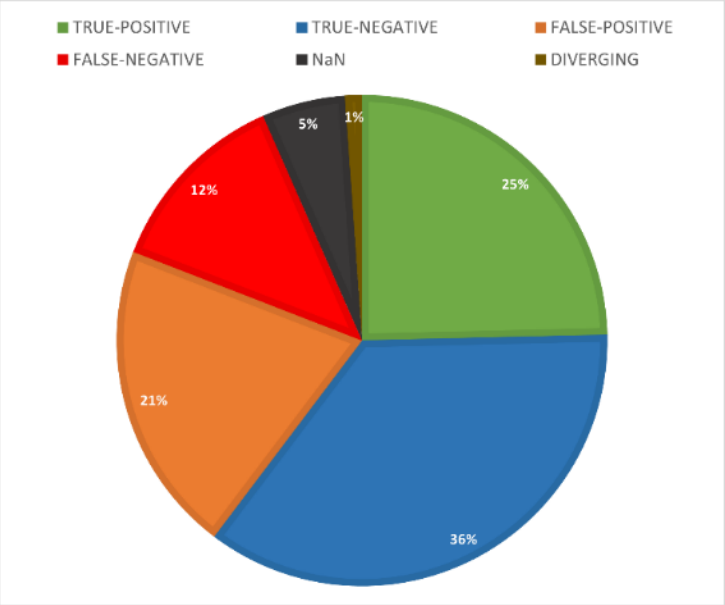


Figure 30: Distribution of the initial and final prediction analysis results (12.5NM)

Table 25 summarizes the results for the 3 analysed cases, with the difference being the chosen limit for the situation of interest.

Table 25: Summary of Type I error and Type II error results

Situation of interest limit	TRUE POSITIVE [%]	TRUE NEGATIVE [%]	FALSE POSITIVE [%]	FALSE NEGATIVE [%]
7.5 NM	8	69	5	11
10 NM	16	53	13	11
12.5 NM	25	36	21	12

4.6.2 Independence of the Knowledge Graph and Task Analysis Accuracy from the Scenario

One of the goals for the AI SA system is for it to be equally effective regardless of the traffic situation at hand. In Section 4.4.1, the overall comparison of the machine and human situation awareness was described. The graphs will focus on the consistency of the machine situation awareness throughout different scenarios, but limited to monitoring tasks only, with no conflict prediction.

The independence of the KG and task analysis accuracy from the scenario can be seen from the fact that there are no cases of degraded KG system situational awareness in any of the scenarios, regardless the fact that each contains a different traffic situation.

Figure 31 presents the objective measure of KG system and human situational awareness for all analysed E1S1 scenarios. Likewise, Figure 32 presents the results for the E1S2 scenario, Figure 33 the E1S3 scenario and Figure 34 the E1S4 scenario.

Both left- and right-hand graphs count all the situations in those scenarios where there were changes in the traffic data concerning either heading, flight level, rate, speed, or the state of the aircraft regarding it being assumed or transferred to the next frequency, which is the same approach that was already described in Section 4.4.1. The left-hand side in the following figures shows how many times those changes happened in the scenarios and have been correctly recognised by both the ATCO and the KG system. The right-hand side shows how many times the ATCO or the KG system did not recognise those changes or did not act on them, leading to a degradation in situational awareness.

What can be seen from the following graphs is that there are different areas of focus in each of the scenarios. For example, in the E1S1 and E1S2 scenario, the highest count of situation awareness degradations falls under the "Level change" category since those scenarios include a pseudo-pilot non-compliance regarding the level instructions. Likewise, the E1S3 scenario accounts for all cases of situation awareness degradation in the category of speed change because only that scenario contains a speed non-compliance. Considering the description of all the scenarios in Section (experiment 1), these results are expected.

The independence of the KG and task analysis accuracy from the scenario can be seen from the fact that there are no cases of degraded KG system situational awareness in any of the scenarios, regardless the fact that each contains a different traffic situation.

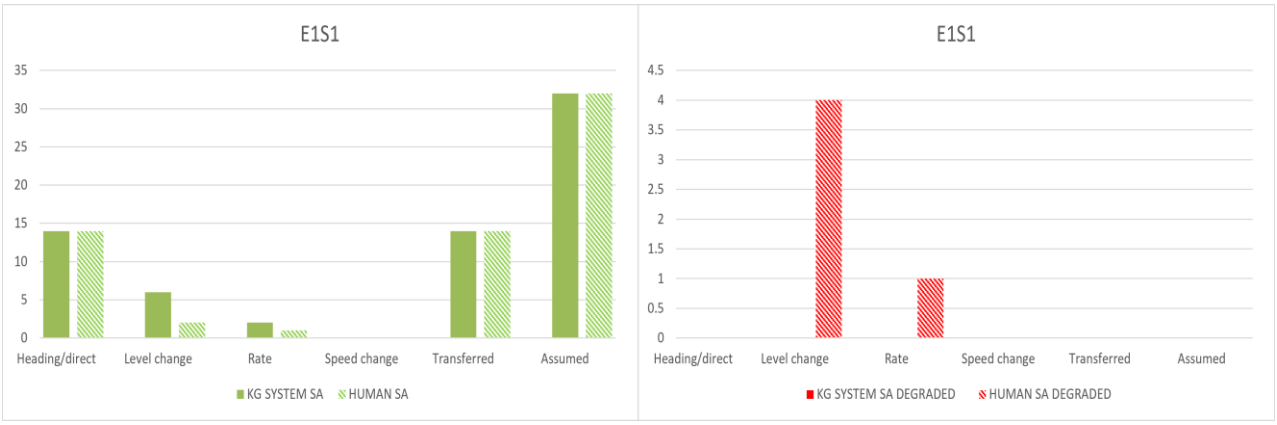


Figure 31: Objective count of occurrences of preserved and degraded situation awareness (E1S1)

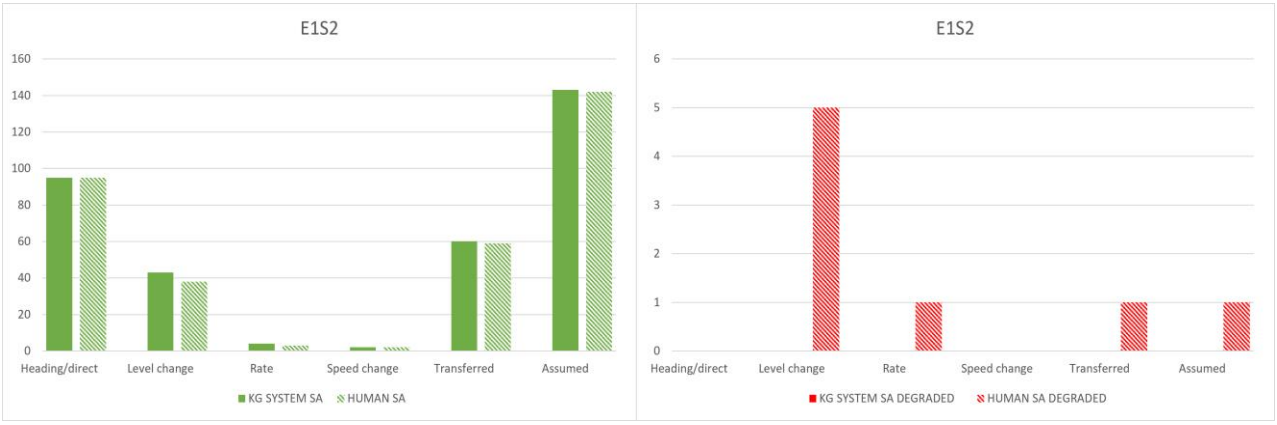


Figure 32: Objective count of occurrences of preserved and degraded situation awareness (E1S2)

EUROPEAN PARTNERSHIP
Co-funded by the European Union

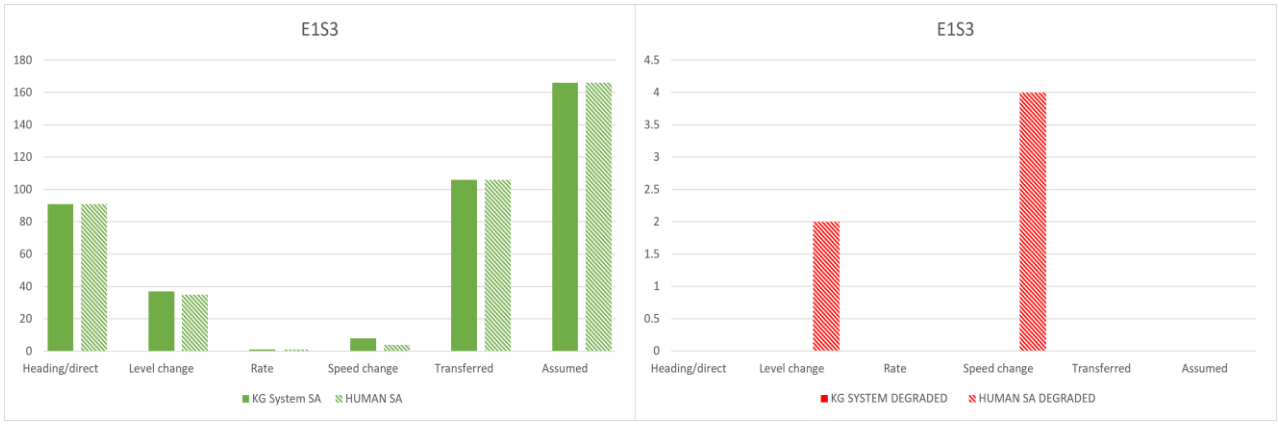


Figure 33: Objective count of occurrences of preserved and degraded situation awareness (E1S3)

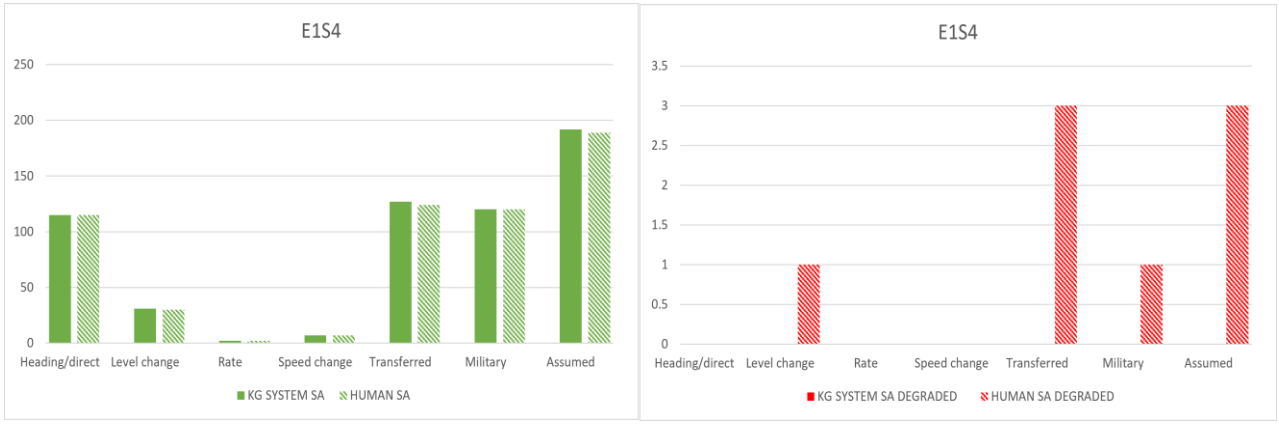


Figure 34: Objective count of occurrences of preserved and degraded situation awareness (E1S4)

EUROPEAN PARTNERSHIP
Co-funded by the European Union

5 Discussion

The main goal of the evaluation in D5.2 is to determine if the concept of human-machine team situation awareness is feasible and what level of accuracy artificial situation awareness can reach today at the project-level of implementation. Because of the early stage of the concept-of-implementation during the human-in-the-loop experiments in January 2022, those tests are of exploratory nature. The AI SA system at the current project-level state of implementation in April 2022 includes a more comprehensive list of monitoring tasks that AI SA system is able to perform (8 tasks in January 2021 compared to 46 out of 57 tasks in April 2022). Quantification of accuracy and analyses of the functionality of the AI SA system are performed with the comprehensive implementation of monitoring tasks.

According to the ConOps (D2.1) the project level of implementation of the AI SA system does not require it to work in real-time. A stepwise approach was chosen. A first human -in-the-loop simulation was done to collect data from 20 ATCOs. Data from simulator sessions were imported into the AI SA system and processed with specific SPARKL queries designed to analyse the correctness of the output of the AI SA system to them and to compare them later with ATCOs answers to the same queries. Experiment 2 included 16 ATCOs who followed in four of five scenarios another ATCO's performance (from experiment 1) in a "watch only" simulation, with the addition of AI SA inputs. One scenario was interactive and allowed participants to be in charge for traffic handling. During the experiment participants were asked the same queries as were processed by AI SA system before. After they answered to those queries, they immediately received the AI SA inputs to the respective query. To guarantee anonymity, voices of ATCO and pseudo-pilot were transformed and reproduced by synthetic voices. "Watch-only" simulations with synthetic voices credibly represented pilot-controller communication. Since the AI SA inputs were also transmitted via audio, individual voice transmissions either had to be accelerated or were not identical to the original simulation transmission. To distinguish between original communication and AI SA inputs, a short chime was used to allow ATCOs to focus more on those audio recordings. A different chime was used to alert ATCOs that the exercise would be frozen and unfrozen, at which time they completed query answers. To reduce the confusion, an example of each type of the chimes, along with examples of the pilot and ATCO voices was played to the participants before the training exercise started. Originally in experiment 1, ATCOs used Datalink for some instructions or clearances. Those clearances also popped out on the screen increasing the workload for the ATCOs in experiment 2.

The main findings for the research questions are summarised in this section and can be overviewed in Table 26.

Table 26 Overview: Research questions and summary of results

Research question	Result	Section
Q1.1. What characterises ATCOs' scanning patterns and priorities?	ATCOs with preserved situation awareness scanned regularly, did not fix their gaze on aircraft or conflicts and applied prioritisation for important and unimportant aircraft.	4.1.3.2 Comparison of ATCO Groups for Gaze-Based Analysis of

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Q1.2. Are different measures for situation awareness[3] significantly interrelated according to their meaning?	Moderate to high consistencies across methods, lowest for subjective rating of situation awareness.	4.1.2 Correlational Results on Situation Awareness Measurement Methods
Q2.1. Are artificial and ATCO situation awareness comparable?	Some agreement on conflict detection between ATCOs and AI SA system. Conflicts missed by either ATCO or AI SA system. AI SA system unable to estimate right time for AI SA inputs and prioritisation.	4.2 Results on Comparison of Human and Machine Situation Awareness
Q2.2. Can the AI SA system provide inputs to situation awareness that ATCOs were not aware of?	Yes. Inputs particularly important for inconspicuous conflicts and non-conformances.	4.2 Results on Comparison of Human and Machine Situation Awareness
Q3.1. Is human performance enhanced by adding machine situation awareness?	Performance enhanced, despite inadequate input modality creating distraction: some conflicts detected earlier and solved faster, others not.	4.3.1 Evaluation of ATCO's Performance Based on Behavioural Coding
Q3.2. Do ATCOs evaluate AI SA inputs as useful and trustworthy contribution to human-machine team situation awareness?	Some AI SA inputs judged helpful, but most inputs considered irrelevant. Inputs partially trusted. AI SA system not yet contributing sufficiently to human-machine team situation awareness. ATCOs willing to trust if system is reliable.	4.3.2 Evaluation of Artificial Situation Awareness Based on Questionnaire Answers
Q3.3. Do ATCOs use AI SA inputs for their situation awareness and decision making?	ATCOs did not validate the AI SA inputs as supportive for their own situational awareness (12.8% rated inputs as supportive) or decision making (8.8% rated inputs as supportive).	4.3.2 Evaluation of Artificial Situation Awareness Based on Questionnaire Answers
Q4.1. Can the monitoring tasks be applied to the KG to achieve situational awareness?	Monitoring tasks have been successfully automated and applied to KG data. Task outputs demonstrate that the system does achieve situational awareness.	4.4.1 Results of Knowledge Graph and Task Analysis
Q4.2. Does the CD machine learning module provide accurate results regarding situations of interest?	CD machine learning module provides 70% accurate predictions compared to the 10 NM miles SI limit.	4.4.2 Results on Conflict Detection ML Module Predictions Analysis Regarding Situations of Interest

[3] situation awareness self-ratings, situation awareness queries, situation awareness based on eye-tracking, implicit measurements of situation awareness

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Q.4.3. Does the CD machine learning module provide accurate results regarding conflicts?	CD machine learning module provides partly accurate results regarding conflicts; there is significant number of inaccurate and inconsistent predictions.	4.4.3 Results on Conflict Detection ML Module Predictions Analysis Regarding Conflicts
Q.4.4. Does the AI SA system check the status of its sub-systems?	CD ML module tasks successfully perform checks on CD module inputs and outputs, thus allowing the AI SA system to have self-awareness regarding its sub-systems.	4.5 Results on AI SA System Performance

The following sections summarise the methodological limitations of the experiments for evaluation of human-machine team situation awareness, the conclusions, implications and provide an outlook.

5.1 Methodological Aspects and Limitations

In the following sub-chapters, limitations are discussed for their potential impact on the results and for consequences regarding the execution of work package 5 with special focus on task 5.1 Comparison of SA between AI and ATCO and task 5.3 Human performance in distributed situation awareness.

5.1.1 Experimental Design

Despite the planned activities of the Grant Agreement to assess the human performance while the participants look at static traffic situations, the ZHAW decided in agreement with the project leader FTTS that human-in-the-loop simulations will be conducted. This major change brought not only benefits with it. Similarity to the working environment was significantly increased using of an interactive simulation tool, since ATCOs are used to work at a working station with dynamically changing traffic situations on monitors and radio communication as an important source of information and communication.

The concept of operation specifies the requirements for system operation of AI SA system. However, the concept was at an exploratory stage during work package 5 evaluation experiments. Therefore, the proof-of-concept system developed corresponds to a low level of technology readiness (TRL 1 or 2) and the evaluation of human-machine team situation awareness is thus very preliminary. No HMI was developed to provide AI SA input. Instead, ATCOs received inputs on the auditory channel. This limits the conclusions on the results, as the choice of modality might have created additional load, because the auditory channel is used extensively in todays' ATC tasks. This could have made it difficult for ATCOs to include the artificial situation awareness in their situation awareness. If the AI SA inputs were communicated visually, the simulation handling would have become more user-friendly. ATCOs would have the possibility to check the inputs selectively and more often adjusted to their needs and with minimal cognitive load. In addition, it would not be distracting from current tasks – and hence disturbing–if the inputs were shown early, offering long-term anticipation to ATCOs. This issue would be less relevant if there was planner controller during experiment 1 and experiment 2. One of his working position tasks is to identify inbound traffic conflicts. AI SA input with large anticipation spans could optimise planning tasks. The ATCOs could then decide for themselves when they want to use

EUROPEAN PARTNERSHIP
Co-funded by the European Union

those inputs. Therefore, the development of an appropriate human machine interface is essential to the effectiveness of the AI SA system.

The project team is aware that AI SA system is intended for a future working environment where communication with flight crew is done by means of Controller Pilot Data Link Communications (CPDLC). Although datalink could be used for issuing clearances, 94% of all ATCO-pilot interactions occurred via radio transmission in the experiments.

The experimental settings in both experiments might have imposed additional limitations. On the one hand, the parallel running of two simulation stations in the same room can be distracting. Up to eight people were in the experimental room and were sometimes chatting amongst themselves. However, this is not unusual for ACC work environment. ATCOs are not always completely silent, especially during shift changes or chats between ATCOs. Taking this into account, data logging in such a context can still be seen suitable.

The manner how queries were asked during simulation and the lack of clear definitions for “conflict” and “non-conformance" in the queries made it difficult to ATCOs to answer. It turned out that the query wording “what do you need to pay attention to?” was inappropriate since any answers could fit. ATCOs have a different understanding of how a crossing is defined. The term non-conformance was new to some of the ATCOs. These findings would need to be addressed in a further similar study.

Lastly, there is to mention that sometimes the radio transmissions were not fully understandable due to technical hardware issues (i.e., head- and micro-phones). This small issue in this context of exploratory research can influence the results to a certain extent but is not as dangerous as in the real working environment. Nevertheless, it should not be neglected that some information might have been lost when studying the results.

5.1.2 Simulation Software

The simulation software used in the experiments, ESCAPE Light, is a platform the participants never used before. It has significant differences compared to Skyvisu, a software developed and implemented by Skyguide. The switch to a new environment was one of the biggest limitations for the participants. The majority of the ATCOs expressed that they had difficulties with the differences of the system during the simulation or in the debriefing questionnaire. This should be kept in mind in view of validity of the results on ATCO situation awareness presented in this deliverable.

5.1.3 Artificial Situation Awareness

Generating AI SA inputs is limited by the need for post-processing of AI SA KG system outputs. This means that it was not possible to generate the inputs for human-machine situation awareness in real time. The “watch-only” scenarios in experiment 2, which led to a trade-off, result of this limitation. They allowed for comparison of human and artificial situation awareness, but at the cost of lower validity for ATCO situation awareness. Because participants were not acting themselves upon their own situation awareness but had to reproduce the actions of another ATCO. Since AISA is in an exploratory stage which is also clearly indicated by its technology readiness level (i.e., TRL 1 or 2) and not meant to be a finalised product, the post-processing is not contradictory to the project goals.

It was challenging to assess artificial situation awareness in experiment 2 because the AI SA system’s technology was not mature in terms of project-level of implementation. Another reason why it is

EUROPEAN PARTNERSHIP
Co-funded by the European Union

difficult to critically assess the AI SA system is that it was already further developed until now (April 2022) than it was at the time of execution of experiment 2 (January 2022). Nevertheless, a summary of the analysis is made regarding the status of AI SA system at the time of the experiment. First, it is important to distinguish what the AI SA system is used for because this influences the analysis. Analyses discussed later assumes that the AI SA system is a stand-alone product and can act in parallel with the ATCO and is not intended to only support the ATCO. Therefore, it is essential to suppose that all observations made by the machine such as detecting conflicts (crossings, level, and speed bust), identifying aircraft on the wrong flight level, and giving aircraft a direct call, could be performed in reality.

To present the limitations of the system generating artificial situational awareness, there are two aspects, explained below, to be considered.

5.1.3.1 Conflict Detection ML Module

Conflict detection ML module provides the KG system with the results on the aircraft pairs which are considered as a situation of interest. Herein, the ML module provides information about pairs whose minimum distance would be less than 20 NM. This value is more than what an ATCO would consider a conflict. To prevent an overload of false predictions, conflict detection module output provides the user with additional information. Firstly, it calculates if the aircraft pair is a situation of interest and then calculates the probability of it. This information would give ATCO an insight into how urgent the addressed conflict is. The results obtained in the analysis could not rely on these calculations because there is not a straightforward connection by which ML module output could be filtered out.

Outputs that could be delivered from the conflict detection module are the distance and time to the predicted minimum distance. These outputs are not constant and proportional to time, i.e., the timestep between two predictions does not correspond to a reduction in time to the minimum distance. The reason for this is the way the conflict detection module determines where the conflict is in space and time. Using the training data rather than aircraft speed results in an output that is not usable for continuous tracking over time. Therefore, all the analyses on the conflict detection ML module were performed on one prediction, not continuously over time. To have a real-time operation, the ML module should be integrated into KG system and predictions in time should be averaged so there is a continuous flow of the time and distance to minimum distance predictions.

CD ML module training data explained in Section 1.2.1.3 is limiting the conflict detection module performance as it only accounts for aircraft types present in training data. Even though based on correlation, training statistics and CD ML module prediction are not statistically related, predicting values without having training statistics for an aircraft type would impair ML module performance.

The “black-box” effect takes away the opportunity to analyse why the conflict detection module delivers inaccurate predictions, why it predicts negative values or, most importantly, why some conflicts are not recognised at all. This reduces the level of confidence for the CD ML module.

Furthermore, the conflict detection module filtered the aircraft pairs to check only traffic that is not expected to be vertically separated. It cannot predict a conflict that would occur if one aircraft changes its flight level before its altitude enters another aircraft's flight level band. Thus, the conflict detection module recognises all traffic to be a situation of interest if one aircraft changes flight level in the vicinity of another one for the whole time until they are vertically separated. It does not consider cleared flight

EUROPEAN PARTNERSHIP
Co-funded by the European Union

level as a reference when checking if traffic is a situation of interest. If it could consider cleared flight level and distance to minimum distance, more relevant aircraft pairs would be predicted.

As well as predicting if two aircraft would be in a conflict if they change their flight level, predicting if two aircraft would be in conflict if cleared routes are issued is also a limitation of the conflict detection module.

5.1.3.2 Monitoring Tasks

The monitoring tasks listed in Table 2 in Section 1.2.1.4 provide an accurate and complete view of the current traffic situation, as seen from results presented in Section 4.1.1. However, some limitations have been identified and need to be considered.

At this stage of development, the KG system is not able to work in real-time. The flight data and the static data about the airspace needs to be stored in RDF graphs before the tasks can be applied to the data to get the outputs. Processes which automatically transform data to RDF form are not yet integrated with the rest of the system (which contains the tasks), so human action is still required.

The way the system currently operates will also prove to be a limitation in the future, considering the storage space available. The KG system saves all input data and all task outputs to the KG. At this point, the negative effect the number of graphs stored in the KG has on the computation speed has already been noted. This is something that should be considered and improved while developing a system working in real-time.

5.1.3.3 Awareness Level Classification

As already stated at multiple points in this document, determining the situation awareness of an AI system is a complex matter. Even if not viewed in the context of ATC and ATCO-machine collaboration, the classification of an AI system requires the application of a general framework to a specific system configuration. In the case of the AI SA KG system, which by design does not run in real time and is not completely integrated, this means that additional steps are required to explain why a system fulfils requirements of various awareness levels. Thus, the nature of the AI SA KG system presents a limitation during awareness level classification.

An example of incompatibility between the AI SA KG system and the framework chosen for awareness level classification is the requirement that a system measure the values of environment parameters, which can then be assigned to environment properties. While it is obvious that AI SA contains and uses values originating in its environment, the literal interpretation of that condition requires the existence of system-integrated sensors which would record those values. Their non-existence is of course valid for a PoC system, but the task remains to explain how the system accomplishes the task set forth by that condition. The same issue is apparent in other conditions and requirements set forth by the chosen framework. A different choice in framework or the creation of a new framework might be necessary if, during further development of AI SA or a similar ATC system, it becomes apparent that this framework continues to be incompatible with the type of system (and system architecture) being researched.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

5.1.4 Limitations Related to Participants

It was obvious that for some participants it was a somehow frustrating experience to have to work with a new tool that is less useful than the system they are familiar with (compare Figure 5). They were impaired in their situation awareness and performance and felt annoyed. This compromised the validity of situation awareness measurement because ATCOs' scanning and recognition of situations from experience was obstructed. It required more effort to develop situation awareness. Not only situation awareness was more difficult with the unfamiliar tool, but also the handling of the traffic required more effort and absorbed cognitive resources. Performing less than they could was a frustrating experienced. Frustration can evoke thoughts of anger, that absorb mental capacity needed for tasks – a vicious circle.

Judgement of the stress level of the subjects after the experiment was heightened. During the experiment they were regularly prompted with queries about situation awareness for which they needed to scan the radar and use an unfamiliar set of tools. ATCOs with difficulties to adapt to the new simulation tool probably felt annoyed. Additionally, the scenarios were rather short in length (between 5 and 23 minutes). ATCOs had to quickly reorient themselves in a series of five different scenarios. There also, adaptability mattered. On the other hand, the majority of ATCOs did not show elevated activation in the psychophysiological measurements. During simulation their activation was on average lower than before and after (compare Figure 16 and Figure 17). This indicates that accomplishing the tasks and concentrating actually lowered ATCOs activation a bit, which could be due to task familiarity and he fact that ATCOs are rigorously selected for their job.

Lower activation could also result from feeling less responsible when performing in a simulation as compared to real work because a mishap would have no consequences for safety nor traffic flow and capacity. ATCOs seemed to operate with a more relaxed attitude than during their work. It would therefore be interesting to compare with arousal measured during work in the ACC. On the other hand, their motivation might have outweighed relaxedness in simulation, when participating in an experimental study.

The operational and technical standards at Skyguide belong to the frontrunners in European. According to subject matter experts, Skyguide's system works very reliably. This may have lowered ATCOs attention towards non-conformances (complacency). ATCOs may have been biased because they expect other products to be equivalent. In fact, some ATCOs were very critical towards the experiment. The design of the experiment partly reinforced this.

Implementation of innovative tools in an existing environment encounters resistance because highly skilled participants lack habits to operate the system. This may evoke aversion against the new tools. Subjective perception of the usefulness AI SA system may have been biased by the fact that ATCOs were not familiar with the simulation tool and needed more effort to fulfil their job. This was not favourable for the evaluation of the AI SA proof-of-concept system conducted in work package 5.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

5.1.5 Behavioural Coding

The approach of creating data frames based on behavioural coding was chosen because it leaves little room for subjective interpretation as compared to judging the level of skilfulness. Only aspects connected to ATCOs' actions were considered. This allowed to analyse precisely what ATCOs were doing.

The coding of ATCO actions was done by three persons. They underwent standardisation training with a subject matter expert to minimise differences in coding style. Small differences in the start and end times of the events may have occurred. Setting of start time for events and actions may be imprecise and therefore not in perfect synchrony with the event in simulation for reasons of delayed reaction time.

It was sometimes difficult to define how long a conflict lasted. The end time of a conflict was set as the time when the ATCOs measured conflicting aircraft with VERA or when they directly reacted to it. In most cases, it was clear, because many ATCOs used the VERA tool as soon as they detected a conflict. However, it could be additionally checked when the corresponding aircraft were scanned.

5.1.6 Biometrical Analysis

It was decided to be refrained from good practice for biometrical measurement to perform a standardised stress calibration test. This had several reasons: Time constraints and - under circumstances of performing two experiments simultaneously in the same room - unnecessary distraction of participants. Calibration tasks can create an atmosphere of being tested.

There was therefore no standardised baseline in terms of resting heart rate nor a reference for being mentally charged to individually interpret the data. To compare activation levels during simulation or – more specifically - in the context of events with an individual baseline is necessary because the characteristics of psychophysiological reactions are idiosyncratic. Hence between-subject comparisons may not be made based on raw values. The choice of the off-task baseline as an alternative certainly needs to be reviewed, as most participants showed lower levels of activation during task accomplishment in simulation compared to baseline. The renouncement of the stress calibration resulted in a need to define an alternative baseline. The procedure to select data before and after scenarios and during breaks between scenarios to constitute an *off-task baseline* may have resulted in a mixture of arousing (e.g., nervosity prior to experiment) and also relaxing time periods (e.g., relief after simulation). A standardised approach with qual time span for a baseline for all ATCOs would be preferable.

To time span used for event-based analysis of psychophysiological parameters might have been too long (up to 10 minutes) to discover elevations in activation. Shorter phases (e.g., 10 seconds) need to be considered in analysis.

To minimise artefacts by means of median instead of procedures to exclude extreme values simplified the analysis but has possible side effects. The median value is the middle value in an ordered set of values arranged by size. Outliers in terms of artefacts (high signals due to motion or similar) and temporary zero-values are omitted and do not falsify analysis as it could happen if mean values were calculated (sum of all values sampled divided by the number of samples). The benefit of controlling artefacts may be diminished by a reduction of validity (not all relevant data included) as interesting

EUROPEAN PARTNERSHIP
Co-funded by the European Union

effects – the peaks of workload –also partly get omitted by this procedure. Removing artefacts could have generated more valid data on workload.

Besides statistical issues in the analysis of biometrical parameters it is important to consider the intrusiveness of attaching biometrical sensors and measurement devices to the participant's body. It can be physically disturbing (cables, patches, and pressure marks, etc.) and may trigger discomfort due to feeling observed and having no control over the measured reactions.

5.1.7 Eye Tracking Analysis

Different software was used to measure and analyse eye tracking data. However, these may have limitations that accumulate the imprecision or result in missing data. They are discussed in the following chapters.

5.1.7.1 Measuring of Eye Movements During Experiment

Gaze analysis with eye-tracking is challenging when distance between eye tracking glasses and radar screen is relatively large. Calibration is important for accuracy and may be disturbed if the glasses get moved, and processing of the data to map the gaze on areas of interest is often effortful.

The Tobii Glasses 3 have an accuracy of 0.5°. Figure 35 shows the difference in accuracy depending on distance. The two ellipses represent the screens. If the distance from the screen to the eye tracker is about 0.7m, then it leads to an expected deviation of around 6 mm (see Equation 4). If the distance is increased to 1.2m deviation is expected to be around 10 mm (see Equation 5).

Equation 4: Deviation of ET for a screen distance of 0.7m

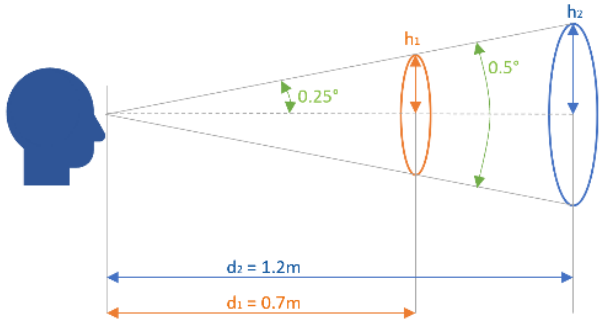
$$h_1 = d_1 \cdot \tan(0.25°) \approx 3mm$$

Equation 5: Deviation of ET for a screen distance of 1.2m

$$h_2 = d_2 \cdot \tan(0.25°) \approx 5mm$$

The label size of an aircraft on the radar screen is relatively small. Relative to the distance between ATCO and screen the expected deviation is considerable and may not allow to exactly determine which area of the label an ATCO was looking at. In addition, it is difficult to distinguish between labels of several aircraft grouped nearby. The use of 43'' screen made it necessary to place ATCOs at a distance about 70 cm so they could overview the radar screen at one glance and get all the information they needed.

EUROPEAN PARTNERSHIP
Co-funded by the European Union



$$\tan(0.25°) = \frac{h_2}{d_2} = \frac{h_1}{d_1}$$

Figure 35: Accuracy of Tobii Glasses 3

Another problem that can occur is inaccurate calibration. Basically, calibration was performed before each scenario, and it was tested if the glances were detected at the expected spots. Inaccuracies can occur if ATCOs moved the glasses in between or change their position. This has been noticed sometimes, when gaze was focused on areas where no information was displayed for ATCOs. This technical problem can be solved by post-correction.

5.1.7.2 Accuracy of the CVT

To analyse exactly when the ATCOs looked at which aircraft and calculate the duration a computer vision tool (CVT) was developed. This tool can recognise the screen in the ET video and the corresponding aircraft using computer vision. However, the accuracy of the screen recognition is greatly hampered by the background of the video and by head movements. This in turn affects the accuracy for recognition of the aircraft being viewed. Because the tool did not work for all ATCO recordings - e.g., because the background contained too many disturbing objects - only nine ATCOs could be analysed with this tool. This resulted in a significant reduction of the sample size for analysis.

The CVT estimates its accuracy with a confidence level. The confidence per scenario was on average 0.5, i.e., the tool is about 50 % sure that it has correctly made the assignments. As a result, 50 % of the data is lost or partially incorrect. It was noted that in some cases the confidence was low, although the screen is well recognised. The underlying short disturbances of the system may have a strong negative impact on confidence.

For the analysis, only ATCOs with an average confidence above 0.5 in the scenario were selected. The data was further cleaned by processing only data with a confidence higher than 0.5. By this the overall confidence increased to 0.7. This reduced uncertain results but also leads to missing data. It can happen that the CVT does not detect an aircraft even though it has been looked at. The tool is therefore too conservative and dismisses some data as uncertain that would actually be a true gaze on an aircraft label. Only choosing values with high confidence can ensure that detection of fixations on aircraft labels was actually true.

Also, the tolerance of the tool must be considered. It is difficult to exactly determine where the ATCOs have looked, therefore a tolerance for recognition of fixations needs to be selected. In this experiment, the tolerance corresponds to the size of the blue circle in Figure 13. If aircraft were far away from each other, this did not cause problems because only one AC could be identified anyway. However, if there

many aircraft were group nearby and the labels overlapped, then it is difficult to clearly assign the gaze. The tool then recognises that several aircraft were looked at once. It can be argued that if aircraft were very close together, then the ATCO perceives certain information from several aircraft through the area of peripheral vision. However, the area cannot be precisely defined.

5.2 Summary and Conclusion

The Single European Sky ATM Research (SESAR) 3 Joint Undertaking founded a research project called AISA to investigate the usefulness of artificial intelligence in ATM. This project developed an AI-based machine situation awareness system and tested its capability to accurately perform monitoring tasks and contribute to human-machine team situation awareness. A first experiment (N=20) made a differentiated analysis of the situation awareness of ATCOs and their working method using multiple methods. A second experiment (N=16) compared the situation awareness of ATCOs and the AI SA system with query probes and investigated the impact of artificial situation awareness inputs on ATCOs performance and judgement. In this simulation the AI SA system pointed out conflicts in advance or alerted about non-compliant aircraft. Including the element of AI in team situational awareness can widen geographical coverage and increase the time span covered by the awareness of the ATCO-AI team. Today however, this widening does not fully meet the ATCOs' needs as executive controller.

Topical section one compared different methods to measure **ATCO situation awareness**. Results showed that specific task-related measures with queries can be a helpful complement to subjective rating methods, as it provided ATCOs detailed feedback and thereby supported their self-monitoring. A comparison of the situation awareness of ATCOs showed that ATCOs with preserved situation awareness scanned and reacted to events neither particularly early nor late. Most of their transmissions integrated multiple instructions (e.g., heading and flight level change). Conflicts were resolved without necessity for later corrections to the initial conflict resolution. To support their situation awareness, they regularly used tools to measure the distances between conflicts, but not excessively. They scanned the radar more regularly, and their gaze fixed for a shorter duration and more specifically on areas of conflicts or high relevance. They correctly prioritized aircraft and saved mental workload. The capability not to rush in face of high task load and traffic complexity allowed those ATCOs to gain overview and solve tasks in a structured and effective way. Despite the use of multiple methods, it is difficult to recognize what caused them to recognise conflicts, while ATCOs with degraded situation awareness did not - even though they were scanning all relevant areas, too. The availability of accurate mental models is important for recognition - Endsley's situation awareness level 2 (see Section 2.1.1.2). In dynamic and complex environments such as ATC it might be especially challenging to develop and update accurate mental models that support situation awareness - because of irregularity, frequent changes and interdependencies requiring extensive mental resources. We therefore conclude that AI SA system inputs on monitoring tasks might act as a "second opinion" and support the development of differentiated and more accurate mental models. As the results showed, AI SA inputs seemed to enhance the self-monitoring capability of ATCOs (see Section 4.1.1.1).

Feedback on performance at work is sometimes complicated by the social context and in cases of hierarchically high-ranked individuals even prevented. This can be dangerous, for instance when people do not dare to speak up about safety relevant issues for social reasons or because they fear personal consequences. In a study by EHS today employees only speak up in 2 out of every 5 unsafe situations (Tucker & Turner, 2013). Transfer of critical information by machines is not reduced for social reasons. We therefore suggest that the future AI SA system facilitates a safety culture that promotes organisational learning and an open attitude towards errors, as the tool provides input on monitoring

EUROPEAN PARTNERSHIP
Co-funded by the European Union

tasks irresectable of the social context. Although automation might never become a “real teammate” to human operators, it will reliably contribute to safety and learning.

Topical section two made a **comparison of human and AI situation awareness** based on outputs from AI SA system stage I implementation and ATCO answers to identical queries about specific monitoring tasks. Concerning hits (correct positive) and correct rejections (correct negative) we found that some AI SA outputs did fully agree with ATCO answers, for other aspects it sometimes captured aspects that ATCOs had missed, but it also produced errors (false alarm and misses). The accuracy of the AI SA system performance was initially low and could be improved at stage II implementation including 46 of 57 monitoring tasks compared to 8 tasks at stage I implementation. In future, accuracy will be further increased, but the timing of artificial situation awareness inputs also plays and essential role, and - with it - the design of the HMI, too. If ATCOs with degraded situation awareness should be able to profit from AI SA inputs, a sophisticated design would be necessary that guides attention to specific relevant aspects -remember, those ATCOs scanned the relevant area but did not recognize.

Topical section three on **human-machine team situation awareness** explored the ATCO’s judgement about the usefulness of AI SA inputs. After initial enthusiasm about simulating a “fake” real-time interaction of ATCO and AI SA system we soon stumbled on requirements we could not meet with our rather rudimentary HMI design. The HMI design was not part of the project and therefore consisted only of oral inputs provided to ATCOs in a simulated work environment that was already rich of auditory information. This has led to distractions, annoyance, and additional mental load. Despite this, ATCOs receiving AI SA inputs - on average - discovered some conflicts earlier and noticed non-compliance by pilots more often upon warnings from AI SA system than ATCOs without AI SA inputs. From this we conclude that human-machine team situation awareness can in fact improve safety and performance in ATC provided - provided it is designed properly, and machine situation awareness inputs are reliably accurate.

Our opposite approach to deliberately disregard principles of good HMI design taught us some lessons about what happens if artificial situation awareness inputs usurp ATCOs’ attention at the wrong moment and load their minds when they are overly busy to take notice about information that becomes relevant for them only later. For instance, some of the artificial situation awareness inputs on conflicts were provided too early and others too late. From these results we conclude that inappropriate HMI design is counterproductive: It may distract ATCOs’ attention, load their memory with information that has little or no use for later recall and - in practice - may even be dangerous (see D5.2). We suggest that the synchronization of machine situation awareness with ATCOs’ minds and needs - as a result from their rhythm and style of work - is a major requirement for HMI design for human-machine team situation awareness.

To optimally deploy ATCOs’ attention for important information it is necessary to further develop the AI SA system to include mechanisms for prioritisation of information inputs according to ATCOs’ needs. This means we need to expand and include AI SA system’s awareness level about ATCOs for adaptive human-machine interaction. There are “windows of opportunity”, when ATCOs attention is available for information relevant to the current tasks – be it solving an impending conflict, optimising services or planning. Without the ability to filter information more specifically, it is necessary to design artificial situation awareness inputs in a way, that does not impose an immediate need for attentional resources, offers information in sensory modalities, that are least occupied by the tasks at hand, but alerts about urgent aspects. HMI solutions might for instance use different modes of information presentation for short- and long-term aspects. Today ATCOs mainly interact with pilots via radio transmission, putting high load on auditory attention. This might also be necessary in cases, where

EUROPEAN PARTNERSHIP
Co-funded by the European Union

ATCOs' attention is overloaded and risk for situation awareness errors is high. In near future, with more aircraft equipped for Controller Pilot Data Link Communications (CPDLC) providing oral artificial situation awareness inputs might be more beneficial than today. With increasing complexity and higher density of traffic it might be desirable for a future AI SA system to offer simulated forecasts of consequences for options in a visualized form and upon ATCOs' request. That way ATCOs could evaluate the favourability of options whenever they had time or were curious to see and learn.

Overall, ATCOs were not convinced by the performance of AI SA system at stage I implementation during experiment 2, but they were willing to trust an AI-based machine situation awareness in future. We conclude and recommend that the next step of automation could - and therefore should - enable ATCOs to develop and enhance situation awareness and in the long-term their expertise. It should not degrade them to supervisory controllers of ever more clever machines. An AI-based system combined with a reasoning engine might in future be able to learn and understand how ATCOs approach traffic situations and what information they need to be aware of to do so effectively and safely. That would then allow for a true human-machine collaboration based on human-machine team situation awareness. Importantly, this approach could prevent possible adverse effects of automation that also apply for an AI-based SA system: i.e., deskilling by lack of regular practice on the job, complacency due to a felt superfluousness of human control of highly reliable technical systems, fatigue and lowered readiness to perform due to inactivity, and confusion due to low involvement ("human-almost-out-of-the-loop") and high complexity of technology.

Topical section four dealt with the **accuracy of the estimations and predictions of the AI SA system as well as the level of awareness achieved it**. Monitoring KG tasks have proved a high level of accuracy. Analyses have shown that AI SA system noticed every deviation, violation, or non-compliance promptly and without any vagueness. Correctness of the system is not affected by workload, traffic count, or scenario design. The system allows ATCOs to become aware of their omissions in perception - Endsley's situation awareness level one – and select more accurate work strategies, and thereby differentiating ATCOs routine repertoires.

We close our reflections with a resumé: Increases in air traffic will require additional tools to keep the ATCOs' situation awareness and workload at a manageable level. The AI SA system proof-of-concept has demonstrated to be a promising way towards that goal. Machine situation awareness is more meticulous than a human operator with limited attentional resources can ever be, and it never turns off nor switches to other things. As expected, machine situation awareness outperformed ATCOs in some aspects of medium-term anticipation and detection of non-conformances in the second experiment. Nevertheless, designing artificial situation awareness inputs to contribute to human-machine team situation awareness is challenging by itself. Accuracy of machine situation awareness inputs is not enough; human prioritisation of information needs to be considered according to its logic for task fulfillment and scarce mental resources to accomplish this. Human beings are forced to compensate the lack of processing capacity by means of a repertoire of fine-tuned procedures for goal attainment that is stored in long-term memory and activated when information triggers a suitable procedure. What information is focused on is essential for selecting an adequate procedure to appropriately addresses task requirements and the context. AI SA system can help ensure ATCOs become aware of the relevant information to select appropriate procedures, thereby broadening, and refining their repertoire of procedures to become adaptive experts.

As Jenny Burkhalter – experienced ATCO and writer of the preface - would probably say: "That is, what ATC is about."

EUROPEAN PARTNERSHIP
Co-funded by the European Union

5.3 Outlook

More effort is needed for adaptable, human-centred design of automation. In future, the AI SA system should be able to recognise when information is relevant for ATCOs and support their awareness about relevant information for adequate decision making. The HMI needs to provide visual inputs on artificial situation awareness, thereby allowing ATCOs to choose when to process them. That way, inputs regarding monitoring tasks will not require controllers to immediately shift their focus of attention, and information remains available whenever needed. The audio inputs could be used to alert about critical situations such as a loss of safe separation.

An HMI design supporting ATCOs awareness for relevant information ensures ATCOs stay “in the loop” and can make appropriate decisions. This minimizes the dependence on the AI SA system in case of loss of functionalities or the whole system. How the ATCO can take over all monitoring tasks if AI SA fails will be solved by graceful degradation of the AI SA system, thanks to its self-awareness about subsystem failures. The solution is to build robust (redundancies), self-aware systems, to continuously expand the KG and to add new queries.

Today, the AI situation awareness system is not able to perform real-time. Capability for real-time processing can be reached by advanced optimisation, architectural changes, and improved hardware. In our opinion the concept of the AI SA system has proven to fulfil the requirements that could be implemented and evaluated within the scope of the project. In addition, its architecture revealed auspicious qualities (i.e., higher levels of awareness) for the realisation of more advanced human-centred design to reach human-machine team situation awareness. Therefore, we recommend pursuing the elaboration of these levels before heading towards a higher level of technological readiness (TRL 2 and more).

Another application of the artificial situation awareness system might be in training ATCO students in their scanning for monitoring tasks and to implement evidence-based training for licensed ATCOs. International Civil Aviation Organization (ICAO) promotes this shift in training paradigm to individually enhance competencies. The AI SA system could create a *digital twin* of each ATCO about their actual performance in comparison to best practice for monitoring task performance that AI SA system has learnt across ATCOs.

From the point of view of the socio-technical system, the AI SA system has the potential to amalgamate a team of ATCOs - representing unique fast decision-makers and experts - with a machine that can reinforce these strengths by providing information relevant for their adaptation. Safety – especially in dynamic fields – isn’t only a matter of reliability and accuracy of system components. To pay attention to relevant details and keep up with changes in a complex and dynamic environment is challenging (as you might have read in the preface of Jenny). This skill is improved with self-monitoring, that can be fostered by the feedback from the AI SA system. If the future AI SA system allows ATCOs to keep an active role, this will nourish their professional pride as a team capable to manage an incredible job. Pride is an emotion that promotes thoroughness, effort, and learning – and ultimately satisfaction with life. These are all factors protecting ATCOs against overload, stress, and burnout. A truly human-centred system like this would go beyond ergonomic design of HMI and circumvent efficiency-driven automation solutions that turn out to be a trap for human effectiveness.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

6 References

Ahlstrom, U., & Friedmann-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, *36*, 623–636.

AISA Consortium. (2019). *Grant Agreement—AISA*.

AISA Consortium. (2020a). *Concept of Operations for AI Situational Awareness (D2.1)*. SESAR Joint Undertaking. https://aisa-project.eu/downloads/AISA_D2.1_CONOPS.pdf

AISA Consortium (Ed.). (2020b). *Requirements for automation of monitoring tasks via AI SA (D2.2)*. 68.

AISA Consortium. (2021a). *4D Trajectory Prediction Module (D3.1)* (Ref. Ares (2021) 4892774; D3.1). SESAR Joint Undertaking. https://aisa-project.eu/downloads/AISA%20D3.1.pdf

AISA Consortium. (2021b). *Air traffic complexity estimation module (D3.3)* (REF. ARES (2021) 2272791; D3.3). SESAR Joint Undertaking. https://aisa-project.eu/downloads/AISA%20D3.3.pdf

AISA Consortium. (2021c). *Air traffic complexity estimation module (D3.3)* (REF. ARES (2021) 2272791; D3.3, Issue REF. ARES (2021) 2272791). SESAR Joint Undertaking. https://aisa-project.eu/downloads/AISA%20D3.3.pdf

AISA Consortium. (2021d). *Conflict Detection Module (D3.2)* (Ref. Ares (2021) 2337841). SESAR Joint Undertaking. https://aisa-project.eu/downloads/AISA%20D3.2.pdf

AISA Consortium. (2021e). *Facts, Rules and Queries Capturing En-Route ATC Operations (D4.4)*.

AISA Consortium. (2021f). *Populated Knowledge Graph (D4.3)*. SESAR Joint Undertaking. https://aisa-project.eu/downloads/AISA%20D4.3.pdf

AISA Consortium. (2021g). *Proof-of-concept KG system (D4.1)*. https://aisa-project.eu/downloads/AISA_4.1.pdf

AISA Consortium. (2022). *Risk assessment of AISA*. 81.

Alertness. (n.d.). In *APA Dictionary of Psychology*. https://dictionary.apa.org/alertness

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Anderson, j. R. (1982). Acquisition of Cognitive Skill. *Psychological Review*, 369–406.

Attention. (n.d.). In *Encyclopedia Britannica*. https://www.britannica.com/science/attention/The-intensity-of-attention#ref383433

Backs, R. W., & Boucsein, W. (2000). *Engineering Psychophysiology. Issues and Applications*. Taylor & Francis.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson Correlation Coefficient. In I. Cohen, Y. Huang, J. Chen, & J. Benesty (Eds.), *Noise Reduction in Speech Processing* (pp. 1–4). Springer. https://doi.org/10.1007/978-3-642-00296-0_5

Brannick, M. T., & Prince, C. (1997). An overview of team performance measurement. In M. T. Brannick, E. Salas, & C. Prince, *Team performance assessment and measurement: Theory, methods, and applications* (pp. 3–19). awrence Erlbaum Associates.

Broadbent, D. E. (1971). *Decision and stress*. Academic Press.

Buxbaum, H.-J. (2020). *Mensch-Roboter-Kollaboration*. Springer Gabler.

Cain, B. (2007). *A Review of the Mental Workload Literature* (Defence Research and Development Canada).

Can AI Systems Match Human-Level Situational Awareness? | Bench T. (n.d.). Retrieved 5 May 2022, from https://www.mouser.com/blog/can-ai-systems-match-human-level-situational-awareness

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin*, *104*, 163–191.

Create a new ggplot—Ggplot. (n.d.). Retrieved 5 May 2022, from https://ggplot2.tidyverse.org/reference/ggplot.html

Csikszentmihalyi, M. (1987). *Das flow-Erlebnis: Jenseits von Angst und Langeweile: Im Tun aufgehen*. Klett-Cotta.

Csikszentmihalyi, M. (2000). *Beyond boredom and anxiety, Jossey-bass publisher.*

Darr, S., Ricks, W., & Lemos, K. A. (2008). *Safer Systems: A nextgen aviation safety strategic goal*. 27th Digital Avionics Systems Conference October 26-30, 2008.

Dehn, D. M. (2008). Assessing the Impact of Automation on the Air Traffic Controller: The SHAPE Questionnaires. *Air Traffic Control Quarterly*, *16*(2), 127–146. https://doi.org/10.2514/atcq.16.2.127

Dekker, S. W. (2015). The danger of losing situation awareness. *Cognit Technol Work*, *17*, 159–161.

Di Lollo, V. (2018). Attention is a sterile concept; iterative reentry is a fertile substitute. *Consciousness and Cognition*, *64*, 45–49. https://doi.org/doi.org/10.1016/j.concog.2018.02.005

Duan, D., Xie, M., Mo, Q., Han, Z., & Wan, Y. (2010). An improved Hough transform for line detection. *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, *2*, V2-354-V2-357. https://doi.org/10.1109/ICCASM.2010.5620827

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J. N. D., & Maning, C. A. (1999). Situation awareness as a predictor of performance for en route air traffic controllers. Air Traffic Control. *Air Traffic Control Q*, *6*, 1–20.

Durso, F. T., & Manning, C. A. (2008). Air Traffic Control. *Reviews of Human Factors and Ergonomics*, *4*(1), 195–244. https://doi.org/10.1518/155723408X342853

Durso, F. T., & Sethumadhavan, A. (2008). Situation awareness: Understanding dynamic environments. *Hum Factors*, *50*, 442–448.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *Proceedings of the Human Factors Society 32nd Annual Meeting*, 97–101.

Endsley, M. R. (1995a). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*(1), 65–84.

Endsley, M. R. (1995b). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors Jorunal 37(1)*, 32–64.

Endsley, M. R. (1995c). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*, 32–64.

Endsley, M. R. (2006). Situation awareness. In *Handbook of human factors and ergonomics* (pp. 528–542). John Wiley & Sons.

Endsley, M. R. (2018). Expertise and Situation Awareness. In *The Cambridge Handbook of Expertise and Expert Performance* (2nd edition, pp. 714–741). Cambridge University Press.

Endsley, M. R., & Robertson, M. M. (2000). Situation awareness in aircraft maintenance teams. *International Journal of Industrial Ergonomics*, *26*, 301–325.

Ernst, D., Becker, S., & Horstmann, G. (2020). Novelty competes with saliency for attention. *Vision Research*, *168*, 42–52. https://doi.org/doi.org/10.1016/j.visres.2020.01.004.

Feltovich, P. J., Ford, K. M., & Hoffman, R. R. (1997). *Expertise in Context*. AAAI Press.

Flach, J. M. (1995). Situation Awareness: Proceed with Caution. *Human Factors*, *37*(1), 149–157. https://doi.org/10.1518/001872095779049480

Flavell, J. H. (1979). Metacognition and Cognitive Monitoring. A New Area of Cognitive-Developmental Inquiry. *American Psychologist*, *34*(10), 906–911.

Friedrich, M., Biermann, M., Gontar, P., Biella, M., & Bengler, K. (2018). The influence of task load on situation awareness and control strategy in the ATC tower environment. *Cognition, Technology & Work*, *20*, 205–217.

Gaze. (2022, April 22). https://en.wikipedia.org/wiki/Gaze

Groover, M. P. (2019). *Fundamentals of Modern Manufacturing: Materials, Processes, and Systems* (7th ed.). John Wiley and Sons Ltd.

Haeusler, R., Hermann, E., Bienefeld, N., & Semmer, N. (2012). How cockpit crews successfully cope with high task demands. In A. De Voogt & T. C. D'Oliveira, *Mechanisms in the chain of safety. Research and operational experience in aviation psychology*. Ashgate.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload. Advances in Psychology* (Vol. 52, pp. 139–183). North Holland.

Häusler, R. (2006). *Cockpit Crews auf Erfolgskurs—Eine videobasierte Analyse adaptiver Aufgabenstrategien in beanspruchenden Situationen* [Unveröffentlichte Dissertation am Lehrstuhl für Arbeits- und Organisationspsychologie der Universität Bern]. Bern.

Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J. H., & Welsh, T. N. (2019). No one knows what attention is. *Attention, Perception, & Psychophysics*, *81*(7), 2288–2303.

International Organization for Standardization. (1991). *Ergonomic principles related to mental work-load vol ISO 10075:1991*.

James, W., Burkhardt, F., Bowers, F., & Skrupskelis, I. K. (1890). *The principles of psychology*. Macmillan.

Janis, I. L. (1982). Decision making under stress. In *Handbook of stress: Theoretical and clinical aspects* (pp. 69–87). Free Press.

Jantsch, A., & Tammemäe, K. (2014a). *A framework of awareness for artificial subjects*. 1–3. https://doi.org/10.1145/2656075.2661644

Jantsch, A., & Tammemäe, K. (2014b). A framework of awareness for artificial subjects. *Proceedings of the 2014 International Conference on Hardware/Software Codesign and System Synthesis*, 1–3. https://doi.org/10.1145/2656075.2661644

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Jeannot, E., Kelly, C., & Thompson, D. (2003). *The Development of Situation Awareness Measures in ATM Systems* (HRS/HSP-005-REP-01) [Data set]. Koninklijke Brill NV. https://doi.org/10.1163/1570-6664_iyb_SIM_org_39214

Jones, D. G., & Endsley, M. R. (1996). Sources of situation awareness errors in aviation. *Aviation, Space and Environmental Medicine*, *67*, 507–512.

Kahnemann, D. (1973). *Attention and Effort*. PRENTICE-HALL INC.

Keinan, G. (1987). Decision making under stress: Scanning of alternatives under controllable and uncontrollable threats. *Journal of Personality and Social Psychology*, *52*, 639–644.

Lindsay, G. W. (2020). Attention in Psychology, Neuroscience, and Machine Learning. *Frontiers in Computational Neuroscience*, *14*, 1–21.

Luethi, M., Meier, B., & Sandi, C. (2009). Stress effects on working memory, explicit memory, and implicit memory for neutral and emotional stimuli in healthy men. *Frontiers in Behavioral Neuroscience*, *2*, 1–9.

Matthews, G., Davies, D. R., Westerman, S. J., & Stammers, R. B. (2000). *Human Performance. Cognition, stress and individual differences*. Psychology Press.

Miller, G., A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review*, *63*, 81–97.

Munir, A., Aved, A., & Blasch, E. (2022). Situational Awareness: Techniques, Challenges, and Prospects. *AI*, *3*, 55–77.

Palma Fraga, R., & Kang, Z. (2021). Visual Search and Conflict Mitigation Strategies Used by Expert en Route Air Traffic Controllers. *Aerospace*, *8*(170). https://doi.org/10.3390/ aerospace807017

Pullum, L. (2021). *Verification and Validation of Systems in which AI is a Key Element*. Oak Ridge National Lab. https://www.ornl.gov/publication/verification-and-validation-systems-which-ai-key-element

Rong, W., Li, Z., Zhang, W., & Sun, L. (2014). An improved Canny edge detection algorithm. *2014 IEEE International Conference on Mechatronics and Automation*, 577–582. https://doi.org/10.1109/ICMA.2014.6885761

Rouse, W. B., & Morris, N. M. (1985). *On looking into the black box: Prospects and limits in the search for mental models* (DTIC #AD-A159080). Center for Man–Machine Systems Research, Georgia Institute of Technology.

Salmon, P. M., Stanton, N. A., Walker, G. H., Jenkins, D., Ladva, D., Rafferty, L., & Young, M. (2009). Measuring situation awareness in complex systems: Comparison of measures study. *International Journal of Industrial Ergonomics*, *39*, 490–500.

Schaninger, A., & Hofer, F. (2004). Evaluation of CBT for increasing threat detection performance in X-ray screening. In K. Morgan & M. J. Spector, *The Internet Society: Advances in Learning, Commerce and Security* (pp. 147–156). WIT Press.

Schell, A., & Dawson, M. E. (2001). Psychophysiology. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 12448–12452).

Shahid, A., Wilkinson, K., Marcu, S., & Shapiro, C. M. (2012). Karolinska Sleepiness Scale (KSS). In A. Shahid, K. Wilkinson, S. Marcu, & C. M. Shapiro (Eds.), *STOP, THAT and One Hundred Other Sleep Scales* (pp. 209–210). Springer. https://doi.org/10.1007/978-1-4419-9893-4_47

Shotcut—Home. (n.d.). Retrieved 5 May 2022, from https://shotcut.org/

Shu, Y., & Furuta, K. (2005). An inference method of team situation awareness based on mutual awareness. *Cognition Technology & Work*, *7*, 272–287.

Smith, R. (2007). An Overview of the Tesseract OCR Engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, *2*, 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Sohn, Y. W., & Doane, S. M. (2004). Memory processes of flight situation awareness: Interactive roles of working memory capacity, long-term working memory and expertise. *Human Factors*, *46*, 461–475.

Sorensen, L. J., & Stanton, N. A. (2015). Exploring compatible and incompatible transactions in teams. *Cognition, Technology and Work*, *17*(3), 367–380.

Sperandio, A. (1978). The Regulation of Working Methods as a Function of Workload among Air Traffic Controllers. *Ergonomics*, *21*, 367–390.

Stanton, N. A., Salmon, P. M., Walker, G. H., Salas, E., & Hancock, P. A. (2017). State-of-science: Situation awareness in individuals, teams and systems. *Ergonomics*, 1–18.

Taylor, R. M. (1990). *Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design.* (AGARDCP-478; Situational Awareness in Aerospace Operations, pp. 1–3). NATO-AGARD.

Tucker, S., & Turner, N. (2013). Waiting for safety: Responses by young Canadian workers to unsafe work. *Journal of Safety Research*, *45*, 103–110. https://doi.org/10.1016/j.jsr.2013.01.006

Vetter, C. T. (2022). *Acquisition of Situation Awareness and Performance Parameters from En-Route Air Traffic Controllers and Comparison with AI's Situation Awareness* [Unpublished Master Thesis].

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies, *Varieties of attention*. Academic Press.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theor. Issues in Ergon. Sci.*, *3*(2), 159–177.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, *58*, 1–17.

Zangmeister, W. H., & Stark, L. (1982). Types of gaze movement: Variable interactions of eye and head movements. *Experimental Neurology*, *77*(3), 563–577.

Zsambok, C. E., & Klein, G. (1997). *Naturalistic Decision Making*. Psychology Press.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

7 Glossary

This section provides an overview of the terminology used across different deliverables of the AISA project and definitions.

Term	Definition
AI Situation Awareness System (AI situational awareness system)	The operating system that will be implemented by ATM system providers in the future. It means the future ATC system together with an AISA AI engine. In some cases, the system is referred to as “AI based support system”
AI Situation Awareness Model (AI SAM)	The model developed within AISA which represents core functions of the future system (AI SAS) relevant for the project.
AI SA KG system	System developed during the AISA project, composed of a KG, AI SA tasks and ML modules. Also referred to as “the AI SA system”, “the system”, and “PoC system”.
Artificial Intelligence	The ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalise, or learn from experience. (Source: Britannica)
Artificial situation awareness	This term is used interchangeably with the term “machine situation awareness”. Both terms refer to situation awareness that is generated by the AI situation awareness system.
ATM environment	The overall set of systems, processes, functions, and infrastructure where air traffic control takes place. The current environment describes the status during the preparation of this document (i.e., 2020, 2022), whereas the future environment predicts the situation in 2035-2040.
ATC system	The set of systems the ATCO is using, including the ones which are working in the background and are directly linked to those visible to the ATCO.
Automation	The creation of technology that will execute a certain task or set of tasks automatically.
Conflict	Loss of safe separation (5 NM horizontally, 1000 ft vertically).
Exercise	Product of an ATCO’s interaction with a scenario. Each ATCO input influences the traffic flow, making every ATCO’s exercise unique and different from the original scenario.
Final prediction	The conflict detection ML module output after all ATCO clearances were issued for the aircraft pair and no further trajectory changes were made.
Initial prediction	The conflict detection ML module output before any ATCO clearances were issued for the aircraft pair and no trajectory changes were yet made.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Machine situation awareness	This term is used interchangeably with “artificial situational awareness”. Both terms refer to situation awareness that is generated by the AI situation awareness system.
Monitoring	This term is used in two different manners in this document. First of all, as the work plan indicates, AI SA plans to start primarily with those “monitoring tasks” which currently (2020) require only monitoring type of contribution by the ATCO either due to the relatively significant level of automation or because the task itself is simple and requires no more interaction than monitoring. On the other hand, in terms of classification of future tasks among human and machine, “monitoring” means if in the future (medium or long-term scenario) a task is so highly automatised (with AI involvement), ATCOs will only need to perform monitoring activities.
Scenario	The design of the airspace and air traffic in the ESCAPE Light simulator. Each scenario is defined by a specific traffic mix.
Situation Awareness	SA is the perception of environmental elements and events with respect to time or space, the comprehension of their meaning, and the projection of their future status. (Source: Mica R. Endsley, “Toward a theory of situation awareness in dynamic systems”)
Situation of interest	Traffic situation relevant for the analysis of situational awareness defined by the minimum distance between two aircraft. Note: the limit for the situation of interest may be different regarding the needs of the analysis (i.e., the conflict detection ML module uses 25 NM as the limit, whereas in the SA comparison analysis, 10, 7.5 and 12.5 NM were used).
Shared situation awareness	Shared perception of environmental elements and events with respect to time or space, the comprehension of their meaning, and the projection of their future status within a team.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Appendix A Extracts of the SHAPE Questionnaire

A.1 SASHA_Q

After the simulation

In the previous working period ...

	0 (never)	1	2	3	4	5	6 (always)
... I was ahead of the traffic.	○	○	○	○	○	○	○
... I started to focus on a single problem or a specific area of the sector.	○	○	○	○	○	○	○
... there was a risk of forgetting something important (like transferring an a/c on time or communicating a change to an adjacent sector).	○	○	○	○	○	○	○
... I was able to plan and organise my work as I wanted.	○	○	○	○	○	○	○
... I was surprised by an event I did not expect (like an a/c call).	○	○	○	○	○	○	○
... I had to search for an item of information.*	○	○	○	○	○	○	○

*This statement is not concerning with the ESCAPE simulation environment but rather with air traffic in the simulation itself.

How satisfied are you with your performance?
1 = not satisfied at all
5 = completely satisfied

1 2 3 4 5

3

A.2 SASHA_L

1. E2S2.1:
    - 1.1. Which aircraft are in conflict?
    - 1.2. What additional conflicts do you see? (2x)
    - 1.3. Which aircraft have an exit conflict?
    - 1.4. Is there any non-conformance? (trick question)
2. E2S2.2:
    - 2.1. Which aircraft are in conflict?
    - 2.2. What do you need to pay attention to? (2x) (1x trick question)
    - 2.3. Is there any non-conformance? (2x) (1x trick question)
3. E2S3:
    - 3.1. Which aircraft are in conflict at exit point?
    - 3.2. Is there any conflict? (2x)
    - 3.3. Did you notice anything?
    - 3.4. Is there a non-conformance?
    - 3.5. What do you need to pay attention to? (trick question)
    - 3.6. Which aircraft can climb?
4. E2S4.1
    - 4.1. Which aircraft need to descent?
    - 4.2. Center is now available. Which aircraft can turn for a direct to?
    - 4.3. Is there a conflict?
    - 4.4. Is there any non-conformance? (trick question)

EUROPEAN PARTNERSHIP
Co-funded by the European Union

5. E2S4.2
    5.1. Which aircraft have an exit conflict?
    5.2. Mil East will be on in 2 minutes. Who is flying through the military?
    5.3. Which aircraft are in conflict?
    5.4. Is there any non-conformance? (trick question)

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Appendix B Description of Computer Vision Tool (CVT)

The next paragraphs describe on a technical low-level how CVT works. The whole process is divided into five steps

- Screen detection
- Coordinate system mapping from ”global space” to ”screen space”
- Importing airplane positions, compensating user screen movement, and zooming, map geographic coordinates to screen coordinates
- Label detection
- Detecting when an airplane was looked at

The first step is technically the most complex. Since the ATCO’s gaze is stored in the coordinate system of the ET glasses, it cannot simply be transferred to the screen recording. Instead, the screen must first be identified in the ET video to make this possible. Complex computer vision algorithms were used for this purpose. The different steps are visualised in Figure 36 and are shortly described in the list below.

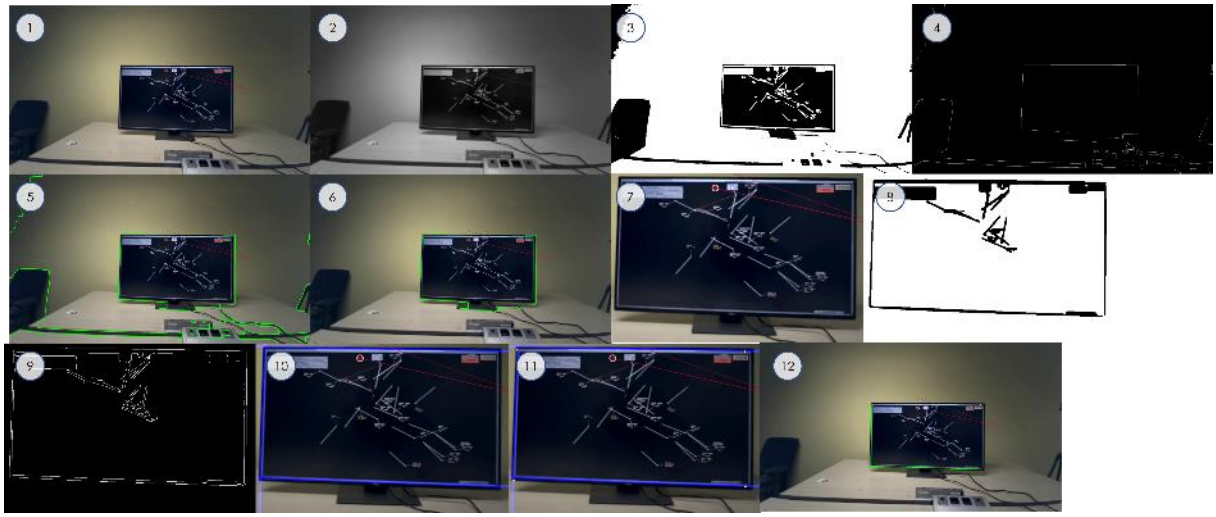


Figure 36: Steps for the screen detection of the Computer Vision Tool

1. The raw input image
2. Gray scaling and Gaussian blurring
3. Binary thresholding (to black and white)
4. Canny Edge Detection (more information is described in the article by W. Rong et al. (Rong et al., 2014)): to detect edges in the video
5. Contour detection: detect connected areas
6. Contour filtering: The screen is one of the largest contours and is roughly in the center. That can be used to identify the screen contour
7. Cutting the original and thresholded grayscale image to the bounding boxes of the detected contour
8. Remove the largest and smallest connected areas from the image (to reduce noise)
9. Canny Edge detection
10. Hough transform (non-probabilistic) to detect line-candidates (more information is described in the article by D. Duan et al. (Duan et al., 2010))
11. Calculate line intersections (= screen corners), classify them to screen corners, average the individual corner positions, and low-pass filter their positions (Assumption is that head

EUROPEAN PARTNERSHIP
Co-funded by the European Union

movement is smooth, so the corners should not move too much between frames. That makes the estimation a bit more robust towards line outliers)

12. Map the detected corners back into the global space. Now it is possible to get the screen space coordinates of a point in the Eye-Tracking-glass-space.
13. After the screen has been identified in the ET video, the gaze can be transformed from the glasses' coordinate system to the screen's coordinate system. The transformation makes it possible to display the gaze in the screen recording.

Next, the aircraft must be detected and marked. For this, the ESCAPE Light logs are used, which contain the positions of the aircraft. However, since these are given in geographic coordinates, these coordinates must also be used in the screen recording. For this, the position of Switzerland was used. Since Switzerland always keeps the same shape, except in the military scenario, it is possible to define two always visible points and to which the real longitudinal and latitude coordinates are assigned. Since Switzerland's contrast and surroundings are always the same and unique, it is not a problem to always recognise Switzerland and, therefore, the two points. The recognition of Switzerland is also vital to identifying the zoom factor. After defining these two points, the ESCAPE Light log data can tag the aircraft.

The next step is to identify the label. Since the ATCO often does not look at the airplane itself (white dot) but concentrates on the label, it is crucial to be able to identify the label to the corresponding aircraft. For this, computer vision algorithms were used again, and Tesseract OCR, an optical character recognition engine. The article by R. Smith (Smith, 2007) describes more information to Tesseract OCR. With the engine, it is possible to read text from an image, in this case, the callsign in the labels. Thus, it is possible to assign the corresponding label to each aircraft.

In the upper left corner in Figure 37, some information about the progress and confidence of the tool is given. In the lower part of Figure 37, the tool's confidence is shown again graphically. What confidence means is explained in the following paragraph. Figure 38 shows what information the tool has detected on the screen. The gaze of the ATCO is displayed with a big blue ring. This ring moves with the real view from the ET recording.

Figure 37: Information on ET-Recording in CVT

Furthermore, the detected aircraft are marked with a small green ring. The aircraft is then additionally marked again with the callsign. However, as shown in the picture, the green rings do not perfectly

EUROPEAN PARTNERSHIP
Co-funded by the European Union

match the position of the aircraft. This is because the synchronisation of the df with the actual gaze from ET-recording is not done properly. What this means exactly will be explained in a later paragraph. Also, the found labels are marked with a small blue ring. If the label was not found, the small blue ring remains on the green ring. Additionally, Switzerland is highlighted with a solid green line.

Figure 38: Information on Screen recording in CVT

Finally, all that remains is to store the created data into a new df. For this purpose, the tool is started, and it stores in a list when which aircraft is viewed. The label and the aircraft itself are taken into account for the detection. For this, every aircraft and label is taken into account, which lies within the gazes (big blue ring).

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Appendix C Graphs

In the following, graphs are shown which are interesting to examine the behaviour of the ATCO before and after the experiments. For this purpose, a graph for experiment 1 and 2 is shown each time.

C.1 Karolinska Sleepiness Scale

The scale measures the subjective level of sleepiness at the time when the question was asked. On this scale subjects indicate which level best reflects the psycho-physical (Shahid et al., 2012). The first five levels refer to an active state whereas the last 4 levels refer to a sleepy state.

Participants were asked "How do you feel at the moment?"

It can be seen in Figure 39 that the average refers to an active state, but some were more tired after the experiment 1. It can be seen in Figure 40 that the average refers to an active state, and the experiment 2 made the participants more tired, but on a low level.

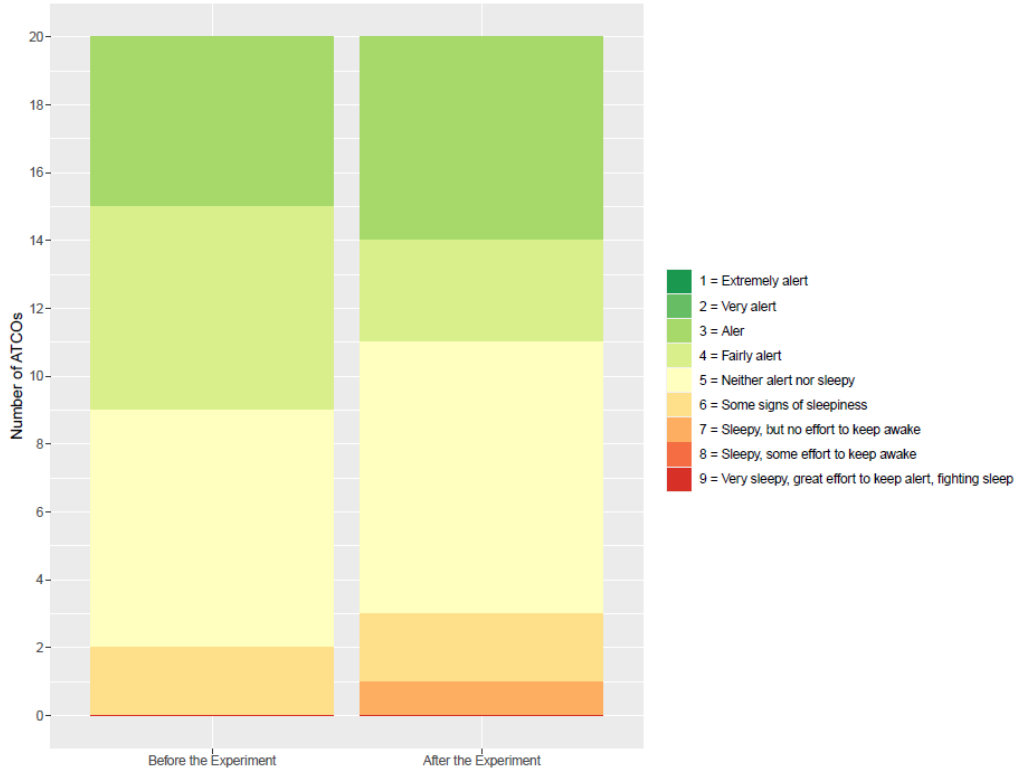


Figure 39: Sleepiness before and after experiment 1

EUROPEAN PARTNERSHIP
Co-funded by the European Union

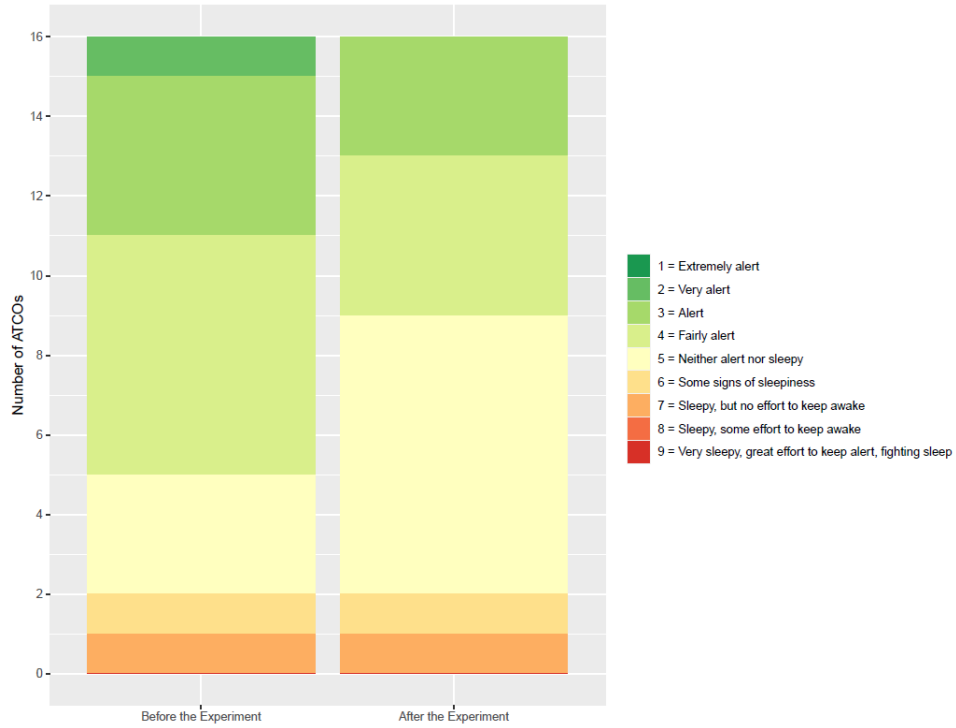


Figure 40: Sleepiness before and after experiment 2

C.2 Stress

The aim was to measure how stressed the participants felt during and after the experiment. Question "How stressed do you feel overall?" was asked for this purpose.

Figure 41 shows that the participants were less stressed after experiment 1, and the extremum is higher after it. experiment 2 is different from experiment 1 as seen in Figure 42. The participants are less stressed before experiment 2, but more extremes show up here.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

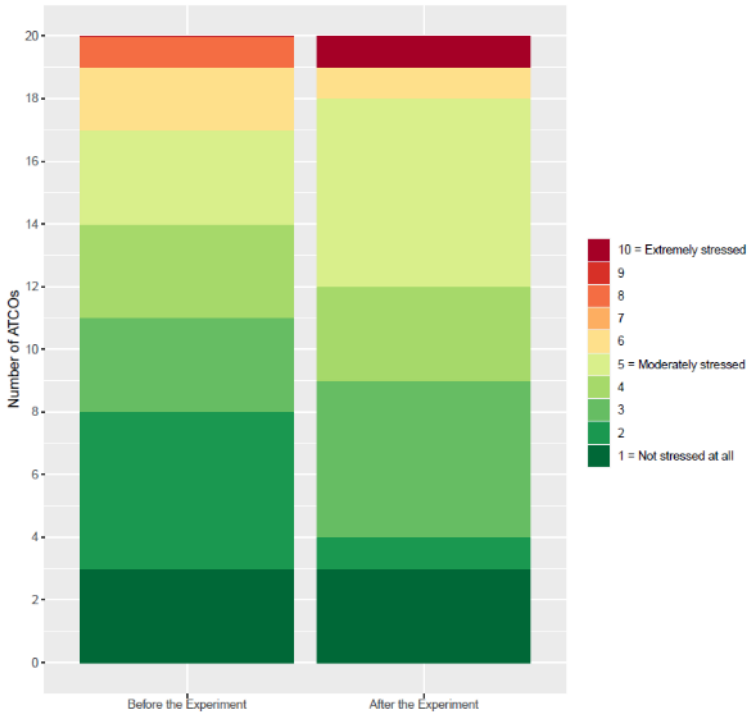


Figure 41: Stress level before and after experiment 1

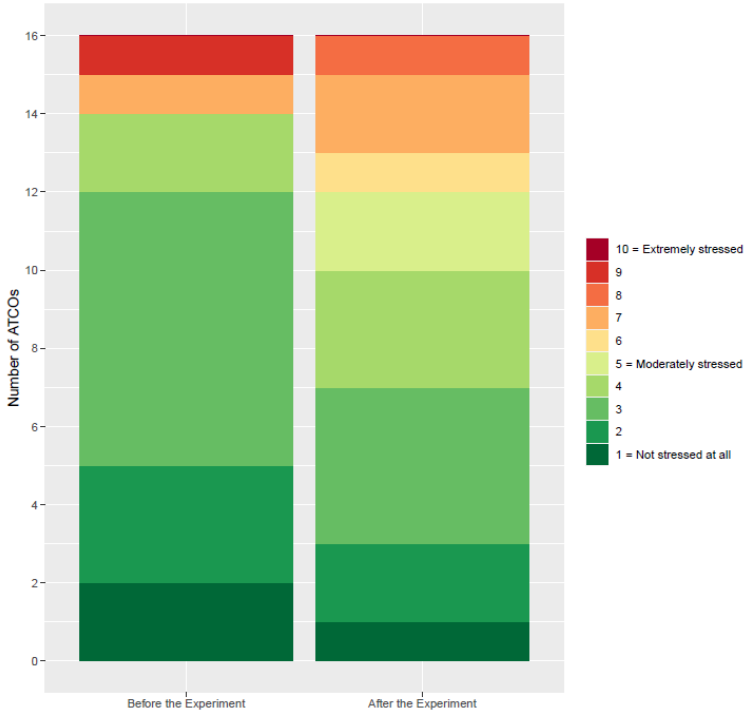


Figure 42: Stress level before and after experiment 2

EUROPEAN PARTNERSHIP
Co-funded by the European Union

C.3 Satisfaction

The ATCOs were asked after each scenario how satisfied they were with their own performance. Figure 43 shows the results of the question in experiment 1 and Figure 44 shows the results for experiment 2. Overall satisfaction was higher in experiment 1 than in experiment 2. This may be because the ATCOs themselves did not have control in experiment 2 but carried out commands. In experiment 1 (Figure 43) it can be seen that satisfaction was very high and that, in general, the scenarios in which the conflicts were better resolved performed better. In experiment 2, it is not clear how the results differed, as the scenarios lasted for different lengths of time, were answered differently in relation to the queries, and were not interactive.

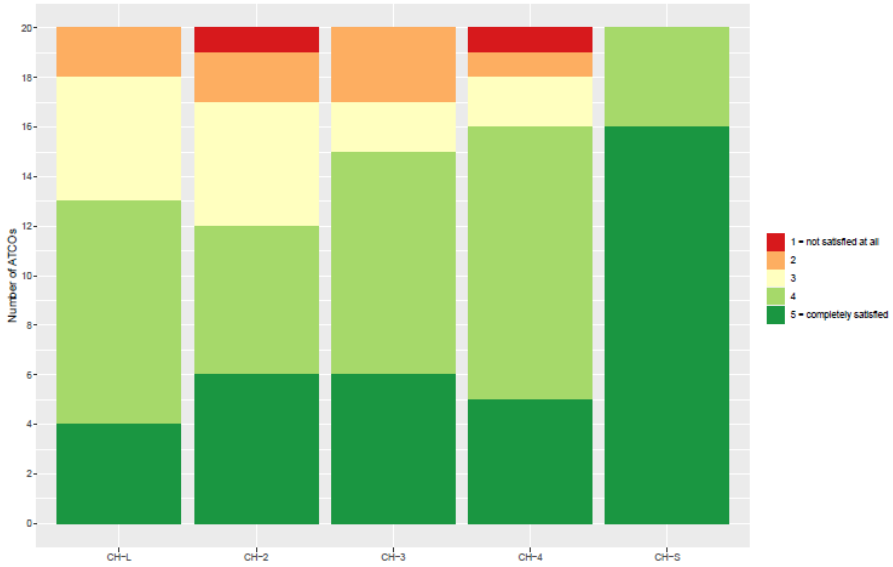


Figure 43: Satisfaction of ATCOs for each scenario of experiment 1

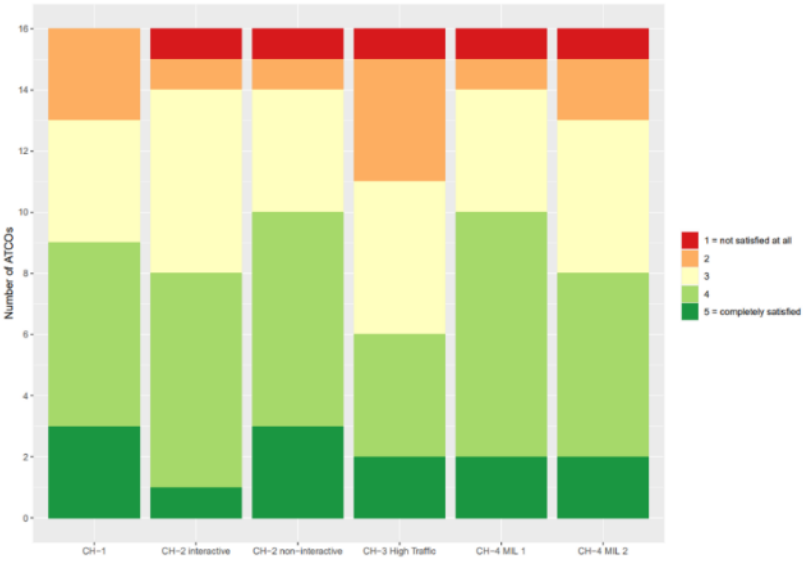


Figure 44: Satisfaction of ATCOs for each scenario of experiment 2

Co-funded by the European Union

Appendix D Data Frame for Count – Experiment 1

This data frame counts how often certain events (e.g, assume, transfer, use of CPDLC, initial calls, use of VERA, use speed vector change) occurred. Based on this data frame, it is possible to roughly estimate the mental workload of the ATCOs, since the assume and initial calls should occur equally often for all ATCOs, because the scenarios and the occurring aircraft were the same.

The following shows the data frame for the count of assume for four different ATCOs. The other data frames are developed in the same way.

Counter Assume

ATCO_ID	Light	Crossing	MIL	Short	High
1	12	18	22	6	15
2	11	18	22	5	15
3	12	17	22	5	16
4	11	17	22	5	15
...					
Mean	11.28	17.28	22.39	5.22	15.17

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Appendix E Data Frame for Conflict Comparison – Experiment 1

This data frame determines the duration of the conflicts and the solution. It can be used to quickly determine which ATCO has identified the conflicts quickly/slowly and which solutions do not meet the standard. It shows when the conflict started and when it was finished (i.e. when the ATCO detected the conflict). Furthermore, it shows which solution the ATCO chose to solve the conflict.

Only the first four ATCOs are shown to present the idea of the data frame, but to prevent direct attribution of the ATCO. The figure below shows the data frame for the E1S2 scenario. For the other scenarios it looks the same with the corresponding conflicts.

	Crossing: TVF4740 & IBK1CH				Crossing: FPO85J & BTI8EP			
ATCO	StartTime [min]	EndTime [min]	Duration [min]	Solution	StartTime [min]	EndTime [min]	Duration [min]	Solution
1	5.923	7.173	1.25	Direct	5.923	8.134	2.211	no solution needed
2	3.699	9.623	5.924	Direct	3.699	7.193	3.494	Direct
3	3.773	4.28	0.507	Direct	3.773	4.109	0.336	Direct
4	2.967	3.053	0.086	Direct	3.612	4.593	0.981	Direct
...								

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Appendix F Mean Reaction times data frame of experiment 1

This data frame compares how fast the ATCOs reacted to an initial call in one scenario. It means from when they have detected the aircraft when the pseudo-pilot calls the Swiss radar for the first time. ATCOs who see the aircraft quickly remembered it and therefore statements can be made about the SA. High negative numbers implicate that the ATCO assumed an aircraft before the pilot called.

Only the first four ATCOs are shown to present the idea of the data frame, but to prevent direct attribution of the ATCO.

ATCO_ID	Light	Crossing	MIL	Short	High	Mean all scenarios
1	2.474	3.26	11.554	1.723	3.315	4.465
2	0.127	-3.581	0.989	0.152	1.117	-0.239
3	1.571	3.12	2.739	1.573	3.008	2.402
4	2.339	1.802	7.225	-10.75	2.384	0.6
...						

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Appendix G Data Frame for Checkbox – Experiment 1

With this data frame, the ATCOs can be compared by checking the commands for each aircraft. Communications were checked for different code words like assume, transfer, speed, HDG, direct, climb, and descent. Whenever one of these terms was used for an aircraft, it was noted for the ATCO:

- If the ATCO called the aircraft and the codeword was mentioned, it is marked with a checkmark symbol,
- If the ATCO called the aircraft and the codeword was not mentioned, it is marked with a cross symbol,
- If the ATCO did not call the aircraft at all, a hyphen sign is displayed.

In the next step, however, only those events were analysed for which it can be said with certainty that they should have occurred.

Only the first four ATCOs are shown in the figure below to present the idea of the data frame, but to prevent direct attribution of the ATCO. Also, not all aircraft are shown, because it is the same principle for all of them. The figure shows the data frame for the E1S2 scenario, but the structure for the other scenarios is the same.

A/C an Konflikt beteiligt	Callsigns	Event	1	2	3	4	...	Count Check	Count Cross
x	AFL2548	assume	✓	✓	✓	✓		18	0
		transfer	×	×	×	×		1	17
		speed	×	×	✓	×		2	16
		direct	×	✓	✓	✓		16	2
		HDG	×	✓	✓	✓		13	5
	BAW577	assume	×	×	×	×		0	18
		transfer	✓	✓	✓	✓		17	1
		HDG	×	×	×	×		0	18
		direct	✓	✓	✓	✓		14	4
	BAW842P	assume	✓	✓	✓	✓		18	0
		transfer	×	×	×	×		0	18
		direct	✓	✓	✓	✓		10	8
		HDG	×	×	×	×		1	17
	BAW881V	assume	×	×	×	×		0	18
		transfer	✓	✓	✓	✓		18	0
		direct	×	✓	×	×		6	12
		HDG	×	×	×	×		0	18
x	BTI8EP	assume	✓	✓	✓	✓		18	0
		transfer	×	×	✓	×		10	8
		direct	✓	✓	✓	✓		15	3
		HDG	×	×	×	×		3	15
	CCA862	assume	✓	✓	✓	✓		18	0
		transfer	✓	✓	✓	✓		18	0
		direct	×	×	×	✓		1	17
		HDG	×	×	×	×		0	18
		climb	✓	✓	✓	✓		18	0
	DLH1158	assume	×	×	×	×		0	18
		transfer	✓	✓	✓	✓		18	0
	EJU26BX	assume	✓	✓	✓	✓		18	0
		transfer	×	×	×	×		0	18
...									

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Appendix H Data Frame to Compare Time – Experiment 1

This data frame is the extension of the checkbox data frame. Each checkmark was replaced by the time when the event took place. This way, it can be determined which ATCO will carry out events sooner or later. The time when the transfer call was made is especially interesting because it can be deduced how long the aircraft was on the frequency.

Only the first four ATCOs are shown in the figure below to present the idea of the data frame, but to prevent direct attribution of the ATCO. Also, not all aircraft are shown, because it is the same principle for all of them. The figure shows the data frame for the E1S2 scenario, but the structure for the other scenarios is the same.

conflict involved	Callsigns	Event	1	2	3	4	...	Mean	SD	Max	Min
x	AFL2548	assume	7.9	8.3	6.6	7.6		7.4	0.9	8.9	5.6
		transfer	x	x	x	x		15.0	-	15.0	15.0
		speed	x	x	12.1	x		13.4	1.8	14.6	12.1
		direct	x	12.1	6.6	10.4		11.1	2.6	13.8	6.6
		HDG	x	8.3	8.5	7.6		8.6	1.0	10.5	6.3
	BAW577	assume	x	x	x	x		-	-	0.0	0.0
		transfer	9.9	6.2	7.0	5.1		7.1	1.1	9.9	5.1
		direct	2.1	0.2	0.6	0.4		1.0	0.8	2.9	0.2
		climb	x	x	x	x		-	-	0.0	0.0
		descent	x	x	x	x		-	-	0.0	0.0
		speed	x	x	x	x		-	-	0.0	0.0
		HDG	x	x	x	x		-	-	0.0	0.0
	BAW842P	assume	2.4	3.1	2.2	1.5		2.3	0.6	3.3	1.3
		transfer	x	x	x	x		-	-	0.0	0.0
		direct	2.6	3.1	2.2	1.6		3.1	0.9	4.4	1.6
		climb	x	x	x	x		-	-	0.0	0.0
		descent	x	x	x	x		-	-	0.0	0.0
		speed	x	x	x	x		-	-	0.0	0.0
		HDG	x	x	x	x		14.3	-	14.2	14.2

...

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Appendix I Data Frame for Number of Events per Call & Number of Conflict Solutions – Experiment 1

From the following data frames, it can be counted how many commands are given to the pseudo-pilot in a single call. With an ATCO that has a good SA, it is expected that some commands will be given in one call, e.g., a direct command in the initial call.

The data frame also counts how many solutions have been applied until the conflicts are finally resolved. This allows for the ATCO's misperceptions of the situation to be identified and thus conclusions to be drawn about the SA.

A # symbol (“hashtag”) marks each time a new call has taken place. The & symbol (“ampersand”), on the other hand, marks when several events have been transmitted to the pilot in one call. If a # symbol is followed by a - (dash), then the ATCO has not issued any further commands.

Only the first two ATCOs are shown in the figure below to present the idea of the data frame, but to prevent direct attribution of the ATCO. The figure shows the data frame for the E1S2 scenario, but the structure for the other scenarios is the same.

conflict involved	Callsigns	ATCO 1	ATCO 2	...
x	AFL2548	assume # - # - # -	assume & change HDG # direct BENOT # - # -	
	BAW577	direct LUL # transfer # - # -	direct LUL # transfer # - # -	
	BAW842P	assume # direct RESIA # - # -	assume & direct RESIA # - # - # -	
	BAW881V	transfer # - # - # -	direct LUL # transfer # - # -	
x	BTI8EP	assume & direct BENOT # - # - # -	assume # direct BENOT # - # -	
	CCA862	assume & climb # transfer # - # -	assume & climb FL370 # transfer # - # -	
	DLH1158	transfer # - # - # -	transfer # - # - # -	
	EJU26BX	assume # direct LUL # climb # -	assume # climb FL380 # direct LUL # -	
	EJU67NL	climb & direct LOKTA # - # - # -	climb FL400 # transfer # - # -	
	EXS440	assume & direct LUL # - # - # -	assume & direct LUL # - # - # -	
	EXS82W	assume and direct LUL # - # - # -	assume # direct LUL # - # -	
x	FPO85J	assume # change HDG # assume # direct RESIA via CPDLC	assume # assume # - # -	
	HOP413F	direct BENOT # descent # transfer # -	direct BENOT # descent FL360 # transfer # -	
x	IBK1CH	assume & direct BENOT # change HDG # direct BENOT # transfer	assume # change HDG # direct GVA # transfer	
	IBK36FS	direct LOKTA # - # - # -	direct LOKTA # transfer # - # -	
	IFA229	assume and direct BENOT # - # - # -	assume & direct BENOT # - # - # -	
	JAF81J	assume & direct LUL # descent # force ACT # -	calling # assume & direct LUL # change HDG # -	
	MSOBM	assume and direct LUL # - # - # -	assume & direct LUL # - # - # -	
	RYR5QX	assume & direct BENOT # change HDG # direct BENOT # -	assume & direct BENOT # - # - # -	
	RYR7PN	assume # climb # - # -	assume # climb FL360 # - # -	
	RYR7RC	check # assume # direct LUL # -	assume & direct HOC # - # - # -	
	TOM51G	assume & direct LUL # climb # - # -	assume & climb FL380 & direct LUL # - # - # -	
x	TVF31WW	direct SUXAN # climb # transfer # -	direct SUXAN # climb FL390 # transfer # -	
x	TVF4740	assume & direct GAMSA # change HDG # direct MADEB # -	assume # wrong Call # direct MADEB # -	

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Appendix J Data Frame for Pearson Correlation

The data frame describes the scores for the five different categories (SASHA_Q, SASHA_L correct, ET conflict detection, implicit measurements) for the Pearson correlation. Only the first four ATCOs are shown in the figure (E1S2) below to present the idea of the data frame, but to prevent direct attribution of the ATCO. The structure is the same for the E2S2.1 data frame. The only difference is that that the SASHA_L score is included in the E2S2.1 scenario.

unit	score	%	min	score
ATCO	SASHA_Q	SASHA_L correct	ET conflict detection	implicit measurements
1	11	-	41.04	3.41
2	16	-	40.27	3.37
3	25	-	13.78	-6.04
4	36	-	24.35	-2.71
...				

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Appendix K Correlation of Conflict Detection Module Input Values Deviation with Prediction Error

A brief analysis was made to determine whether using the input values to ML conflict detection (CD) module that deviate from the training dataset correlate with the accuracy of the CD module. Each of the input variables for each of the conflicts was analysed to determine how much it deviates from the training data. Input variables were related to flight parameters of aircraft involved in particular conflict (altitude, track, speed, and rate of climb/descent). Training data was characterised by mean value and standard deviation, assuming normal distribution as explained in Section 1.2.1.3. Input variables were scaled according to number of standard deviations away from thus described training data inputs.

On the other hand, CD module conflict prediction error was calculated in terms of difference between the predicted minimum distance between aircraft and actual minimum distance. Correlation between deviation of input variables values and CD module's prediction error was performed in IBM SPSS. The results are presented in Table 27. No significant correlation was found.

Table 27: Correlation of Deviation of Input Variable Values with Conflict Detection Module Prediction Error

Variable	Pearson Correlation Coefficient	Sig. (2-tailed)
Altitude of 1st Aircraft	-0.026	0.864
Altitude of 2nd Aircraft	-0.107	0.482
Speed of 1st Aircraft	0.143	0.349
Speed of 2nd Aircraft	0.122	0.426
Track of 1st Aircraft	-0.246	0.103
Track of 2nd Aircraft	0.054	0.725
ROCD of 1st Aircraft	-0.279	0.063
ROCD of 2nd Aircraft	0.103	0.499

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Appendix L Experimental Plan of the AISA Project

The AISA project aims to study the effects that automation of ATCO monitoring tasks has on human-machine team situation awareness. The overarching research question is stated in the project's fundamental documents – can the AI system be made aware of the traffic situation (in the narrow field of en-route ATC operations) and can that artificial awareness provide transparency and generalization to automated systems? We posit that the "machine can be aware of the situation, including its state, in a domain-specific way, and it can take part in the team situational awareness and that such system can be used to automate monitoring tasks in transparent manner." (AISA Consortium, 2019). This document serves as an overview of all research questions – technical ones that guide the development of system components and general ones which guide the overall project. Each research question, combined with related requirements, informed the creation of objectives. Therefore, identified objectives can be used to confirm technical achievements (technical objectives) or can be used to validate overall system achievements (validation objectives). The objectives are described along with methods proposed for their achievement and relevant results available at the time of submission of this document (May 2022). For detailed results explanation, each objective has a reference to the relevant deliverable.

Technical objectives are mostly related to the AISA system components, so the relevant documents are WP2, WP3, and WP4 deliverables. Deliverable 2.2 presented the requirements for the system described in D2.1, so they informed the creation of technical objectives. After the development of the components (during WP3 and WP4), execution of 2 experiments helped answer the guiding question of the project and complete the validation objectives. As opposed to the technical objectives, no requirements were set for this part of the project, so they were not used for validation objective creation. The most important document pertaining to those parts of the Experimental plan is this deliverable (D5.2). It contains the sum of all experiment research questions (grouped by topics), complete methodology, and results obtained from the two experiments.

The descriptions of the two experiments differ slightly - experiment 2 has an additional section describing the differences between the initial plan and the final one, due to modifications necessary to answer relevant questions on ATCO and team SA. Both experiments, along with explanations for the introduced changes, are presented in the associated chapter.

The technical and validation objectives' sections will follow a similar structure by including research questions and relevant requirements (if any), the planned methodology, and results/data which answers the research questions and objectives.

L.1 Overall objectives

Overall system development was guided by the following research questions:

- *Is the proposed architecture flexible and fast enough for this purpose?* (Flexibility of the framework for knowledge graph management and reasoning)
- *What are the risks associated with the AI situational awareness system?* (Risk assessment)

Translation of these research questions into objectives and their solving is done in Table 28.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Table 28 The overall system research questions, technical objectivees, and results

Related research question	Objective related to Grant agreement	Methodology	Result
Is the proposed architecture flexible and fast enough for this purpose?	1.1. The proposed architecture should be flexible enough for the purpose of knowledge graph management and reasoning	Development of KG in such a way that it enables flexible adding/changing of data	Deliverable 4.3 (AISA Consortium, 2021f) shows the methodology and instances of data included in the knowledge graph, and what they are based on. Since some data did not exist in AIXM and FIXM, this deliverable proves the flexibility of the KG as more data ("plain" data) has been added, and some parts of FIXM and AIXM have been modified into a joint UML diagram on which the KG is based on. System performance analysis in this deliverable showed that the system provides answers at approximately the same rate as the ATCO radar refresh rate (approx. 5 s).
	1.2. The proposed architecture should be fast enough for the purpose of knowledge graph management and reasoning	Comparison of system performance (ATC knowledge graph creation, reasoning execution) to an equivalent system with a similar purpose	
Which are the risks associated with the AI situational awareness system?	1.3. Perform the risk assessment to determine the risks associated with the AISA system	Identification of potential hazards via application of an existing methodology. If an appropriate methodology cannot be found, a new one should be developed	Deliverable 5.1 (AISA Consortium, 2022)applied an existing ICAO methodology for risk assessment. Two metrics were defined to characterize each risk – likelihood and severity. In all, 74 hazards were identified, with 150 mitigation measures proposed which

EUROPEAN PARTNERSHIP
Co-funded by the European Union

			significantly reduced the number of non-acceptable and tolerable risks. Additionally, an AI hazard library was created to serve a risk assessment starting point for other AI projects.

System components can be seen in Figure 45, depicting the conceptual diagram of the proof-of-concept system described in the AISA ConOps. Figure 45 shows the main parts of the system – knowledge graph, reasoning engine, and machine learning modules. Research questions and technical objectives in this section are grouped according to the system components they pertain to, along with related sub-systems.

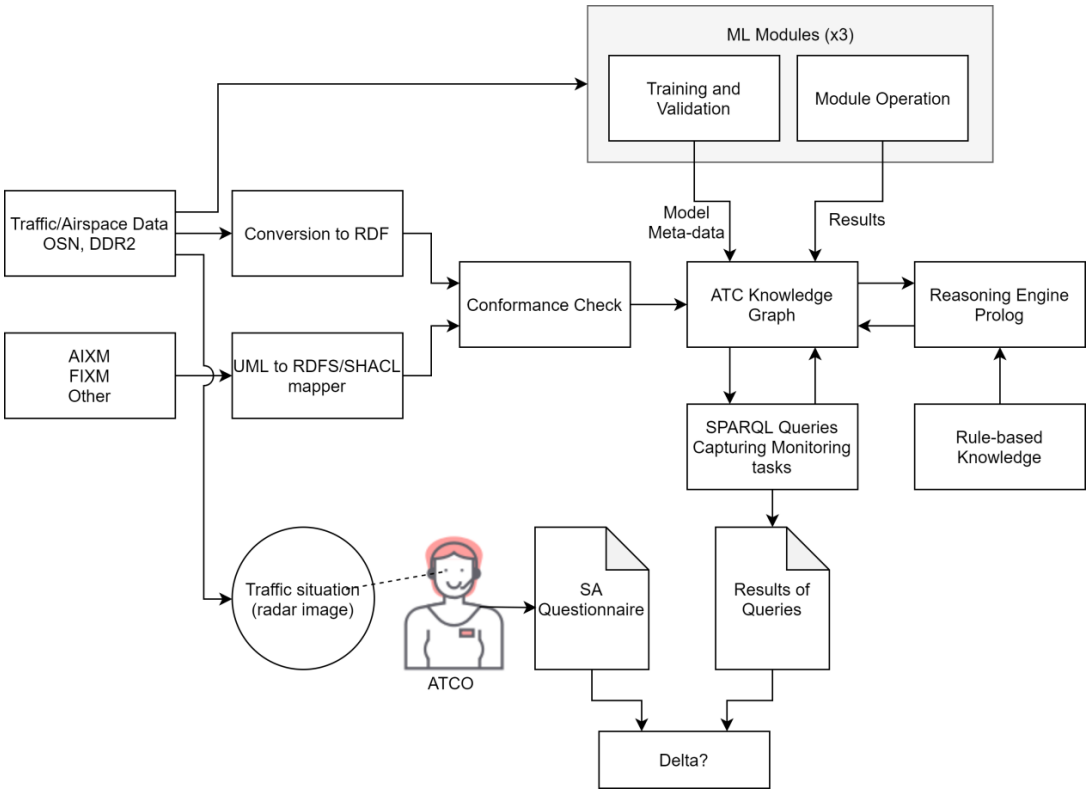


Figure 45 Conceptual diagram of the proof-of-concept machine situation awareness system

L.1.1 Knowledge graph and reasoning engine

The ATC knowledge graph, which serves as the central part of the AISA system, stores air traffic data and knowledge and provides meaning to the stored data. The developed AISA system's reasoning engine is based on knowledge graphs and interacts with the machine learning modules. A reasoning

EUROPEAN PARTNERSHIP
Co-funded by the European Union

engine is capable of explaining the obtained results and provide insight into inconsistencies and improbable results. This allows the system to be close to the process of thinking that ATCOs have when making decisions and conclusions. Research questions related to the knowledge graph and reasoning engine, as presented in the Grant Agreement (AISA Consortium, 2019), are:

- *How feasible is it to encode all required semantics for ATC en-route operations?*
- *Is first-order logic powerful enough for all types of queries that will be needed?*
- *Can a sufficiently fast query execution over a large triples store be achieved?*

The encoding of semantics for ATC en-route operations proved to be very feasible albeit time-consuming. Most of the required semantics could be found in Aeronautical Information Exchange Model (AIXM) and Flight Information Exchange Model (FIXM) and were assumed from these models. The missing information (such as Flight Level Allocation Scheme/FLAS) was identified and then added by creating a unified UML diagram based on which the encoding was done.

The first-order logic application ensures the system's SA reliability even in multiple modified environments. This logic is proven to be powerful enough to perform all achieved monitoring task queries. Tasks which were not achieved are mostly related to ML modules which hadn't been integrated with the rest of the system.

Fast query execution enables the continuous monitoring of traffic situations. When introducing large stores of triples, time consumption can be a limiting factor for a system to operate in real-time. The processing time analysis implies that the system is sufficiently fast as it takes 5s to process a single traffic situation graph. At the time of the analysis, hardware was adapted for the processing by using multiple cores. Query execution time can be further improved with hardware adaptation.

Knowledge-graph-related requirements from Deliverable 2.2 (AISA Consortium, 2020b) are divided into these categories:

- Requirements for Knowledge Graph and Reasoning Engine,
- Requirements for UML to RDFS/SHACL Mapper,
- Requirements for Proof-Of-Concept KG-System,
- Requirements for KG-Prolog Mapper,
- Requirements for Populating the KG,
- Requirements for Automation of Monitoring Tasks in Proof-of-Concept System,
- Requirements for Knowledge Engineering for En-route ATC Operations.

By combining the relevant research questions and requirements, objectives related to the knowledge graph and its sub-systems are presented in Table 29.

Table 29 The KG and its subsystem research questions, objectives, and results

Research question or requirement	Technical objectives related to Requirements in D2.2	Methodology	Result
Classes and properties used in the knowledge graph shall be	2.1. RDF Schema must be used in the KG to define classes and properties	Appropriate RDF Schema development prior to creating traffic data instance graphs	The UML to RDFS/SHACL Mapper was developed. The RDF Schema consists of existing classes defined by aeronautical exchange

EUROPEAN PARTNERSHIP
Co-funded by the European Union

defined using RDF Schema (RDFS)			formats (AIXM, FIXM) and user-defined classes. More information about the Mapper can be found in Deliverable 4.1 (AISA Consortium, 2021g) while Deliverable 4.3(AISA Consortium, 2021f)contains more information about RDF instances, classes and properties.
KG shall be queried with SPARQL	2.2. KG must be queried with SPARQL	Use of SPARQL queries in reasoning engine development/ATCO task implementations	Java task implementations contain and execute SPARQL queries to access data stored in the KG. Used queries can be found in Deliverable 4.4. (AISA Consortium, 2021e)
Data instances shall be provided in RDF	2.3. Data instances must be provided in RDF	ESCAPE Light data log will be exported to XML format and converted to RDF	Knowledge graphs are converted from XML to RDF format to be used as AISA system inputs. Additionally, the conversion from XML to RDF has been automatized. Used RDF instances can be found in Deliverable 4.3 (AISA Consortium, 2021f)
SHACL shall be used to check the instance data against the constraints and validation report shall be provided in RDF	2.4. SHACL must be used to check the instance data against the constraints and validation report must be provided in RDF	Appropriate development of UML to RDFS/SHACL Mapper	The UML to RDFS/SHACL Mapper has been developed. It successfully checks the instance data against the SHACL constraints, and the validation report is shown in RDF. Details can be found in Deliverable 4.1 (AISA Consortium, 2021g)
Mapper shall process UML class diagrams in XMI format	2.5. Mapper must process UML class diagrams in XMI format	Appropriate development of UML to RDFS/SHACL Mapper	The UML to RDFS/SHACL Mapper has been developed (AISA Consortium, 2021g) and it is also referred to as the AISA XMI mapper. The mapper takes as input an UML class diagram that is represented in XMI format.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Mapper shall process AIXM and FIXM UML diagrams in full	2.6. Mapper must process AIXM and FIXM UML diagrams in full	Appropriate development of UML to RDFS/SHACL Mapper	The mapper successfully processes AIXM and FIXM UML diagrams in full as it comes with a plug-in architecture and has a FIXM plug-in and an AIXM plug-in. The plugins are described in Deliverable 4.1 (AISA Consortium, 2021g)
User should be able to select a subset of AIXM and FIXM to process	2.7. User can select a subset of AIXM and FIXM to process	Appropriate development of UML to RDFS/SHACL Mapper	A configuration file is provided as input to the mapper and based on it, selected subsets of models (chosen by the user in the configuration file) are extracted by the extractor module. These extracted subsets of models are mapped by model-specific plugins to RDFS/SHACL documents and provided as RDF/XML files. More details can be found in Deliverable 4.1(AISA Consortium, 2021g).
Mapper shall process other UML diagrams (outside of AIXM and FIXM) if provided in the same format as AIXM and FIXM.	2.8. Mapper must process other UML diagrams (outside of AIXM and FIXM) if provided in the same format as AIXM and FIXM	Appropriate development of UML to RDFS/SHACL Mapper	plain.xq is one of the three plugins for the mapper and it targets models which are not based on AIXM and FIXM (but are in the same format) and do not use stereotypes meaning they can be processed as well. The mapper and the plug-in are explained in Deliverable 4.1 (AISA Consortium, 2021g).
Instance data shall be imported into the KG in RDF	2.9. Instance data must be imported into the KG in RDF	Exported traffic data should be converted to RDF format during pre-processing. RDF graphs should be stored in AISA system input folders to enable their import into the KG	The RDFS/SHACL documents are in RDF/XML format, but it is easy to transform from it to Turtle RDF syntax. One approach is the functionality from Apache Jena and this approach is demonstrated by the TransformXML2TTL.java program further explained in

EUROPEAN PARTNERSHIP
Co-funded by the European Union

			Deliverable 4.1 (AISA Consortium, 2021g).
KG shall provide RDF graph store	2.10. KG must provide RDF graph store	The KG uses an RDF graph store on the Apache Jena TDB (a free RDF database used to store and query RDF data)	All the basic functionality for working with RDF, RDFS, SPARQL, and SHACL in Java is provided by Apache Jena and as the RDF graph store Jena TDB is used. Chapter 3 of the Deliverable 4.1 talks about the graph storage. (AISA Consortium, 2021g).
KG shall provide SPARQL endpoints	2.11. KG must provide SPARQL endpoints	Creation of SPARQL endpoints	A local Jena Fuseki server is used to store data by using the predefined dataset URL. Three SPARQL endpoints (for updates, queries, and graph store protocols) are also defined.
KG shall provide reasoning/entailment over RDF graphs	2.12. KG must provide reasoning/entailment over RDF graphs	Use of KG-side reasoning during ATCO task implementation	KG-side reasoning is used in some ATCO task implementations(AISA Consortium, 2021f, 2021e).
KG shall provide SHACL processors for checking conformance between the knowledge graph and the schema and for executing inference rules encoded in SHACL	2.13. KG must provide SHACL processors for checking conformance between the knowledge graph and the schema and for executing inference rules encoded in SHACL	Appropriate development of UML to RDFS/SHACL Mapper.	The UML to RDFS/SHACL Mapper checks conformance between the KG and the schema and checks the instance data against the SHACL rules. Deliverable 4.1 provides more information about the subject. (AISA Consortium, 2021g).
KG-Prolog mapper shall receive results of the SPARQL queries or complete KG and convert them into predicates	2.14. KG-Prolog mapper must receive results of the SPARQL queries or complete KG and convert them into predicates	Appropriate development of KG-Prolog Mapper	Alternative ATCO task implementation method (Java classes) was chosen and used. Chapter 4 of Deliverable 4.4 gives a full explanation for the decision (AISA Consortium, 2021e).

Based on the results of the logic programs, mapper shall produce SPARQL update requests.	2.15. Mapper must produce SPARQL update requests.	Appropriate development of KG-Prolog Mapper	Alternative ATCO task implementation method (Java classes) was chosen and used. Chapter 4 of Deliverable 4.4 gives a full explanation for the decision (AISA Consortium, 2021e).
Based on the results of the logic programs, mapper shall produce Prolog rules	2.16. Mapper must produce Prolog rules	Appropriate development of KG-Prolog Mapper	Alternative ATCO task implementation method (Java classes) was chosen and used. Chapter 4 of Deliverable 4.4 gives a full explanation for the decision (AISA Consortium, 2021e).
Mapper shall update KG with results of SPARQL update requests and Prolog rules	2.17. Mapper must update KG with results of SPARQL update requests and Prolog rules	Appropriate development of KG-Prolog Mapper	Alternative ATCO task implementation method (Java classes) was chosen and used. Chapter 4 of Deliverable 4.4 gives a full explanation for the decision (AISA Consortium, 2021e).
Mapper shall derive shape of predicates (arity and ordering of attributes) not only from validating SHACL properties but also from non-validating properties	2.18. Mapper must derive shape of predicates not only from validating SHACL properties but also from non-validating properties	Appropriate development of KG-Prolog Mapper	Alternative ATCO task implementation method (Java classes) was chosen and used. Chapter 4 of Deliverable 4.4 gives a full explanation for the decision (AISA Consortium, 2021e).
Populating the KG shall be based on RDFS produced from AIXM, FIXM, ML outputs	2.19. Populating the KG must be based on RDFS produced from AIXM, FIXM, ML outputs	Basing the KG population on the RDFS produced from AIXM, FIXM, ML outputs, and other sources	The population of the KG was initially done manually and was later automated. Both methods relied on the RDFS produced from AIXM, FIXM, ML outputs and other sources.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

			Deliverable 4.3 explains this process (AISA Consortium, 2021f).
Populating the KG should be performed by data translators	2.20. Populating the KG may be performed by data translators	Data processing and conversion to be performed by data translators.	Initial traffic data processing and conversion to RDF was performed by data translators. The procedures were later automatized, which allows for quicker system operation. Chapter 5 of Deliverable 4.3 gives an overview of the beginnings of KG population (AISA Consortium, 2021f).
ML outputs and other aeronautical information not contained within the AIXM and FIXM shall be adapted for KG	2.21. ML outputs and other aeronautical information not contained within the AIXM and FIXM must be adapted for KG	Addition of ML module data and metadata and missing aeronautical information to the KG during RDFS development and KG population	During RDFS development and KG population missing information was added and fully adapted. Deliverable 4.3 mentions this missing information and its adaptation (AISA Consortium, 2021f).
A subset of monitoring tasks will be implemented during the project	2.22. More than 75% of defined monitoring tasks must be implemented	Development and implementation of monitoring tasks in the most efficient way	Around 80% of monitoring tasks have been successfully implemented in the Java programming language. Section 1.2.1.4 talks about the automated and implemented tasks.

L.1.2 Machine learning modules

The project uses the given traffic situation, predictions, and assessments based on machine learning to produce KG system outputs. Training data and module descriptions are provided to users via meta-data to assess and follow the module's performance. Three ML modules - conflict detection, trajectory prediction, and complexity assessment module - use different approaches for predictions to produce results from different sources. Research questions related to ML modules are:

- *To what extent can functioning of a ML sub-system be verified by a reasoning engine?*
- *What is the best way to integrate ML modules into the KG-based system?*

Machine learning objectives, shown in Table 30, are guided by the following groups of requirements:

- Trajectory Prediction ML Module requirements,
- Requirements for Conflict Detection module,
- Requirements for Air Traffic Complexity Estimation Module.

Table 30 The ML modules research questions, objectives, and results

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Related research question or requirement	Technical objectives related to Requirements (D2.2)	Methodology	Result
	Trajectory prediction machine learning module		
What is the best way to integrate ML modules into the KG-based system?	3.1. The module must receive input information from the KG.	The module uses a filed flight plan and actual aircraft state as input information	ADS-B data from The OpenSky Network and flight plan data from the DDR2 database from EUROCONTROL are used as input data (AISA Consortium, 2021a).
To what extent can functioning of a ML sub-system be verified by a reasoning engine?	3.2. The module must provide output information in a standardized format that can be exported to the KG.	The development of the module in such a way that it provides its output as a waypoints-grid	The output is given in a waypoints-grid which can be manually imported in the KG (AISA Consortium, 2021a).
The module shall provide KG with model meta-data.	3.3. The module must provide KG with model meta-data.	Export and inclusion of model metadata to the KG	The module metadata can be manually imported in the KG (AISA Consortium, 2021a).
The module shall use last known aircraft position for operation	3.4. The module must use last known aircraft position for operation	Use of actual aircraft position during the development of trajectory prediction	The module uses last known aircraft position for the dynamic part of the trajectory prediction (AISA Consortium, 2021a).
The module shall use weather data if possible	3.5. The module must use weather data if possible	Training the module on traffic data which includes weather influence	The module indirectly takes into account the weather data as it was the cause of certain re-routings that were used for training the neural network (AISA Consortium, 2021a).
The module shall provide trajectory prediction as a set of 4D points	3.6. The module must provide trajectory prediction as a set of 4D points	Development of the module so it provides the trajectory prediction in a geographical region as points in a waypoints-grid	The result of a trajectory prediction are points in a waypoints-grid in the time-domain (AISA Consortium, 2021a).
	Conflict detection machine learning module		
What is the best way to integrate ML modules into	4.1. The module must receive input information from the KG.	Use of the same data in KG system and ML module *operations*	Conflict detection ML module output for each aircraft pair and defined

EUROPEAN PARTNERSHIP
Co-funded by the European Union

the KG-based system?			timestamp (AISA Consortium, 2021d).
To what extent can functioning of a ML sub-system be verified by a reasoning engine?	4.2. The module must provide output information in a standardized format that can be exported to the KG	The module outputs are in xlsx form which are transferred in RDF from which they can be exported to the KG	KG system tasks that use module output to issue system outputs successfully. Deliverable 4.3 provides more information about this result (AISA Consortium, 2021d).
The module shall provide KG with model meta-data	4.3. The module must provide KG with model meta-data	Addition of ML module metadata to the KG	Metadata (in the form of training data statistics) added to the KG. See Section 1.2.1.3 for detailed description.
ML module should be able to perform prediction exploiting open-access libraries (e.g., Scikit-Learn or Tensor Flow)	4.4. ML module can perform prediction exploiting open-access libraries (e.g., Scikit-Learn or Tensor Flow)	Reviewing different open-source libraries and identifying which one can be used for the Conflict detection module prediction	ML algorithm is based on the Scikit Learn mentioned in Deliverable 3.2 (AISA Consortium, 2021d).
ML module shall be able to provide information about conflict or situations of interest between aircraft pairs	4.5. ML module must be able to provide information about conflict or situations of interest between aircraft pairs	Investigation of the new approach based on the ML techniques to identify conflict, SI, or safety metrics	ML module provides information about situations of interest between aircraft pairs (AISA Consortium, 2021d).
ML module should be able to provide information about safety metrics related to conflict or situations of interest between aircraft pairs.	4.6. ML module can provide information about safety metrics related to conflict or situations of interest between aircraft pairs.	New approach based on the ML techniques to provide information about the proposed safety metrics (such as minimum distance, time & distance to CPA)	The ML module provides information about safety metrics related to situations of interest between aircraft pairs (AISA Consortium, 2021d).
Complexity estimation machine learning module			
What is the best way to integrate ML modules into	5.1. The module must receive input	Development of the module and integration with the KG system	Currently, ADS-B data from OpenSky Network is used as ML module input. Traffic situation data stored in the

EUROPEAN PARTNERSHIP
Co-funded by the European Union

the KG-based system?	information from the KG		KG can be modified to serve as module input. (AISA Consortium, 2021c).
To what extent can functioning of a ML sub-system be verified by a reasoning engine?	5.2. The module must provide output information in a standardized format that can be exported to the KG	Development of the module to provide information in a standardized format for KG export	The module generates a Microsoft Excel table containing traffic complexity results (AISA Consortium, 2021c).
The module shall provide KG with model meta-data	5.3. The module must provide KG with model metadata	Definition and addition of module metadata in the KG	Module metadata can be added to the KG (AISA Consortium, 2021c).
The module shall provide sector-level air traffic complexity score on a scale from 1 to 5.	5.4. The module must provide sector-level air traffic complexity score on a scale from 1 to 5.	Appropriate module development	The ML module is not currently able to provide a complexity score on a scale from 1 to 5. Task conversion into complexity score is planned for further module development.

L.1.3 Overall evaluation objectives

Human, machine, and shared human-machine situation awareness are characterised and measured by research questions defined in the D5.2. These questions are identified and presented in Experiment 1 and Experiment 2 descriptions, later in the text. These questions are detailed, exact, and focused on one specific desired outcome. The broader questions, from which the aforementioned questions arose, are:

- *To what extent are human and artificial SA even comparable?*
- *Can questions used to assess SA in humans be effectively translated into SPARQL queries?*
- *What is the maximum level of SA that can be obtained with KGs and/or reasoning engines?*
- *Does the complexity of knowledge engineering make benefits of artificial SA system irrelevant?*
- *Is it safe to include AI situational awareness in conjunction with ATCOs in TSA? (Risk assessment)*

Objectives on human, machine or TSA are presented in the Table 31.

Table 31 The SA research questions, objectives, and results

Related research question	Validation objective	Methodology	Result
What is the maximum level of SA that can be obtained with KGs and/or reasoning engines?	6.1. Assess the accuracy of AI SA system for en-route air traffic monitoring tasks.	Binary categorization for preserved and degraded SA	The accuracy of the AI SA system is high. More about the results can be found in Section 4.4.1.
What is the maximum level of SA	6.2. Assess the accuracy of AI SA	Comparing ML module prediction values with	CD machine learning module provides 70%

EUROPEAN PARTNERSHIP
Co-funded by the European Union

that can be obtained with KGs and/or reasoning engines?	system ML modules	the actual values and categorisation of results based on the defined limit	accurate predictions compared to the 12 NM miles SI limit. The analysis is presented in Section 4.4.2.
What is the maximum level of SA that can be obtained with KGs and/or reasoning engines?	6.3. Assess the level of situation awareness that AI SA system can reach	Assessing awareness level by using specific framework	The AI SA system is conditionally an Awareness Level 5 system. This conclusion is presented in Section 4.4.4.
Is it safe to include AI situational awareness in conjunction with ATCOs in TSA?	6.4. Assess the impact of AI SA system inputs on human situation awareness	Human performance measures based on questionnaire answers	ATCOs judged some AI SA inputs useful, but most inputs considered irrelevant (depending on the AI SA information) Section 4.3.2.
Is it safe to include AI situational awareness in conjunction with ATCOs in TSA?	6.5. Assess the impact of AI SA System on human performance	Measuring the effects of AISA system outputs in human-in-the-loop simulations	AISA inputs helped to perceive and solve conflicts earlier than without the input. Section 4.3.1.
To what extent are human and artificial SA even comparable?	6.6. Assess and compare ATCO and AISA system SA	Compare ATCO SA with AI SA system outputs on identical traffic situations	There is comparability in many answers to SA query. Aside this, AI SA system detected conflicts that none or only some of the ATCOs mentioned. False alarm and misses occurred for both. See Section 4.2.
Is it safe to include AI situational awareness in conjunction with ATCOs in TSA?	6.7. Assess the acceptance and applicability of AI SA System	Organization of risk assessment session	The AISA system could be considered safe with current conditions after the implementation of mitigation measures discussed earlier in this deliverable.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

L.2 Experimental Approach and Research Questions

Experiment 1

Research questions related to experiment 1 are the ones making up topics "Human Situation Awareness" and "Accuracy of Artificial Situation Awareness" in this deliverable. Those questions are:

- 1.1 *What characterizes ATCOs' scanning patterns and priorities?*
- 1.2 *Are different measures for situation awareness (self-ratings, queries, gaze-based analysis, and implicit measurements) significantly interrelated according to their meaning?*
- 4.1 *Can the monitoring tasks be applied to the KG to achieve situational awareness?*
- 4.2 *Does the CD machine learning module provide accurate results regarding situations of interest?*
- 4.3 *Does the CD machine learning module provide accurate results regarding conflicts?*
- 4.4 *Does the AISA system check the status of its sub-systems?*

From the posed research questions, it's obvious that they deal either solely with human SA or solely with machine SA. This is logical because, before the ATCOs complete traffic simulation exercises and generate data, there is no data for the AISA system to process and to generate its SA. For this purpose, the following plan was created:

1) The experiment will be performed on-site, in skyguide facilities in Dübendorf, Switzerland. Skyguide ATCOs should be informed about the experiment in advance and asked for voluntary participation. Volunteers should not be disqualified on the basis of total work experience nor experience with the chosen simulation tool.
2) Traffic scenarios will be developed for the chosen simulation tool – EUROCONTROL Simulation Capabilities And Platform for Experimentation (ESCAPE), Light version. The system usually used by skyguide ATCOs, SkyVisu, is superior regarding available ATCO tools, but it is not available outside of skyguide. ESCAPE Light will be adapted by adding standard measuring tools the ATCOs use in daily operations, but the persisting differences between the two operating systems must be noted as a potential limitation of the experiment.
3) Since skyguide ATCOs will be taking part in the experiments, Swiss en-route traffic data should be used. The original traffic data will have to be modified because it was "controlled" by ATCOs and contains no situations of interest. Changes should be performed by FTTS personnel in accordance with skyguide SME's instructions. Introduced changes must result in specific conflicts and situations, so ATCO answers can be compared to predicted answers. The raw traffic data used for traffic simulations should also be excluded from the training data of the ML modules, so the proof-of-concept system can be tested for generalizability.
4) The experiments will be performed on computer pairs – one will host the experiment and serve as the pseudo-pilot position, while the other will be used exclusively as an ATCO working position. If possible, multiple scenarios may be run simultaneously.
5) ATCOs will complete each prepared scenario, starting with a training scenario which would introduce them to the ESCAPE Light system and the differences between it and SkyVisu. A low complexity scenario should follow the training scenario, to further improve familiarity with the system before starting scenarios whose results will be used for the SA analysis. To avoid the

EUROPEAN PARTNERSHIP
Co-funded by the European Union

learning effect in the results, caused by certain scenarios being performed by all ATCOs later in the experiment and thus being affected by their improved grasp of the system, the remaining scenarios will be performed in a random order generated for each ATCO.

6) Data exported from ESCAPE Light for each scenario completed by an ATCO (referred to as an *exercise*) should be labelled in a way that reflects that information. Data logs should include aircraft positions, callsigns, altitudes (current, requested and cleared), headings (current, requested and cleared), speed (current, requested and cleared), trajectory lists and coordinates of important points (e.g. exit points). This data will be combined with ML modules' outputs and metadata to serve as inputs for the proof-of-concept system.

7) The AISA system tasks should be applied to each exercise separately, with the KG being emptied between each run. This will ensure that there is not data accumulation which could slow down the system nor data duplication which could cause erroneous outputs.

8) AISA system outputs should be analysed to determine if the system is aware of traffic situation elements. This can either be performed by formulating SPARQL queries targeting system outputs stored on the KG server or by printing all system outputs and analysing them to find answers to SA assessment questions. The second experiment should include questions on ATCO SA about the same considered traffic situation but in a shared human-machine SA environment.

Three ATCO SA assessment techniques will be used during the first experiment:

1) Subjective rating – performed after each scenario by ATCOs completing the SASHA_Q questionnaire, this technique consists of 6 questions with behavioural descriptions for aspects of SA. The scale starts at 0, for statements that are *never* true, and goes to 6, representing statements that are *always* true for them.

2) Gaze-based analysis – performed by using eye-tracking glasses and post-processing acquired data using available tools. The plan is to use Tobii Pro Glasses 3 and the associated Tobii Pro Lab software. Eye-tracking software define a concept of *area of interest*, which is the surface covering the screen element/feature that is being researched. Areas of interest can be either statis or dynamic. Since the most important elements of the traffic situation – the aircraft – are moving all the time, dynamic areas of interest will be preferable but can be substituted by sufficiently large static areas of interest. Eye-tracking data and simulation data will need to be synchronized for the analysis.

3) Implicit performance measurement – used for implicit assessment of ATCO SA, behavioural codes can be defined and compared to ATCO radio communications.

Additional analysis can include biometrical analysis (performed before, during, and after the experiment), including but not limited to pulse measurement and skin conductance. Acquisition of biometrical data must not affect ATCO performance – sensors and other equipment should not increase their workload in any way. Baselines for each biometrical parameter should be established to enable calculation of absolute and relative changes.

Experiment 2

To assess ATCO opinion on shared situation awareness, a second experiment is conducted. As the system isn't able to operate in real-time, the SPARQL query outputs of the AISA system are obtained from the previously recorded data. The research questions from the experiment 2 are:

- 2.1 *Are artificial and ATCO situation awareness comparable?*

EUROPEAN PARTNERSHIP
Co-funded by the European Union

- 2.2 *Can the AI SA system provide inputs to situation awareness that ATCOs were not aware of?*
- 3.1 *Is human performance enhanced by adding machine situation awareness?*
- 3.2 *Do ATCOs evaluate artificial situation awareness inputs as useful and trustworthy contribution to human-machine team situation awareness?*
- 3.3 *Do ATCOs use artificial situation awareness inputs for their situation awareness and decision making?*

The main goal of experiment 2 is to investigate if the human and machine situation awareness is comparable and what is the ATCO opinion on the AISA system output.

Initial plan

Originally, the plan was to provide ATCOs with the static figures of traffic situations while introducing AI SA system input. Several issues arose with the initial plan. As those would be the only figures, ATCOs:

- couldn't build SA prior to that traffic situation,
- wouldn't be provided with the tools that could enable measuring distances and hovering over the label to gain additional label data.

As human SA wouldn't be built, the measuring of degrading and preserving SA would be inappropriate to assess in that environment.

Revised plan

To accomplish the validation objectives, experiment 2 should be conducted by using human-in-the-loop simulations. The revised experiment concept provides ATCOs with the ability to build and preserve situation awareness while permitting them to use available system tools. The plan is as follows:

1. In Dübendorf (Switzerland), skyguide premises are used to execute experiment 2. Skyguide's licensed ATCOs are asked to voluntarily participate in experiment 2. Personal data is protected and encoded during the experiment's conduction. The selection of the ATCOs is not based on their experience or working shifts. During the experiment, two ATCOs provide participants with the SME's explanation and support.
2. The simulation platform ESCAPE Light is used again with the same available tools as in experiment 1. During the experiment eye tracking, screen recording, frontal recording and biometrical data are recorded. Two parallel working positions in the same room are accompanied by the pseudo-pilots from the Faculty of Traffic and Transport Sciences.
3. To build ATCO SA, scenarios are designed as follows: the training scenario, human-in-the-loop scenario and "watch-only" scenarios. The main difference from experiment 1 is the *interactivity*. As the AISA system can't operate in real-time, participant's SA is built based on the exercises performed in experiment 1, which means that the interactivity cannot always be accomplished. A training scenario is used to prepare and familiarise participants with the ESCAPE Light system. This is followed by an interactive scenario and a watch-only scenario. Thereafter the order of the remaining watch-only scenarios is randomized. Participants in experiment 2 should be different from participants in experiment 1 to avoid differential gain in experience with the given system. In the "watch-only" scenarios ATCO and pilot voice are synthesised to ensure anonymity.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

4. ATCO situation awareness is measured additionally with probe technique (SASHA_L) during experiment 2. After each query, simulation is shortly frozen (20'', up to 30'' in case of more extensive questions) to allow ATCOs to answer.
5. AISA system inputs are hand-picked to avoid overloading ATCOs with the vast amount of information regarding AISA system monitoring task outputs. The selection of input, time, and notification format is agreed upon and defined with SME's assistance. The selection of system inputs was focused on those traffic situations that imply SA degradation.
6. AISA system inputs are provided to ATCOs after they answered the queries. As the HITL simulation is a modification to the original evaluation plan, and no HMI design was planned for the AISA project, AISA inputs are provided in an audio format and at predefined times - after a query on aspects of situation awareness is answered.

L.3 Variables

The independent variable manipulated is AI SA support. Two conditions are investigated: "no AI SA support" in experiment 1 vs. "AI SA support" in experiment 2 (Figure 46). This will be achieved by providing auditory input on machine situation awareness (AI SA output to SPARQL query on data form experiment 1) to ATCOs in experiment 2.

The dependent variable is human performance. This is measured in terms of start of conflict solution and conflict duration.

Aside this ATCO reactions to AI SA support are investigated with questionnaires during and at the end of experiment 2 to evaluate usefulness and trust.

As a possible intervening variable mediating the relationship of support of AI SA system in situation awareness and human performance scenarios are selected that vary in task load.

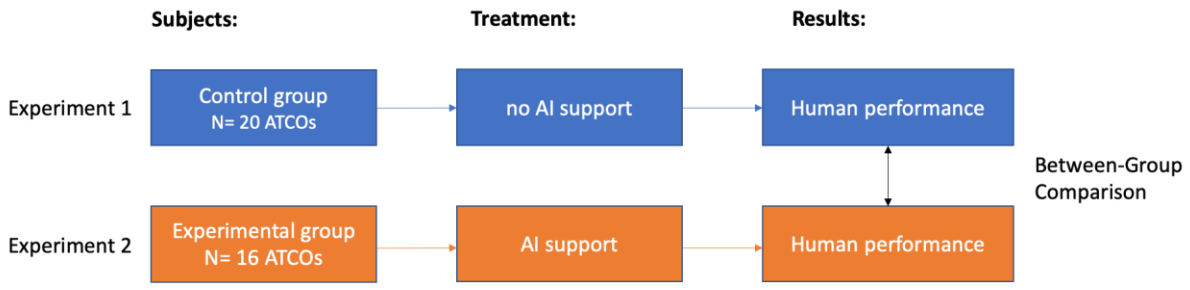


Figure 46 Experimental design to evaluate the effect of AI SA support on human performance

In further simulations using data collected in experiment 1, accuracy of ML estimations and predictions was assessed. For this SPARQL queries were used and ATCO interactions after the time of query were ignored to assess the correctness of AI SA system output to query (e.g. Conflict detection ML module output for predicted distance to minimum distance regarding SI aircraft pair). Variables measured are performance on air traffic monitoring tasks (operationalised for preserved and degraded SA), accuracy of conflict detection ML module to make predictions regarding situations of interest and regarding conflicts and analysis of the level of situation awareness that AI SA system can partially or fully achieve. For this the concept of Jantsch and Tammemäe (2014b) was used that differentiates 5 awareness levels. According to the framework, AISA system is conditionally an awareness level 5 system. *Conditionally,* because of the current method of checking the system inputs – the SHACL rules. If its

EUROPEAN PARTNERSHIP
Co-funded by the European Union

functioning is bolstered by the implementation of another layer of checks, the estimation of the awareness level could be confirmed.

In order to determine whether real-time operations are feasible, an analysis of graph runtime was made. A median of runtime per graph depicted in seconds for number of aircraft shows how the runtime increases with the number of aircraft. Depending on various parameters, a runtime per graph ranges from under 4 s to over 10 s. Considering ATCO station refresh rate is every 5 s, this is the realistic rate with which to compare the AI SA system processing times meaning real-time operations are achievable.

ML module accuracy determination focused not only on the comparison of initial and final predictions to the actual distances and time in the scenario, but also on the analysis of the statistical data of each aircraft in analysed aircraft pairs. This information was used to try to find the correlation between the data used for training the ML module and the accuracy of the predicted minimum distances. Multiple correlation analysis showed that these variables are not statistically related.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Co-funded by
the European Union