D6.2 Validation report

Deliverable ID:	**6.2**
Dissemination Level:	**PU**
Project Acronym:	**ARTIMATION**
Grant:	**894238**
Call:	**H2020-SESAR-2019-01**
Topic:	**Digitalisation and Automation principles for ATM**
Consortium Coordinator:	**MDU**
Edition date:	**27 December 2022**
Edition:	**00.02.00**
Template Edition:	**02.00.03**

EXPLORATORY RESEARCH

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Authoring & Approval

Authors of the document

Name / Beneficiary	Position / Title	Date
Ana Ferreira	Deep Blue	2022-12-06
Nicola Cavagnetto	Deep Blue	2022-12-06
Pietro Aricò	UNISAP	2022-11-20
Augustin Degas	ENAC	2022-11-24
Christophe Hurter	ENAC	2022-10-20
Mobyen Uddin Ahmed	Professor at MDU	2022-11-15
Waleed Jmoona	Researcher at MDU	2022-12-05
Shaibal Barua	Researcher at MDU	2022-12-04

Reviewers internal to the project

Name / Beneficiary	Position / Title	Date
Stefano Bonelli	Deep Blue	2022-11-06
Pietro Aricò	UNISAP	2022-12-07
Christophe Hurter	ENAC	2022-12-07

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
Mobyen Uddin Ahmed	MDU	2022-11-06
Christophe Hurter	ENAC	2022-11-06
Pietro Arico'	UNISAP	2022-11-06
Ana Ferreira	Deep Blue	2022-11-06

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	2022-09-05	Draft	Ana Ferreira	Initial structure
00.00.02	2022-10-25	Draft	Ana Ferreira	Second structure
00.00.03	2022-10-27	Draft	Augustin Degas	First draft of Section 3.1.1
00.00.03	2022-10-28	Draft	Augustin Degas	First draft of Section 3.4
00.00.04	2022-11-17	Draft	Waleed	Chapter 4
00.00.05	2022-11-19	Draft	Mobyen	4.2.1
00.00.06	2022-11-22	Draft	Pietro Aricò	3.1.3, 3.2.5, Results
00.00.07	2022-11-22	Draft	Ana Ferreira	3.3
00.00.08	2022-11-22	Draft	Nicola Cavagnetto	4.4
00.00.09	2022-12-06	Draft	Nicola	Update with review
00.01.00	2022-12-07	Final	Mobyen	Ready for Submission
00.01.01	2022-12-22	Draft	Ana	Updated with SJU comments
00.01.02	2022-12-26	Draft	Augustin	Update with SJU comments
00.02.00	2022-12-27	Final	Mobyen	Ready for submission

EUROPEAN PARTNERSHIP

Co-funded by the European Union

ARTIMATION

TRANSPARENT ARTIFICIAL INTELLIGENCE AND AUTOMATION TO AIR TRAFFIC MANAGEMENT SYSTEMS'

This [Deliverable D6.2 Validation report] is part of a project that has received funding from the SESAR Joint Undertaking under grant agreement No 894238 under European Union's Horizon 2020 research and innovation programme.

Abstract

This report presents the results of the validation activities of both the Use Case 1 – Conflict Resolution and Use Case 2 – Delay Prediction, reporting the deviations from the D6.1 – Validation Plan as well. Both the use cases are introduced in terms of description of the tool, realisation of the tool, experimental setup, activities reporting of the development of the validation activities and final results. Finally, some conclusions about the results of ARTIMATION are drawn.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

List of Contents

EUROPEAN PARTNERSHIP

Co-funded by the European Union

List of Figures

EUROPEAN PARTNERSHIP

Co-funded by the European Union

EUROPEAN PARTNERSHIP

Co-funded by the European Union

List of Tables

EUROPEAN PARTNERSHIP

Co-funded by the European Union

1 Introduction

1.1 Purpose and scope of the project and document

This document presents the results of ARTIMATION concepts validation sessions in terms of Conflict Resolution visualisation tool (UC1) that was developed by ENAC, and the Delay prediction and propagation tools validation (UC2) carried out by MDU. It reports the validation objectives guiding the experimental approach, the validation preparation activities for both XAI concepts, how it was carried out, the results and outcomes achieved and, finally, the conclusions.

Both validation experiments aimed at exploring the impact of Explainable Artificial Intelligence (XAI) assistants on Air Traffic Management (ATM) in two scenarios, en-route, and tower, investigating en-route and tower Air Traffic Controllers (ATCOs) acceptance and human performance in different Explainable AI conditions (i.e., visual and algorithmic explanation).

The validation activities, carried out during month 19, 20, 21 and 22 of the project, deviated from the initial validation plan for both the use cases.

In particular, the Conflict Resolution use case deviated in the experimental design: the 3D visualisation in virtual reality environment was tested in different conditions than the screen-based XAI assistant. Therefore, three conditions were analysed: a black box level, an Explainable AI showing few information, and an Explainable AI showing a lot of information.

The Delay Prediction tool deviated from the validation plan as well. The validation took place online, after the development of a tool usable without the assistance of the researchers, which was sent out to Tower Air Traffic Controllers to test the Delay Prediction tool.

There are few deviations from the validation plan for the KPIs and success criteria as well.

For the Conflict Detection use case, the stress has been measured not only with the Galvanic Skin Response (GSR), but also with the Mindtooth Electroencephalography (EEG) system, developed by the UNISAP group within another EU project (Mindtooth, GA950998). Moreover, a sub-objective for the Human Performance assessment was added: we assessed the impact on work performance through self-report questionnaires as well, in addition to the task performance (from simulator system logs) of the Air Traffic Controllers.

The results and outcomes of this deliverable are a main input for the D7.1 Report on guidelines on AI transparency and Generalisation.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

1.2 Deliverable structure

The deliverable is structured as follows:

Chapter 1 introduces the project and the validation activities, filling the gaps between the Validation Plan and the Validation Report;

Chapter 2 lists the validation objectives from the validation plan

Chapter 3 presents the specifics of the validation activities for the Conflict Resolution Use Case (UC1) and the outcomes from the validation activities.

Chapter 4 illustrates the validation activities for the Delay Prediction Use Case (UC2) and outcomes from the validation activities.

Chapter 5 concludes the document with some considerations, recommendations and lessons learnt.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

1.3 List of acronyms

Table 1. List of Acronyms

AI	Artificial Intelligence
ANACNA	Associazione Nazionale Assistenti e Controllori della Navigazione Aerea
ANOVA	Analysis of Variance
ATC	Air Traffic Control
ATCO(s)	Air Traffic Controller(s)
ATFCM	Air Traffic Flow and Capacity Management
ATM	Air Traffic Management
BB	Black Box
CD&R	Conflict Detection and Resolution
DALEX	model-Agnostic Language for Exploration and eXplanations
DBL	Deep Blue
DP	Delay Prediction
DP&P	Delay Prediction and Propagation
Dx.x	Deliverable x.x
ENAC	Ecole Nationale de l'Aviation Civil
ENE	East-North-East
EEG	Electroencephalography
GA	Genetic Algorithm
GSR	Galvanic Skin Response
HM	Heat Map
HMI	Human-Machine Interface
LIME	Local Interpretable Model-agnostic Explanation
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MDU	Malardalens University
ML	Machine Learning
nDCG	Normalized Discounted Cumulative Gain
NE	North-East
NM	Nautical Miles
NW	North-West
RF	Random Forest
SB	StoryBoard
SE	South-East
SHAP	Shapley Additive Explanations
SW	South-West
TSAT	Target Start-up Approval Time
UCX	Use Case X
UNISAP	Università di Roma "La Sapienza"
WSW	West-South-West
XAI	Explainable Artificial Intelligence
XGBoost	eXtreme Gradient Boosting

EUROPEAN PARTNERSHIP

Co-funded by the European Union

2 Validation objectives

The table below summarises the main objectives that were defined at Validation Plan level (D6.1) for the whole project even if the validations activities were carried out in different scopes and with different maturity levels. The following objectives have been created according to the ARTIMATION Grant Agreement (pag.123) whereas the validation activity consists in the identification of indicators to assess the human performance and the user acceptability qualitatively and quantitatively.

Table 2: Validation objectives for Conflict Resolution use case

ID	Objective	Sub-objective	Methods
VO_CDR_1	Assess the impact of different levels of explainability and different types of visualisation on acceptance	Level of Understanding	Questionnaire
		Level of Agreement	Questionnaire
		Level of Acceptability	Questionnaire Neurometrics (EEG)
VO_CDR_2	Assess the impact of different levels of explainability and different types of visualisations on human performance	Level of Usability	Questionnaire
		Situation Awareness	Questionnaire
		Trust	Questionnaire
		Mental Workload	Neurometrics (EEG)
		Stress	Neurometrics (GSR/EEG)
		Task Performance	Questionnaire and debriefing
VO_CDR_3	Investigate the correlation between acceptability and human performance		All of the above
VO_CDR_4	Assess system performance	Expected impact on the safety	All of the above
		Expected impact on ATM system	All of the above

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Table 3 Validation objectives for Delay prediction validation

ID	Objective	Sub Objective	Methods
VO_DP_1	Assess the impact of different levels of transparency on acceptance.	Level of understanding	Questionnaire
		Level of agreement	Questionnaire
		Level of Usability	Questionnaire (SUS)
		Level of acceptability	Questionnaire

EUROPEAN PARTNERSHIP

Co-funded by the European Union

3 Conflict detection and resolution visualisation results

3.1 Description of CD&R visualization concepts and implementation

3.1.1 Materials

As a reminder, we first briefly explain in this section the genetic algorithm used for the experimentation. For more detail, we ask to see deliverable D4.2 and D5.1 [7][8]. We then present the implemented visualisation, with the rationales behind and the expected benefits from each one.

3.1.1.1 Genetic Algorithm used for the task

For this experimentation, we decided to use the Genetic Algorithm (GA) developed in Durand & Gotteland [15] presented here for ease of reading.

In short, a GA is a population and evolutionary based Meta-Heuristic. This means that a GA tries to iteratively improve a population of candidate solutions according to some predefined criteria, see Figure 1.

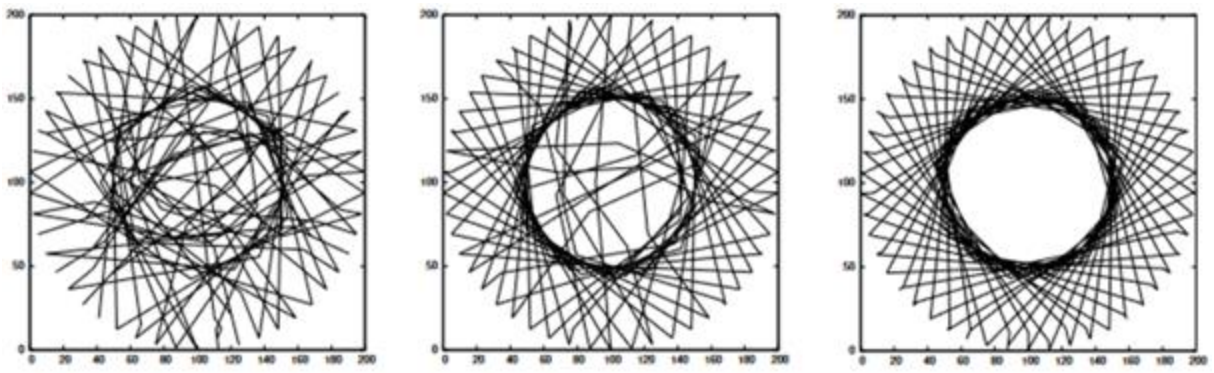


Figure 1: Evolution of the best candidate solution in function of the generation (i.e., iteration) for a conflict with 50 airplanes, from left to right: at the 250th, 500th, and 1000th generation. Taken from Durand & Gotteland [15]

In our conflict resolution case, a candidate solution (a chromosome for the GA) is a set of trajectories, some modified, some not. In the GA version used, the trajectories can only be modified by using turning point manoeuvre (see Figure 2}). As such, airplane trajectories are characterized by three parameters α, t_1, t_2—on top of fixed criteria such as the planned trajectory, the aircraft type, and the speed in this version.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

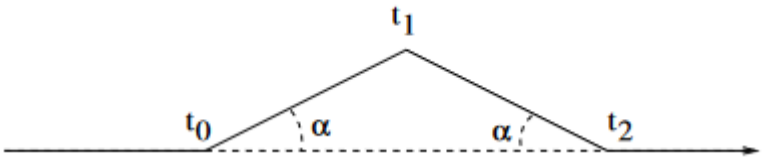


Figure 2: The turning point manoeuvre: the airplane is given a heading change of α degrees at time t0, and then return to its planned trajectory at t1. Taken from Durand & Gotteland [15]

Chromosomes (i.e., candidate solutions) forming the population at a generation (i.e., an iteration) are evaluated in function of three criteria: the duration of the conflicts, if any; the length of the trajectories; and the number of order one ATCO must give to implement this candidate solution. Once the GA has evaluated all candidate solution in the population, it selects a set of candidate solution, mostly the bests, but also other candidate solution to better explore the solution space. The algorithm then applies a set of mutation and crossover operations to enhance the population and to possibly converge toward one of the optimal solutions.

3.1.1.2 Screen based eXplanations

As previously explained, our GA algorithm explored the possible solution for a given conflicting situation between aircraft and extracted one solution which is qualified as the “best” one with the given optimization criteria (number of actions, length of the trajectory, and number of orders). To provide explanation for the proposed solution, we developed three different type of data presentation which are detailed in the following.

Black box: This visualization is a simple as possible and only displays the proposed solution by the GA algorithm, enhanced by instructions to proceed, see Figure 3.

Airplane trajectories are coloured differently. The minimal distance between airplanes is computed and displayed in yellow. The ordering (1st, 2nd …) of the control orders that must be given by the ATCO to the different airplanes are placed along the trajectory. This data presentation is not an explanation by itself but the simple data presentation of the “best” solution the GA algorithm managed to extract. Compared to existing system, the Black Box data representation directly provide a solution to a detected conflict, while the system currently used only displays the detected conflicting aircraft without further information to solve it. The Blackbox was the baseline of the validation and was

EUROPEAN PARTNERSHIP

Co-funded by the European Union

designed to give the solution proposed by the GA in the least intrusive way, providing only essential information.

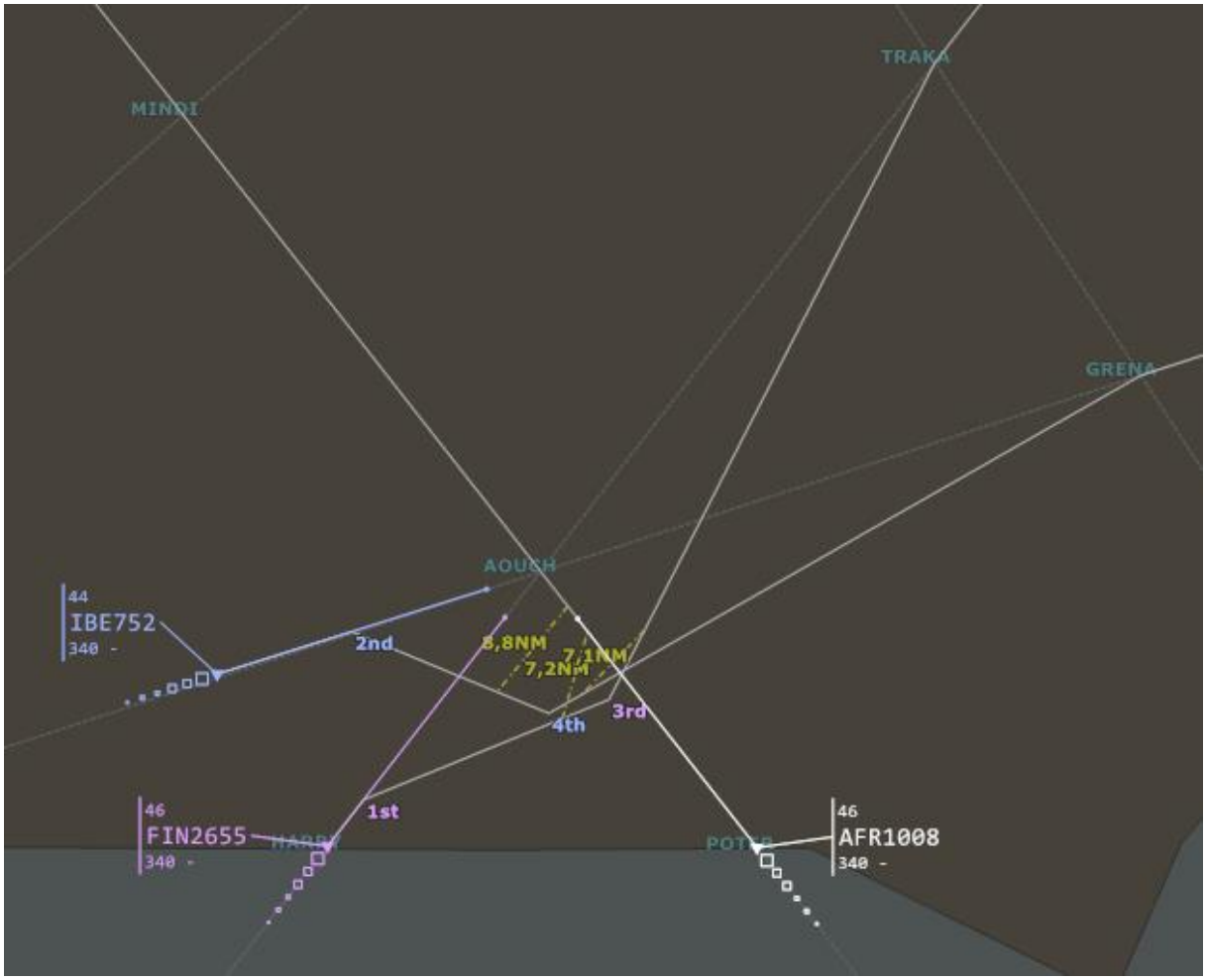


Figure 3: Blackbox visualization the solution proposed by the Genetic Algorithm

Heat map: To better explain the reasoning behind the proposed solution was made, we decided to show on top of the proposed solution what was explored by the GA, and if whether it was good or bad. To do so, we created heatmaps of the explored trajectories showing how aircrafts trajectories can safely be modified.

In Figure 4, the operator can see that: AFR1008 can only follow its trajectory. IBE752 and FIN2655 cannot follow their trajectories and need to turn right (only possibility). In addition, the user can see how much he can wait to turn each airplane, by seeing the end of the “safe zone” (green area) and the begin of the “dangerous zone” (red area). Such data representation is generated with the cumulative view of good and bad solutions. Each solution is convoluted with a gaussian kernel and then accumulated into a density map. Such technique helps do visually define areas also called contour maps (Scheepens et al., 2011). The heatmap was designed to create an uncluttered view of the possible modification of the aircraft trajectories, clearing the least interesting candidate solutions, and supporting contrastive questions such as “does the IBE752 can go left of straight despite the solution

EUROPEAN PARTNERSHIP
Co-funded by the European Union

turning it right?" (it should not, given Figure 4) or "what happens if I leave the IBERIA trajectory changed for longer amount of time" (it does not provoke any conflict with the FIN2655).

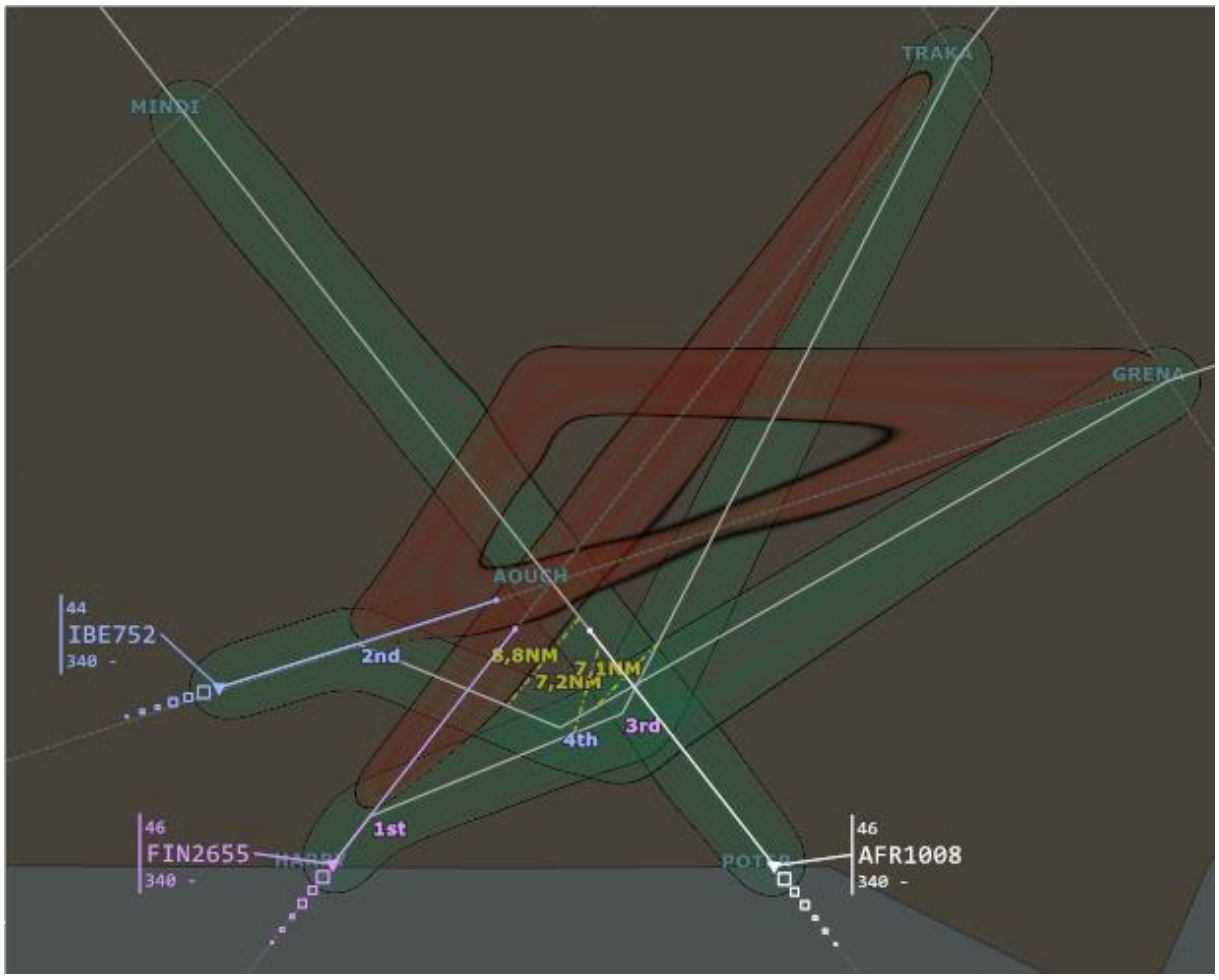


Figure 4: Heat Map visualisation of the solution proposed by the Genetic Algorithm

Storytelling: To better explain why the solution, the last visualization made, called Storyboard, shows two things: The timeline of the proposed solution by the algorithm, to see order after order where are each airplane, see Figure 5. Possibly and alternate solution, showing that other solutions can be made, but are less efficient. Limit solution, showing what needs to be done if the solution is not implemented right away to avoid any conflict. We use existing Data Driven Storytelling technique with step-based explanations and counterfactual explanations [32]. The Storyboard allows to better understand the

EUROPEAN PARTNERSHIP

Co-funded by the European Union

proposed solution by displaying the timeline, to give more trust in it. The alternative solution and limit solution aims at answering contrastive questions, and reinforce proposed solution.

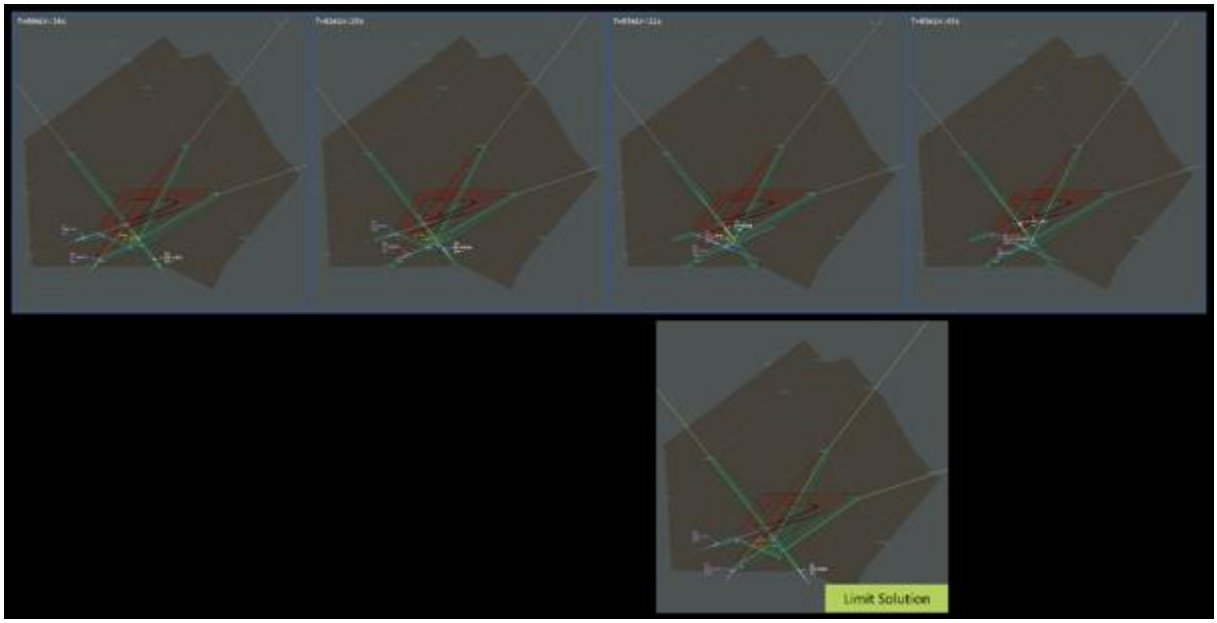


Figure 5: Storyboard presentation of the proposed solution. The sequence of images 1 to 4 shows the temporal steps to solve the conflict. The limit solution shows the good solution but with is close the minimum separation criteria between conflicting aircraft. Detail of the figure can be observed in Appendix D: Storyboard detail.

3.1.2 Validation preparation

The validation procedure of the CD&R tool was designed to 1) primary tests the different level of explainability (Blackbox, Heatmap, Storyboard), while 2) decreasing as much as possible any risk of bias, 3) maximizing the quality of neurophysiological measures, and 4) keeping the experiment short enough.

To entirely focus the validation on the different level of explainability, we decided to focus primary on the levels of explanation. To do so, we decided to remove any activity induced by aircraft continuing to flight while displaying the solution and the level of explanation, by stopping the simulation (see Figure 6). This meant that the participants were focusing solely on the mean to explain, this meant that the neurophysiological measures were not disturbed by other activities, but also that the task was abstracted for usual context.

As such, every scenario of conflict was following 4 different steps represented in Figure 7: 1) video of the simulation with the conflict that has to be solved and the surrounding contextual aircrafts (during $\delta t = 45s$), to gain situational awareness, and emulate the classical work environment; 2) displaying one type of explanation (Blackbox, Heatmap, Storyboard) during a fixed time ($\Delta t = 60s$), with the possibility

EUROPEAN PARTNERSHIP

Co-funded by the European Union

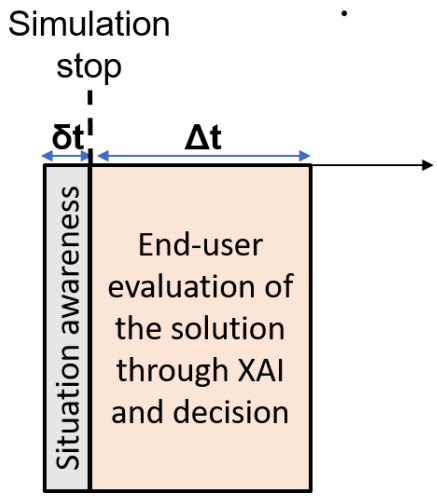


Figure 6: End user evaluation

to go to the next phase after 30s, to avoid boredom and disengagement; 3) ask the user to draw the solution it want to give after seeing the solution proposed; 4) answer questionnaires.

Figure 7: Steps for each scenario

Every level of explanation was tested with three different scenarios. The first scenario was used as a warmup, and the two others where the one data was gathered for analysis (while still gathering data on the first one, as control data). To avoid any bias linked to the order of presentation, fatigue, or scenario, we used a latin square to mix scenarios with level of explanation, and the order of presentation of the different level of explanation. Figure 6 present one such distribution, where Black Box is the first level tested, then Heatmap, the Storyboard.

The number of scenarios presented was decided in such way that the total experiment—from briefing, setting neurophysiological sensors, testing each level of explanation, and final debrief—, was not exceeding 2h.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

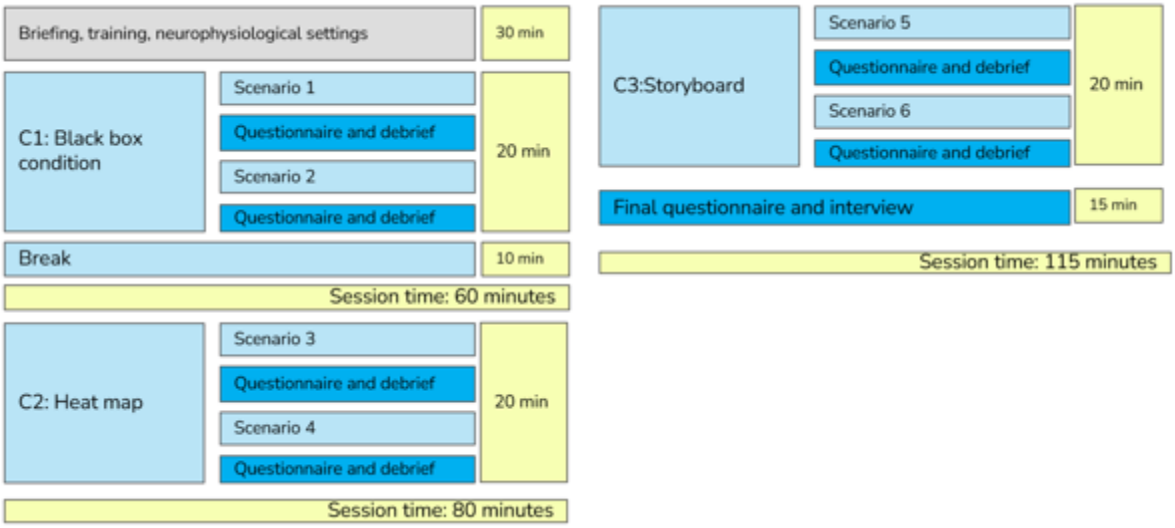


Figure 8: Experimental agenda per participant

3.1.3 Neurophysiological baseline recording

Before starting with the operational experimental activities, each controller has been asked to perform some "baseline" recordings, to collect neurophysiological data specific and of course different for each person, to be used in input to the algorithms for the evaluations of the neurometrics (see D4.1 for all the details regarding neurometrics evaluation algorithms through neurophysiological measures).

In this regard, three 1-minute different recordings have been performed, by collecting the raw EEG and GSR activities of the participants.

- During the first recording, it has been asked the participants to not move, be relaxed, and look at a blank screen. The aim of this recording run was to extract, in particular from the EEG activity, the eye blinks activity at rest, to be used as a template in the operational recordings, to remove eye blinks contribution.
- For the second one-minute recording, the participants have been asked to not move, be relaxed, and keep their eyes closed. This recording run was used to extract the Individual Alpha Frequency (IAF) value, to define all the frequency bands for the calculation of neurometrics.

During the third recording, each participant has been asked to not move, be relaxed, and look at the screen, showing a static representation of the ATM interface. This recording run has been used as a baseline, i.e., to normalize all the neurometrics of each participant with the proper baseline, before running the statistical analysis.

3.1.4 ATCO training

To familiarize the ATCO before the validation day, we sent a presentation about the experimentation, re-explaining the goal of the experimentation, and explaining, in detail, the different level of explanations (see Appendix A: Introduction Notice for CD&R).

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Before the validation, ATCO were shown the same slide to verify their knowledge about the tool, then shown another example with another scenario. Additional slides were presented, to familiarize the ATCO with the experimental setting, in particular the sector they were going to be controlling (see Appendix B: Additional Slide for CD&R on the Validation Day).

To familiarize with the platform, a training was made for the drawing part, and we introduced a warmup scenario for each condition, to strength the theorical training.

3.1.5 Simulation shakedown

Prior to the validation exercise, the validation procedure has been tested in four different phases: 1) the validation was first tested with 2 participants knowledgeable on AI, but not in Air Traffic Control (ATC), to verify logs, and the program was functioning, and create the first introduction slides; 2) the validation was then tested with 3 participants knowledgeable on AI, and in ATC, to start adjusting timings, and assessing the time of the full validation procedure, and adjust the presentation slides; 3) the validation was then tested with one ATCO expert, to finalise adjusting the timings, and verify the entire validation procedure was ready; 4) one last exercise was performed the day prior the start of the validation weeks, to test the validation procedure with the different partners.

3.2 Validation execution

3.2.1 Validation setup

The validation platform is presented in Figure 9. The participants were placed in ACHIL *En-Route* control setting. The control screen was either displaying the simulation, the solution and level of explanation, the drawing, or the survey (see chapter 3.1.2). Prior to the experiment, they were placed the neurophysiological sensors, a Galvanic Skin Response (GSR) recording device (shimmer sensing) and an electroencephalography (EEG) headset (Mindtooth), both linked to a Tablet embedding the

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Mindtooth recording suite, allowing a synchronized recording of both the signals, and to put specific markers used for the following offline analysis.

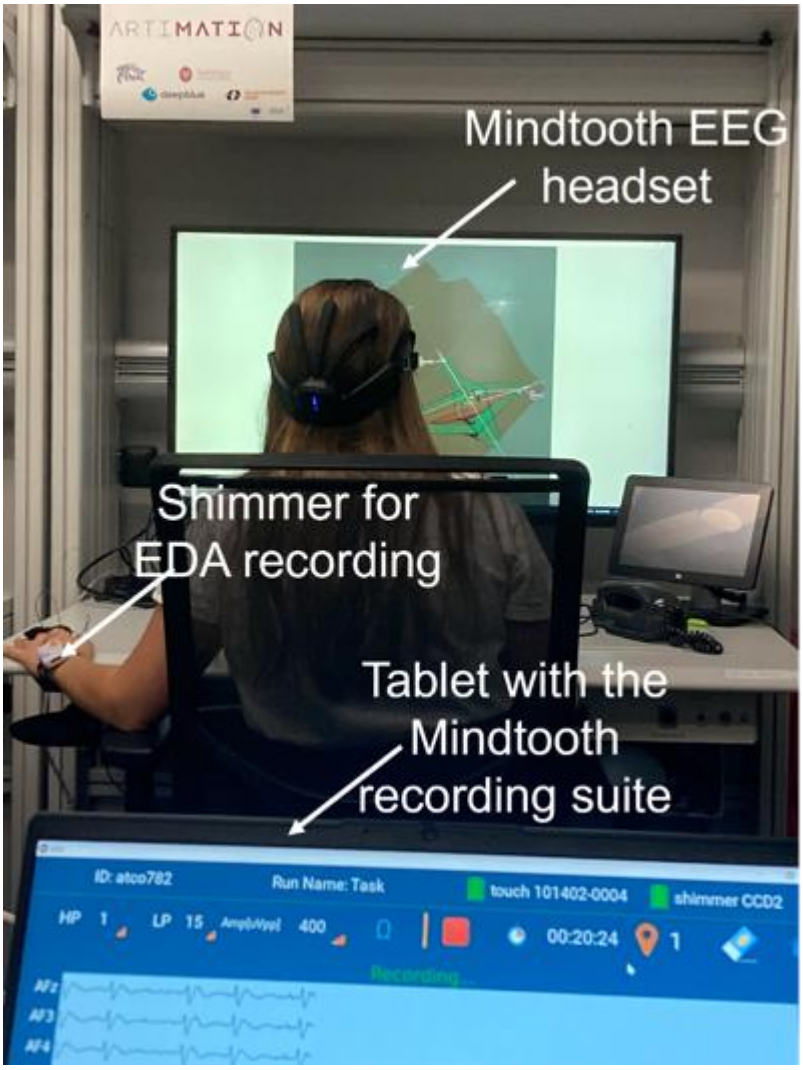


Figure 9: Validation setup for Neurophysiological Measures

3.2.2 Research personnel

Table 4: Research Personnel

Team	Personnel

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Validation simulation and conflict detection and resolution visualisation team [ENAC]	Christophe Hurter Augustin Degas Wided Oueslati
Neurophysiological measurements team	Pietro Aricò
Human Factors specialists' team	Ana Ferreira Nicola Cavagnetto

3.2.3 Participants

In total, 21 participants were recruited to participate in the ARTIMATION validation sessions. Participants were recruited targeting two populations, "Expert" and "Student". The recruitment was made using 1) ENAC internal process, contacting personal, and students, and 2) using researcher network.

For the experts, 11 were recruited. The population was mostly composed of ENAC ATCO instructors (7), former ATCO now in research (2), and former ATCO now in ATCO formation (2).

For the students, 10 were recruited in the oldest formation available at ENAC, just before they left ENAC premises to their affected Control Center.

In total, 21 ATCO participated in the ARTIMATION validation sessions.

Table 5: Participants' background split by expert and student group

Participants	
Expert group (Professional ATCOs)	○ Total of 11 professional ATCOs ○ 3 female (27%) and 8 male (73%) ○ Mean age of 41 years (ranging between 34-51 years old) ○ Mean 15 years of working experience ○ Operational background: Only 2 participants did not have experience in En route ACC but both had Approach ACC and Tower ATC experience.
Student group	○ Total of 10 student ATCOs ○ 4 female (40%) and 6 male (60%) ○ Mean age of 22 years old (ranging between 20-26 years old)

EUROPEAN PARTNERSHIP

Co-funded by the European Union

3.2.4 Validation scenarios

All 10 validation Scenarios created for the validation procedure came from a continuous simulation scenario. This simulation scenario was created by the project using a 2016 traffic record, in a fictious sector created by ENAC for ATCO formation. The record was modified to create the conflict required for the validation. The conflict was designed to create different workload and difficulty of resolution, simplified in two categories "Easy" and "Hard" by adding aircrafts in the conflict—either following, or converging—or by modifying the contextual traffic. The assumption made around the difficulty was verified with 3 ATCOs, collecting their feelings about the difficulty of every conflict.

3.2.5 Measurements

The Human Performance Assessment methodology produced by SESAR 16.04.03/16.06.05 was used as reference for the Human Performance (HP) assessment process [35].

Qualitative data was collected through debriefings at the end of the session. Over-the-shoulder observation was performed by System Engineers. Quantitative data was collected through the questionnaires and neurophysiological measurements.

Observations

Observations allow to address topics related to Human Performance, with the purpose to provide detailed and reliable information on the way the activity is carried out by the user. Direct observation enables gathering a high amount of data, especially qualitative data which cannot be collected through other methods.

In the validation exercises, direct over-the-shoulder observation were used to collect insights about the ATCO's performance, including aspects related to experienced workload, situation awareness, usability, faced difficulties, etc.

Neurophysiological measurements [Unisap]

During each operational condition, the EEG and GSR neurophysiological measures have been collected, and the neurometrics have been calculated through offline analysis. They have been calculated four neurometrics, by following the methodology explained in the D4.1.

The **workload** neurometric, has been calculated as the ratio between the EEG theta activity over the frontal channels, and the alpha EEG activity calculated over the parietal sites.

$$WL = \frac{Theta(Frontal\ Channels)}{Alpha(Parietal\ Channels)}$$

The **(mental) stress** neurometric, has been calculated by using the high beta activity, over the left and rights parietal channels:

$$Stress = BetaHigh\ (P3, P4)$$

The **arousal (emotional stress)** neurometric, has been associated and calculated by using the Tonic component of the GSR signal. While the mental stress neurometric is most related to the cognitive (instantaneous) effect that stressful event may induce, the GSR-based metric, shows the effect that

EUROPEAN PARTNERSHIP
Co-funded by the European Union

the stress may induce in autonomic system (most related to emotional variations), that has been demonstrated to persist over time [11]

$$Arousal = Tonic\ (GSR)$$

Finally, the **approach-withdrawal** neurometric, related to the level of acceptance experienced by the user in front of a specific operational solution, has been calculated by the difference between the EEG alpha activity over the frontal rights sites, and the EEG alpha activity over the frontal left sites.

$$AW = Alpha\ (Frontal\ Dx) - Alpha\ (Frontal\ Sx)$$

Questionnaires

After each condition (i.e., conflict difficulty condition and visualisation condition), the participants were requested to fill in a questionnaire to provide their feedback on aspects related to the assessment like understanding, agreement with the solution, acceptability, cognitive workload, situation awareness, trust, usability and impact on work performance. These constructs were used as sub-dimensions to assess a self-report index of acceptance and human performance.

Debriefings

Debriefings, questionnaires, and over-the-shoulders observations are interconnected techniques. This means that on the one hand, data collected through observations and questionnaires were verified and discussed during the debriefings. On the other hand, insights extracted from the debriefings were used to guide the following observations. This combination of techniques can complement and reinforce the quality of the quantitative data collected and contributes to achieving more reliable results.

At the end of the validation's debriefings were used to gather further input from the participants on the different solutions and conditions and to complement participants' answers to the questionnaire.

3.3 Conflict resolution visualisation tools results

A total of 21 participants, including professional ATCOs and student ATCOs participated in the validation sessions. As mentioned above, the results presented below were derived from a combination of neurophysiological measurements, questionnaires (post-run and post-condition) and debriefings with the participants. The statistical analyses have been performed using Jamovi 2.2.5 [22][30].

3.3.1 Acceptance

3.3.1.1 Level of understanding

Understanding is defined as how much the provided explanation is clear and understandable by the ATCO.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

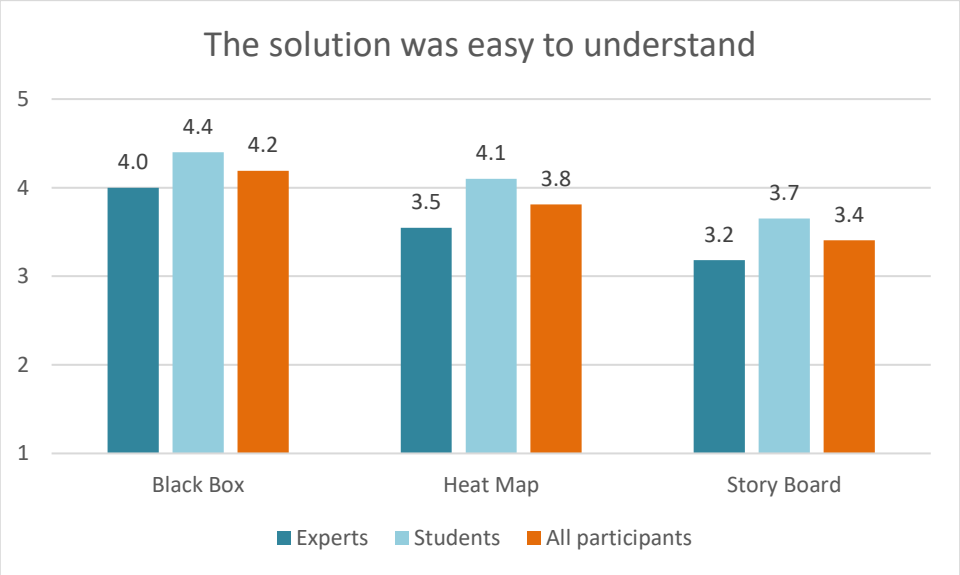


Figure 10: Post-run questionnaire 'Q: The solution was easy to understand' N=21.

In general, the expert ATCOs reported a lower level of understanding compared to students. The Black box (BB) (4.0 for experts and 4.4 for students) and the Heat Map (HM) (3.5 for experts and 4.1 for students) were considered easy to understand. The Story board (SB) condition had a mean rating closer to 3 for experts, which was considered neutral 'Neither agree nor disagree'. The level understanding ratings were higher for the BB condition, followed by HM and finally SB.

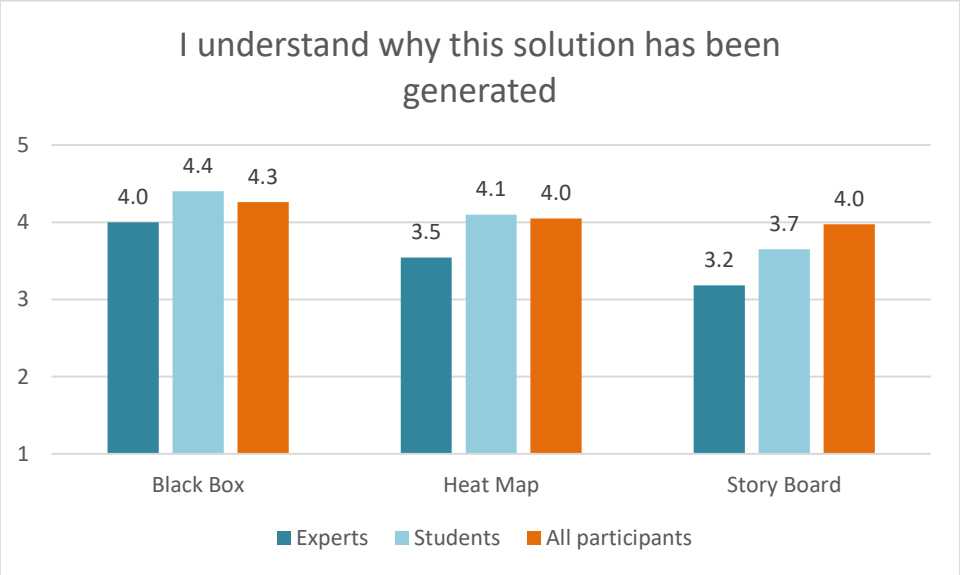


Figure 11: Post-run questionnaire 'Q: The solution was easy to understand' N=21.

In terms of understanding of how conflict resolution has been generated there were no significant ratings between conditions (visualisation modalities) for the experts, but the outcome is that they agreed to understanding how the solutions were generated (BB:4.0; HM:3.5;SB:3.2). Students

Co-funded by the European Union

reported a slightly higher understanding of the resolution generation in the BB condition (BB: 4.4; HM: 4.1; SB: 3.7).

A significant positive correlation between the two items of understanding has been found. Therefore, the two items have been aggregated to ease analyses.
To assess the differences between the three levels of visual explainability, an analysis of variance (ANOVA) test has been conducted for the understanding variable. After a post-hoc comparison between the three explainability levels, the Black Box (BB) condition resulted in a more understood AI outcome by all the sample than the Storyboard (SB) condition ($p = 0.030$). Moreover, the post-hoc comparison between the expertise level of the participants showed that there is a significant difference in understanding between students and experts, resulting in the student group having a higher understanding of the solution ($p = 0.018$). No other significative differences were found.

3.3.1.2 Level of agreement

Agreement is considered the state for which a participant agrees with a specific solution provided by the AI.

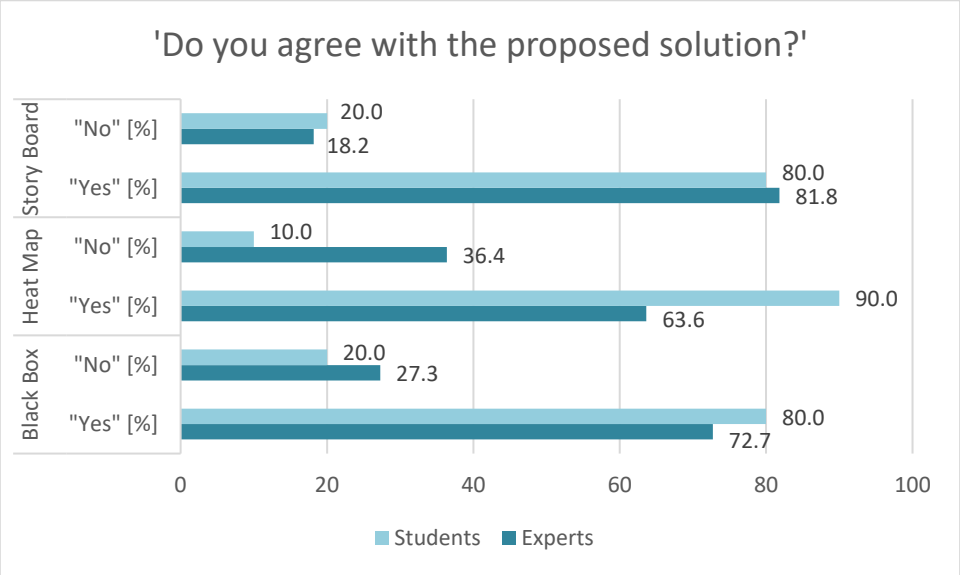


Figure 12: Post-run questionnaire 'Q: Do you agree with the solution?' N=21.

In general, experts were accepting/agreeing with the proposed AI resolution less frequently. We can see this particularly in the Heat Map (HM) condition where students reported a clearly higher level or agreement compared with experts (Students = 90% of agreement rate vs experts = 63.6% agreement rate). These results might be explained by the fact that students were less critical of the HM visualization modality and some of them even mentioned that it was the resolution visualization condition they preferred because it was visually appealing.

To justify the general lower acceptance rate from experts' side we can think about claims during the debriefings that experts mentioned they were **reluctant to accept a solution that is not their own** simply because they might lose more time trying to understand what is being proposed by the tool, ultimately, running the risk of finding themselves **'out of the loop'**.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Another aspect that might have impacted is that experienced ATCOs usually apply strategies that are not considered in the algorithm generating the solutions. For instance, some mentioned they preferred to intervene in more than one conflict not to penalize too much only one flight. Also, they were not sure which parameters the ML algorithm was considering in order to generate the visual conflict resolution proposal, but the focus of the experiment. The solutions generated by the Genetic algorithm used were often considering intervening in less aircraft as possible.

In some cases, participants mentioned that just did not manage to have time to analyse and integrate what the AI solution was proposing.

3.3.1.3 Level of Acceptability

Acceptability can be defined as the intention to accept a new technology, meaning that people must perceive usefulness and intuitive usability in the technology other than having favourable attitudes to adopt it. The individual's feelings, favourable or unfavourable, about aspects of the environment or objects related to the environment.

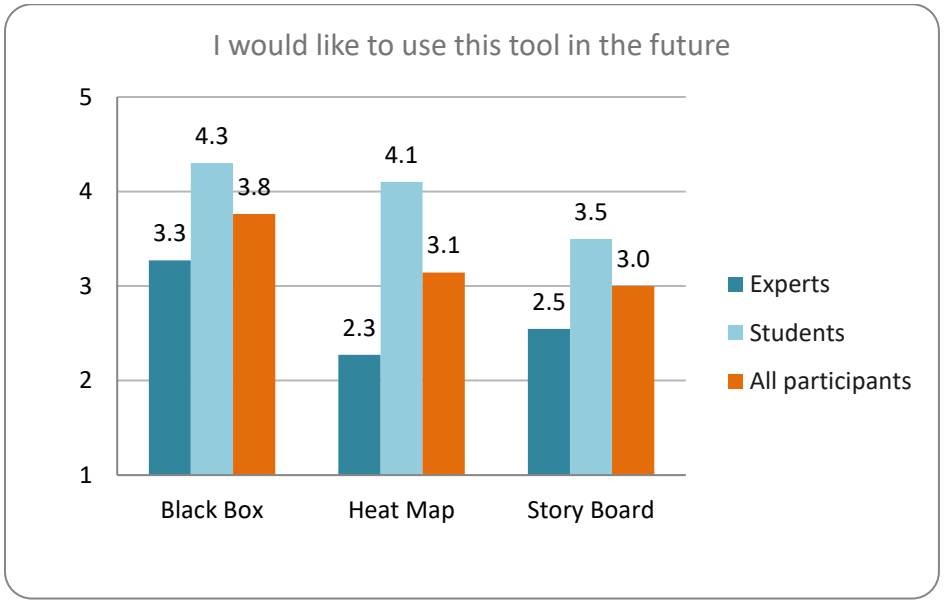


Figure 13: Post-condition questionnaire 'Q: I would like to use this tool in the future' N=21.

When asked if they would like to use this tool in the future once again the opinions differed between experts and students. Students agreed that they would like to use the BB and the HM conflict resolution visualisation tools even if need for improvements were reported during the debriefing. During the debriefings, 6/10 students reported their preference for the BB condition, 3/10 the HM and 1/10 the SB preference.

Experts had a neutral answer for the BB condition, and they disagreed that they would like to use the HM and SB tools. During the debriefings all experts expressed their preference towards the BB condition.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

A significant correlation between the 2 acceptability items ("I would like to use this tool in the future", "I like the new decision support interface") has been found. Therefore, the two acceptability items have been merged to facilitate analyses and data interpretation.
For the acceptability items, a post-hoc comparison between the three conditions has been conducted. The BB condition resulted being significantly more acceptable than both the HM ($p = 0.033$) and the SB ($p < 0.001$). No significative differences in the interaction between the condition and the complexity of the scenarios has been found.
To assess the difference in the acceptability items between the expertise and between conditions, an ANOVA test has been conducted. After a post-hoc comparison, a significant effect of the expertise on the acceptability of the visual explanation has been found: the students found the interfaces globally more acceptable than the experts ($p < 0.001$).
A post-hoc comparison assessing the interaction between the expertise level and the condition has been conducted. All the explainability conditions resulted significantly more acceptable for the students than for the experts. The black box was significantly more acceptable for students than for the experts ($p = 0.020$), as for the heat map condition ($p < 0.001$) and for the Storyboard as well ($p = 0.001$).
Between the experts, the black box condition resulted being significantly more acceptable than both the heat map ($p = 0.021$) and the storyboard ($p = 0.002$). In the expert group, no significative differences between the heat map and the storyboard condition have been found.
Between the students, no significant differences between conditions have been found.

Neurophysiological results

A repeated measures ANOVA (CI=0.95) has been performed, by considering the two factors (i.e., conditions *[Black Box; Heat Map; Storyboard]* and repetitions *[1st and 2nd]*). The statistics has been performed for each experimental group (i.e., students and experts), to highlight any different.

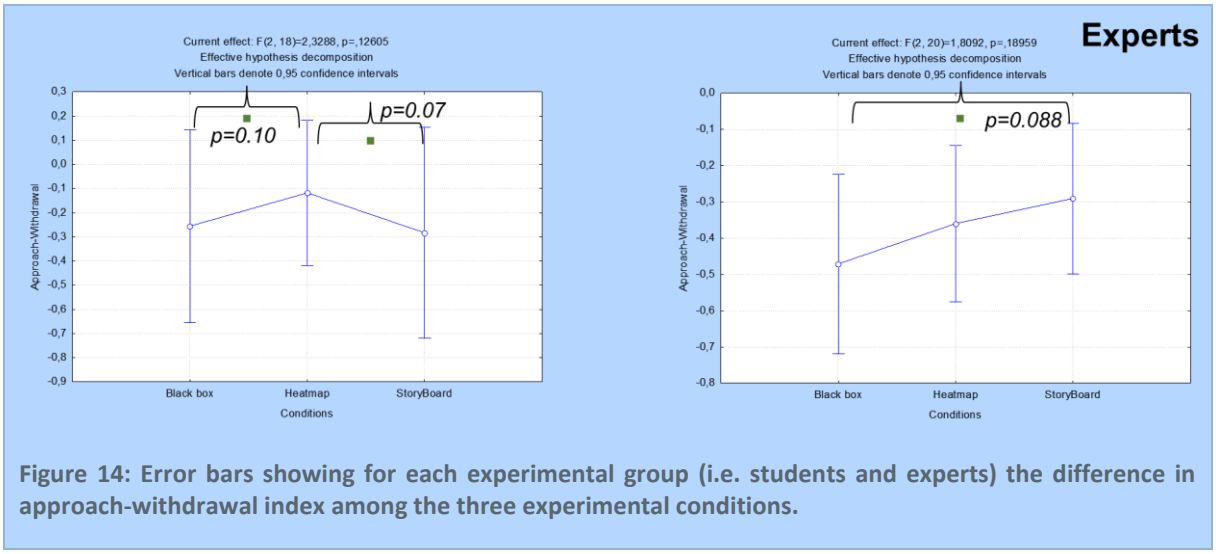


Figure 14: Error bars showing for each experimental group (i.e. students and experts) the difference in approach-withdrawal index among the three experimental conditions.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

The results showed a different behaviour between students and experts. In particular, the students exhibited the highest approach-withdrawal on the HM solution, with respect to the other two conditions (higher than BB solution, p=0.1; higher than SB, p=0.07). Otherwise, experts experienced a higher approach-withdrawal in correspondence of the SB solution, that was higher (trend, p=0.088) then the BB condition.

In detail, students exhibited the highest approach-withdrawal (acceptability) on the HM solution. From the debriefings it was clear that student's acceptability from the HM solution was higher when compared to experts. They mentioned that it was visually appealing, interesting and the use of colour was appreciated. Another thing that was appreciated was that it gives them more flexibility and does not point them towards a single resolution, they can analyse and come up with their own solution, but this could become critical in terms of an overall high workload scenario.

The task that participants were performing in this experiment focused on the resolution of single conflicts, but it does not correspond to the overall role of an En route ATCO, so the results also might have been affected by the fact that it was focused on a single task and therefore the analysis should take this inconsideration.

The task that participants were performing in this experiment focused on the resolution of single conflicts, but it does not correspond to the overall role of an En route ATCO, so the results also might have been affected by the fact that it was focused on a single task and therefore the analysis should take this inconsideration.

Experts experienced a higher approach-withdrawal (acceptability) towards the SB solution, which was higher (trend) then the black-box. Here the differences between conditions were not significant but they are just a trend.

Both in questionnaires ratings an in the debriefings experts mentioned their preference towards the BB solution, they were that it was more straight forward, easy to understand and mainly it allowed them to make their decision in less time compared to the heat map (HM) or the storyboard (SB) solution.

One possible explanation for this discrepancy between results of experts could be related to the intrinsic bias induced by the BB condition, especially on experts, that are of course experienced and used to face with this kind of solutions, with respect to the other two conditions, that, despite the training, were still new. The approach-withdrawal index is able to catch intrinsic (and instantaneous) reactions coming from the user brain, that are not by definition biased the experience, or by the long thinking regarding the possible operational use of this solution (that is instead measured by questionnaire post experiment). In other words, the instinct and instantaneous reaction, suggest that the storyboard (and on average also the HM, but not significantly) could potentially be well accepted by the operators, even more with respect to the BB, but the long thinking of operators, suggests instead a possible lack in effectiveness.

During the debriefings, when asked about their preference 11/11 ATCOs reported that they preferred the Black box (BB) solution, even if one of them also liked the concept of the Heat map (HM).

The main reasons for the BB preference were that it was more straight forward, easy to understand and mainly it allowed them to make their decision in less time compared to the heat map (HM) or the storyboard (SB) solution.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

The students that preferred the HM mentioned the fact that it was visually appealing, interesting and the use of colour was appreciated. Participants that preferred also mentioned that it gives them more flexibility, because they can analyse and come up with their own solution. On the downside it takes more time to analyse in more complex conflicts or conflict with aircraft and that makes it less suitable in situations in which the ATCO would need to make a fast decision.

3.3.2 Human Performance (VO_CDR_2)

3.3.2.1 Mental Workload

The workload index showed a lower significant decrease during the HM condition, with respect to the other two conditions. This behaviour was highlighted in particular during the 1st repetition (i.e., 2 aircrafts crossing).

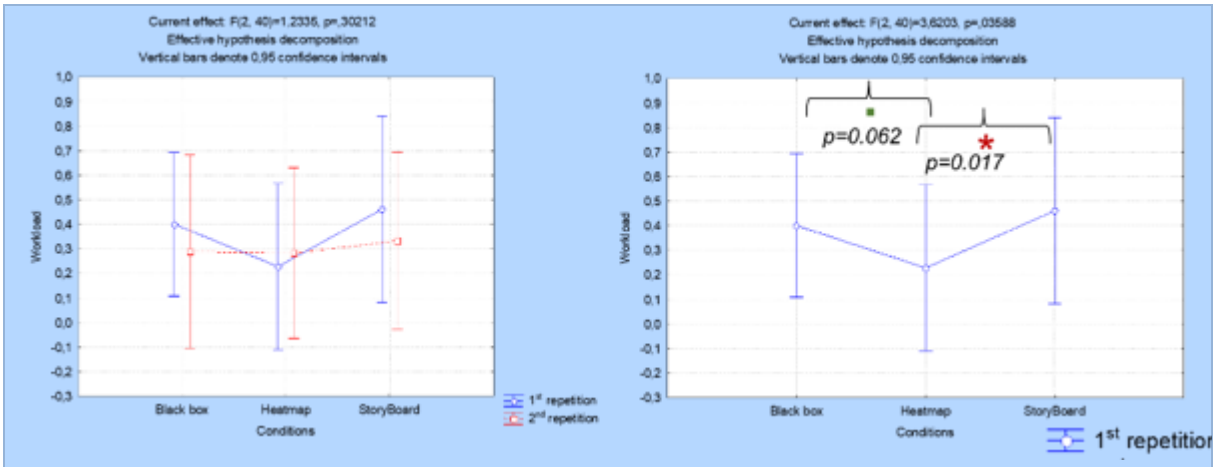


Figure 15: Error bars showing for each experimental group (students and experts) the neurophysiological workload index among the three experimental conditions, in both repetitions.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

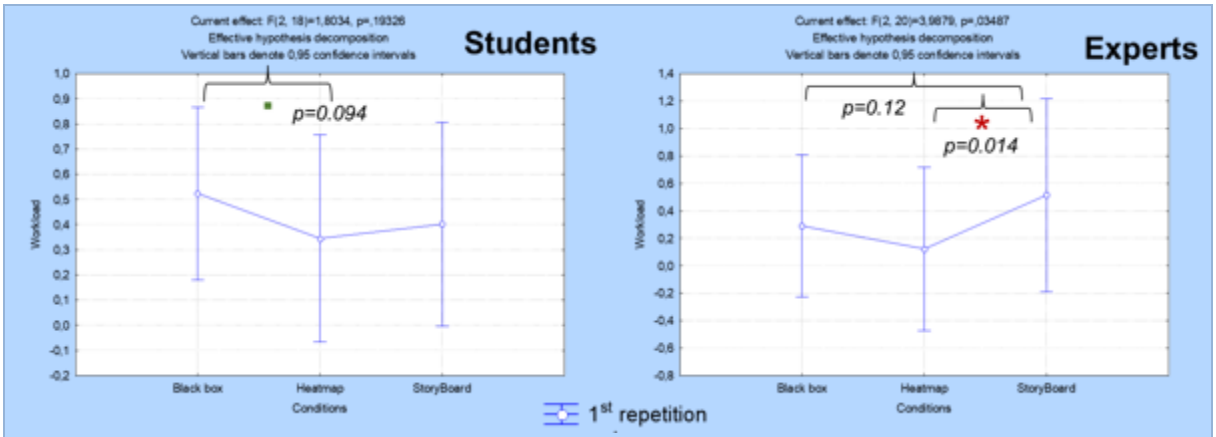


Figure 16: Error bars showing for each experimental group (students and experts) the neurophysiological workload index among the three experimental conditions.

For both the students and the experts' groups, the HM exhibited the lowest value of workload on average.

The students experienced the highest level of workload during the BB condition (p=0.094).

The experts experienced the highest level of workload during the Storyboard condition (p=0.014 with respect to HM), and the BB exhibited higher workload with respect to the HM solution.

3.3.2.2 Stress

3.3.2.2.1 Mental stress

Co-funded by the European Union

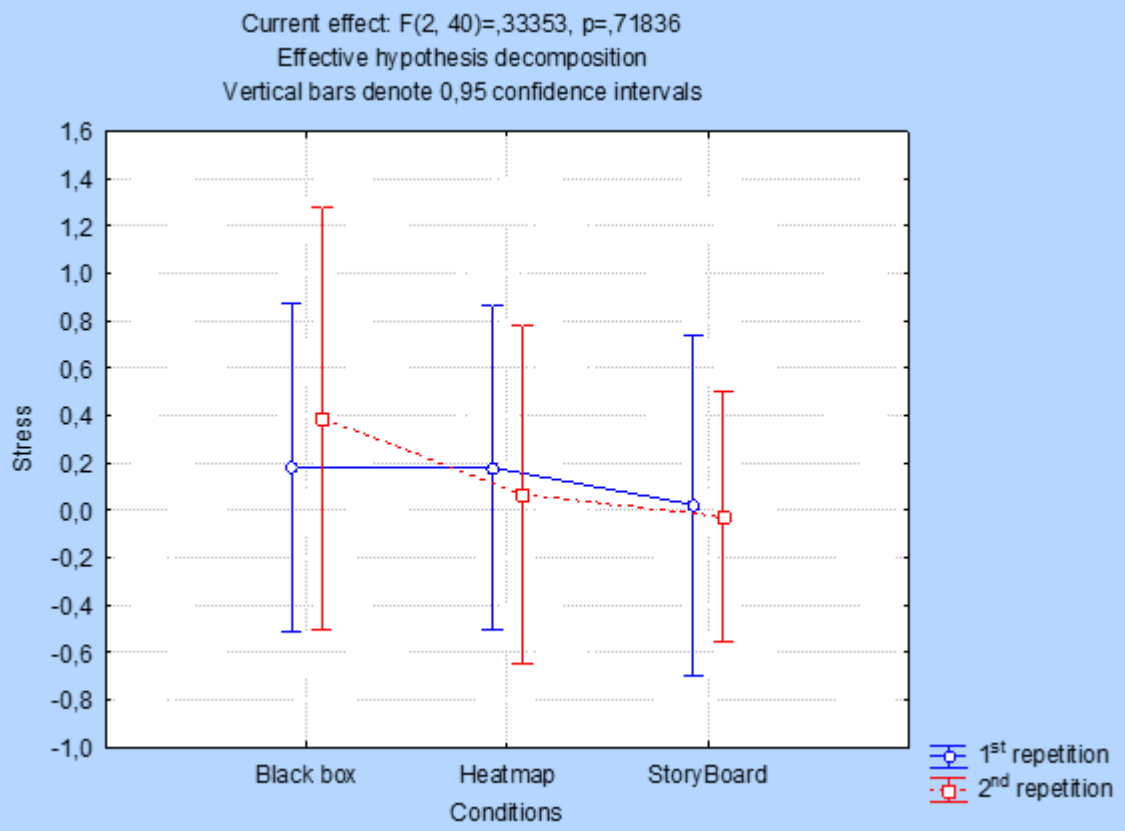


Figure 17: Error bars showing for each repetition the neurophysiological cognitive stress index among the three experimental conditions.

The stress index, apparently, did not show any significant trend among the conditions and between repetitions.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

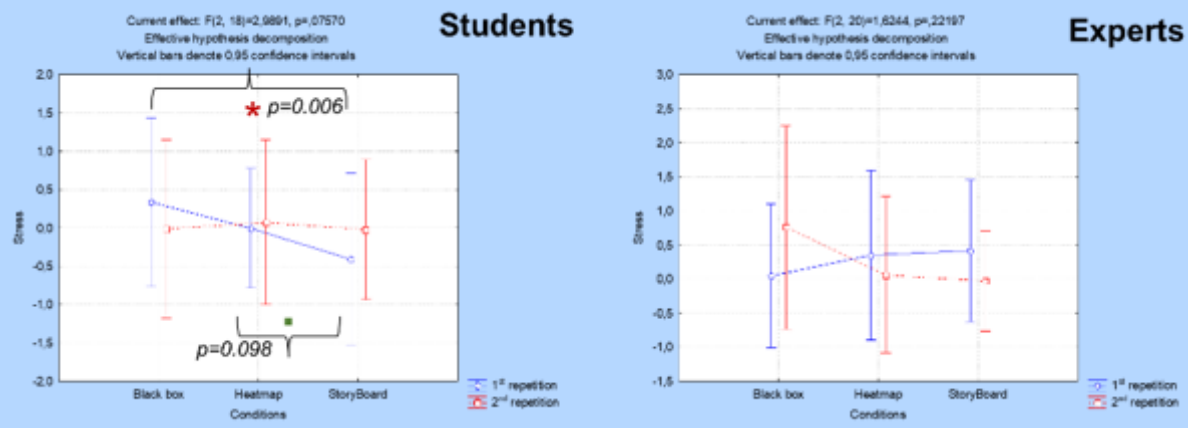


Figure 18: Error bars showing for each experimental group (students and experts) the neurophysiological cognitive stress index among the three experimental conditions.

Students exhibited a significant decreasing in stress experienced during storyboard with respect to the other two conditions, during the first repetition.

On the contrary, Experts did not show any significant trend among the conditions

3.3.2.2.2 Emotional stress (arousal)

EUROPEAN PARTNERSHIP

Co-funded by the European Union

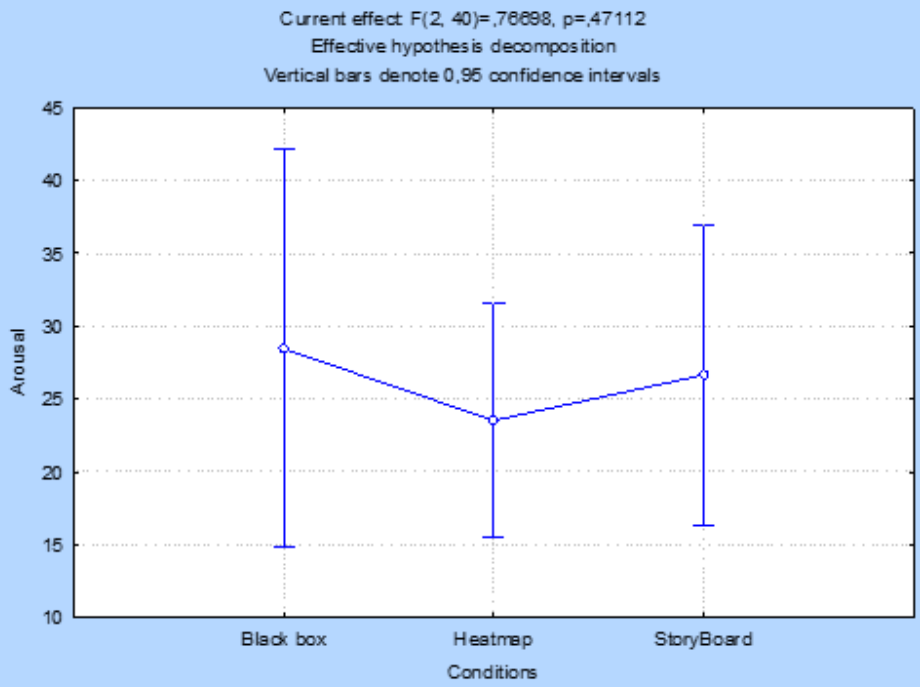


Figure 19: Error bars showing for each repetition the arousal index among the three experimental conditions.

In terms of arousal, no significant difference has been showed among the conditions.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

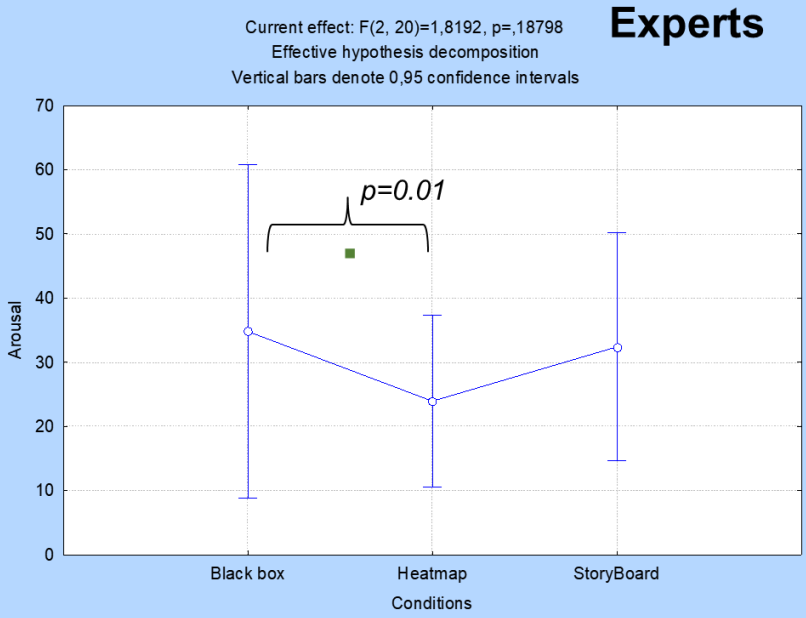


Figure 20: Error bars showing for each repetition the arousal index among the three experimental conditions, for the experts.

Experts exhibited an arousal significantly higher during the black box condition, with respect to the HM condition.

3.3.2.3 Situation awareness

Situational Awareness is the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and a projection of their status soon.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

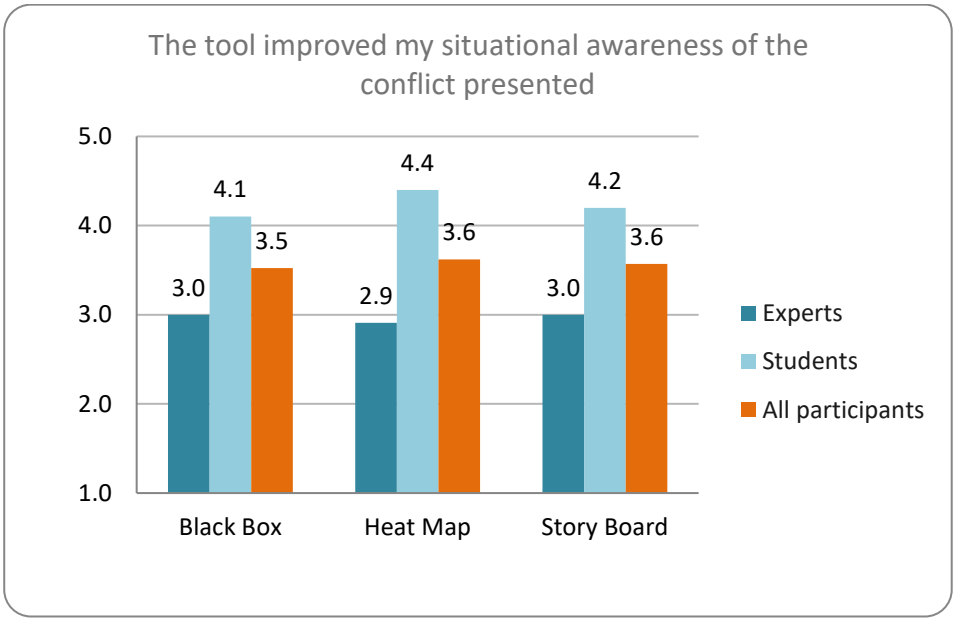


Figure 21: Post-condition questionnaire 'Q: The tool improved my Situation awareness of the conflict presented.' N=21.

For experts no improvements in terms of Situational Awareness (SA) between the different conditions were reported (BB: 3.0; HM: 2.9; SB: 3.0). These results might have been impacted by the experimental limitations and lack of realism in the task, since they had to solve the conflict presented without access to the radar screen situation, they had to remember. Some ATCOs reported that this impacted their SA and possibly their choices and ratings.

Students on the other hand had a more positive attitude by reporting that they agree that the tools improved their SA (BB: 4.0; MP: 4.4; SB: 4.2). Indeed, the tool is providing them with a resolution that they would have to come up by themselves.

A significant difference between students and experts has been found on the Situational Awareness items, resulting in students having an improvement of situational awareness significantly higher than the experts ($p < 0.001$).

For the Heat Map condition, students resulted having a significantly higher improvement in situational awareness than experts ($p = 0.023$).

3.3.2.4 Usability

Usability is a quality attribute that assesses how easy user interfaces are to use. It describes the level of ease with which a system allows a user to get to that goal.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

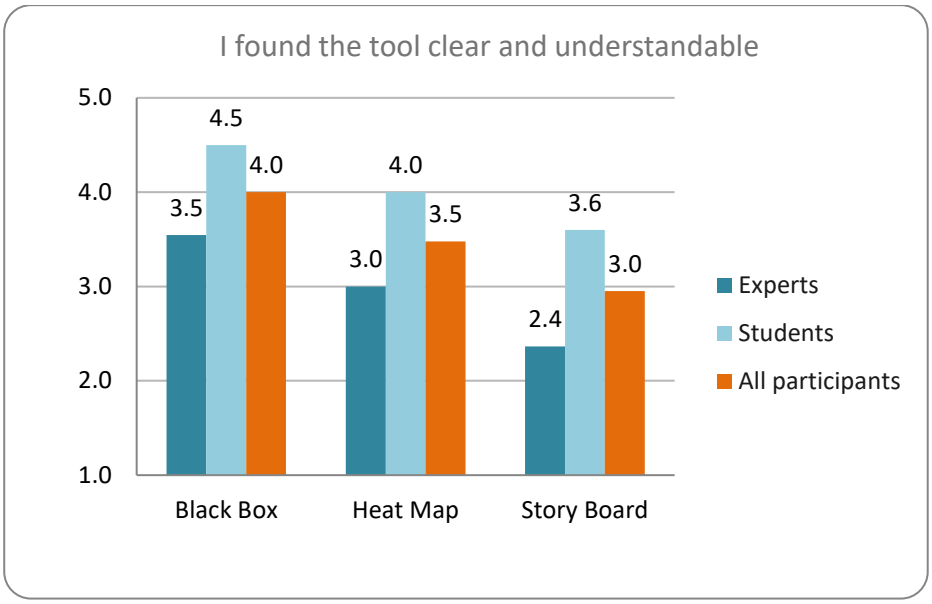


Figure 22: Post-condition questionnaire 'Q: I found the tool clear and understandable.' N=21.

Experts tended to agree that the BB condition was clear and understandable (BB: 3.5). The HM and SB had slightly lower results. The HM solution was considered not to have an effect (HM: 3.0) while the SB was considered to have a negative impact on SA (SB: 2.4).

Students found all the conditions clear and understandable but the BB and HM with higher scores (BB: 4.5; HM: 4; SB: 3.6).

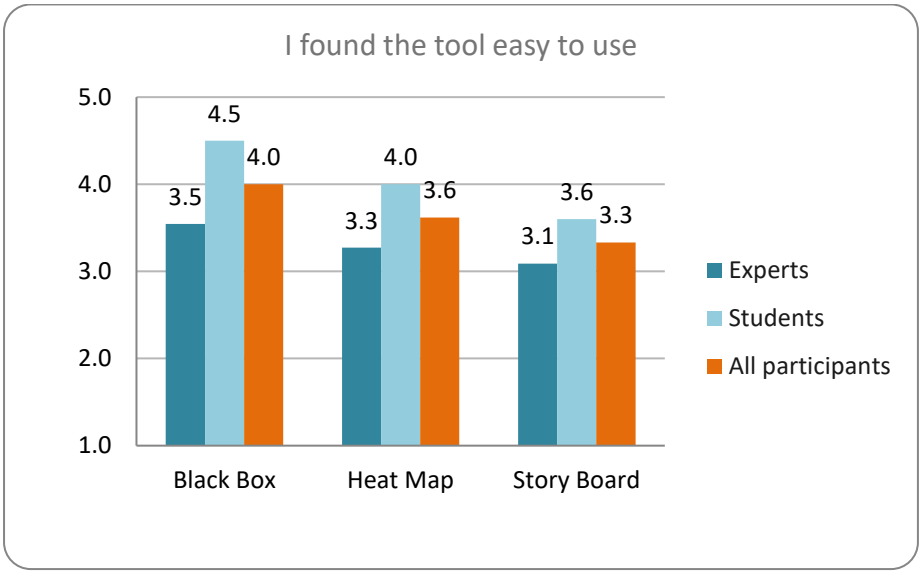


Figure 23: Post-condition questionnaire 'Q: I found the tool easy to use.' N=21.

Experts tended to agree that the BB solution was easy to use (BB: 3.5). The HM and SB had slightly lower results- neutral (HM: 3.3; SB: 3.2).

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Students found all the solutions would be easy to use but the BB (4.5 –strongly agree) with slightly more positive results (HM: 4; SB: 3.6).

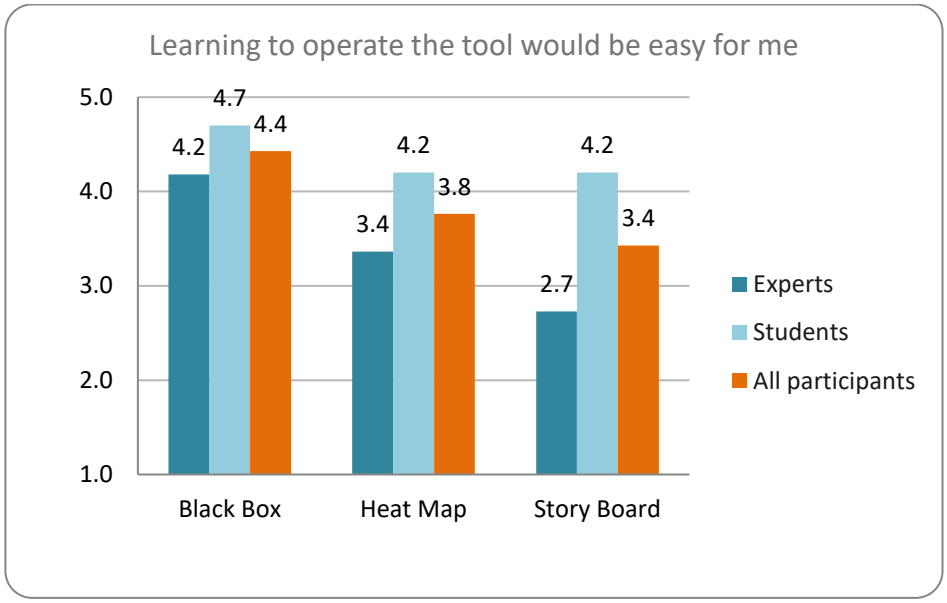


Figure 24: Post-condition questionnaire 'Q: Learning to operate the tool would be easy for me.' N=21.

Experts considered that the BB solution would be easier to learn, followed by the HM. This did not apply to the SB solution.

Students mentioned that all solutions would be easy to learn to operate with a slightly more positive tendency to the BB.

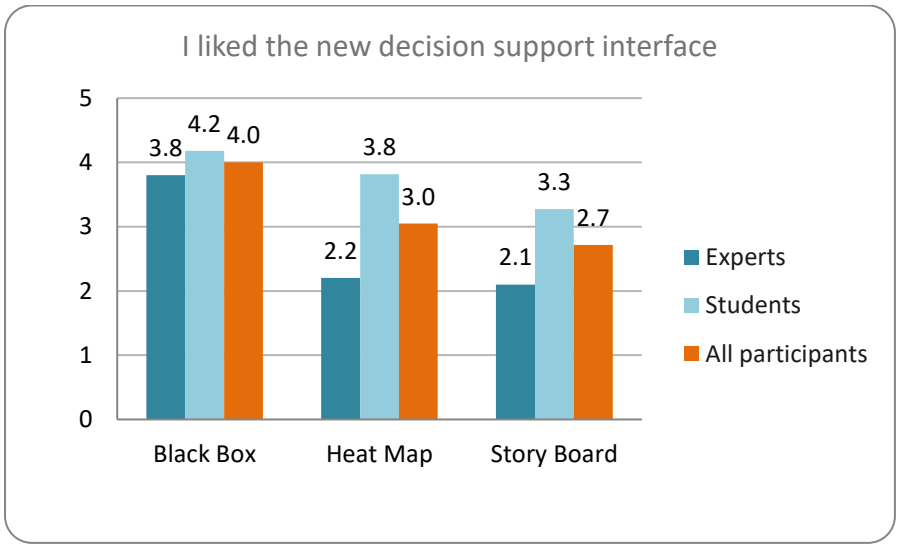


Figure 25: Post-condition questionnaire 'Q: I liked the new decision support interface.' N=21.

Co-funded by the European Union

Experts mentioned that only the BB solution was suitable in terms of decision support interface. The opinion for HM and SB decision support was negative (HM: 2.2; SB: 2.1).

Students mentioned they like the decision support interface of BB (4.2) and HM (3.8) conditions with a preference towards the BB. In general, the SB solution was received with more reservations. ATCOs thought that the SB solution implementation would call for the presentation on a secondary screen.

The BB solution configured as the less disruptive from the solutions since it followed a similar approach to most implemented en-route tools and elements that the ATCOs are familiar with. Most mentioned that it was clear, logical and did not clutter other information on the screen (did not conceal other information). They also mentioned that it was useful to know the turning point of aircraft in the trajectory. The fact that it provided a single resolution solution was appreciated in conflicts that are more complex, also if they involve more than two aircraft.

Students were the ones that highlighted the following benefits of the HM solution. They pointed out that the visual component as an advantage and that they could see which trajectories would be conflicting or not. This visualisation allows for room for other operational factors that might not be computed by the algorithm like bad weather (turbulence). ATCOs can easily trace the zones where the plane can pass in advance. One ATCO mentioned the projection of the envelope was easier to remember than just a few degrees that the BB solution provides, but the level experience probably plays influences this opinion.

An ANOVA has been conducted for the items of usability. After a post-hoc comparison of the results of the questionnaires for the usability items, in the whole sample, the black box resulted significantly more usable than the storyboard condition ($p = 0.049$). At the same time, globally, the students reported a significantly higher usability of all the tools than the experts ($p < 0.001$). Between the experts' group, the black box condition resulted being significantly more usable than the storyboard ($p < 0.001$). No other significant differences have been found.

Black box HMI suggestions for improvement:

- The solution is limited to the horizontal plane (resolution with headings). In TMA sectors this tool will not be useful since the ATCO works more in terms of flight levels. The context and shape of the sector can also influence the need to work more with flight levels.
- Participants mentioned they would like to be provided with a quick way to know which aircraft or routes are completely conflict free and he should not change the trajectory. That might save them some time.
- Instead of numbering one ATCO mentioned he would have liked to see just the proposed route (the new track).
- The trajectory of aircraft should be highlighted.
- The heading angle (in degrees) would be an information that was mentioned as nice to have.
- One participant mentioned that he found the distances hard to read.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

- The lines between speed vectors, trajectories, were considered disturbing by one of the participants (student). Another participant proposed that suggested trajectories should be presented in dotted lines to avoid confusion (with speed vectors).
- One participant suggested that colour on the present solution should be used to differentiate the history of the conflicting pair of aircraft following each other or time-related indications.
- One of the student participants the mentioned the minimum distance or time was secondary information and did not need to be shown.

Heat map HMI suggestions for improvement:

- The solution envelopes take a lot of radar space, it cluttered the screen and might mask other important information.
- Having too much information on the screen does not help in making a quick choice. Many participants mentioned this solution takes too much time to analyse.
- Two ATCOs mentioned they would have liked to see only the green zone/envelope and would avoid further clutter and masking other information. One participant mentioned that it could be more like a radar in a plane: ‘ It's a bit like in a radar, in a plane, to have all the colours of the storm with the green, orange red rather than just the place where you absolutely must not pass.’
- Participants mentioned conflicting pair of aircraft should be easier to distinguish. The colour code could associate and highlight the pair of aircraft in conflict.
- The colour scheme could be improved to contrast more the areas where the planes should not penetrate. When two heatmaps overlap can become unreadable and it's too conspicuous for the radar. One participant mentioned the more colours, the more time trying to analyse it rather than seeing a conflict that's being created at the other end of the sector that you haven't seen.
- In terms of colour scheme, the red colour calls immediate action, so it should not be used for this NM that are considered in this case.
- One participant mentioned the heading was missing.

Storyboard HMI suggestions:

- Too much information and considered difficult to visualise (small size of the conflict frames).
- Impossible to have this visualisation for multiple conflicts detected in the same timeframe.
- This type of visualisation should be presented in a secondary screen, but it is difficult not ideal take the ATCO attention away from the primary screen.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

- Too much time needed to analyse the whole evolution of the conflict, it could introduce safety problems of an executive ATCO working in En route.

In general participants mentioned that they would prefer to have access to the tools on demand because they were considered more useful in complex scenarios or scenario in which they would be experiencing a high level of workload.

3.3.2.5 Trust

Trust is a cognitive state that usually influence the actual, behavioural dependence on automation. The operator's use of automation is related to his or her momentary trust, which in turn is related to the type and frequency of faults and operators' confidence in their own ability.

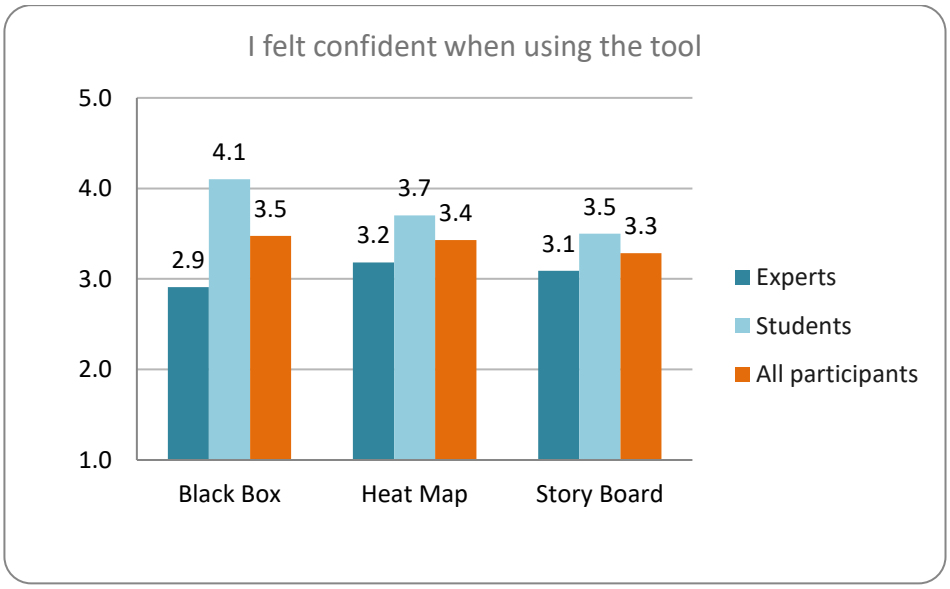


Figure 26: Post-condition questionnaire 'Q: I felt confident when using the tool.' N=21.

Experts reportedly were not confident when using any of the solutions, but this might have been impacted by the fact that they had limited training and explanation. Some ATCOs, especially professional ATCOs, mentioned that they felt they would need more information and training on how the Machine Learning (ML) algorithm to trust it.

To assess the differences between the condition and the expertise level of the sample for the trust items, an ANOVA has been conducted. After a post-hoc comparison, a significant difference between students and experts has been found students had a significantly higher trust in the presented resolution advisory than the experts ($p = 0.009$). No significant differences between conditions have been observed.

ATCOs mentioned is that trust in the solutions is a requirement to use them in operations. That trust must be acquired before or after operational usage, either in training, with briefing or even during

EUROPEAN PARTNERSHIP
Co-funded by the European Union

debriefings. Therefore, we can say that explainability might be more relevant for applications for those purposes.

3.3.2.6 Task performance

Task performance can be defined as the effectiveness with which job incumbents carry out activities that contribute to the organization's "technical core" either directly by executing a part of its technical process or indirectly by providing it with needed materials or services.

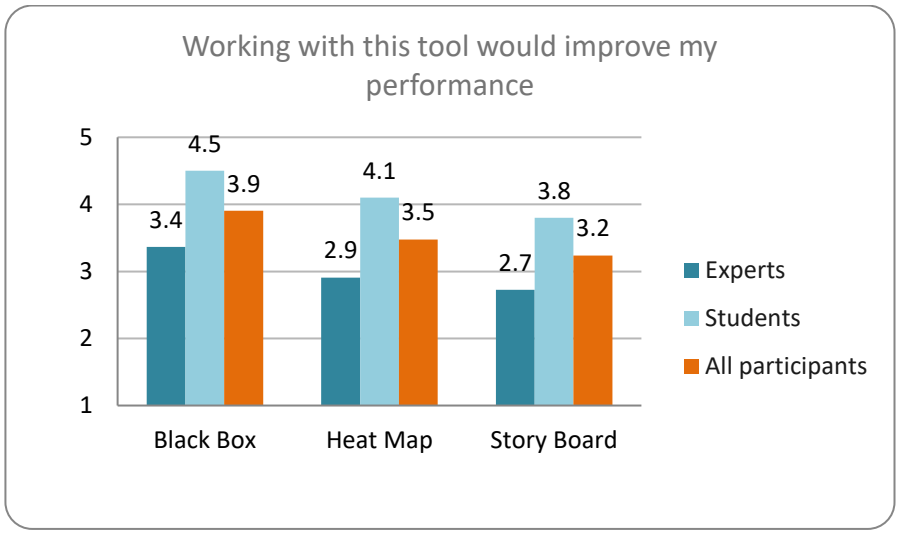


Figure 27: Post-condition questionnaire result 'Q: Working with this tool would improve my performance.' (5-point Likert on level of agreement) N=21.

Almost half of the experts (46%) considered that the BB solution could improve their performance, while the other solutions could have a negative impact. Students had a more positive judgement considering all the solutions would improve their performance, particularly the BB tool.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

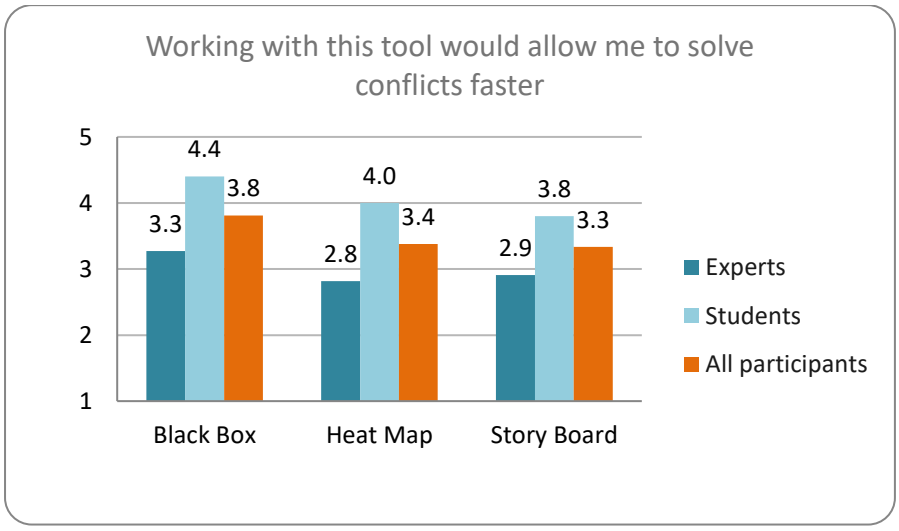


Figure 28: Post-condition questionnaire result 'Q: Working with this tool would allow me to solve conflicts faster.' (5-point Likert on level of agreement) N=21.

Experts did not agree that neither of the conflict resolution visualisation solutions would support them in solving their conflicts in a faster way. On the other hand, students were more positive since they considered all the solutions would support them in solve conflicts in a faster way, particularly the BB tool. During debriefings ATCO students mentioned how the BB solution possible advantages solving conflicts in a faster way.

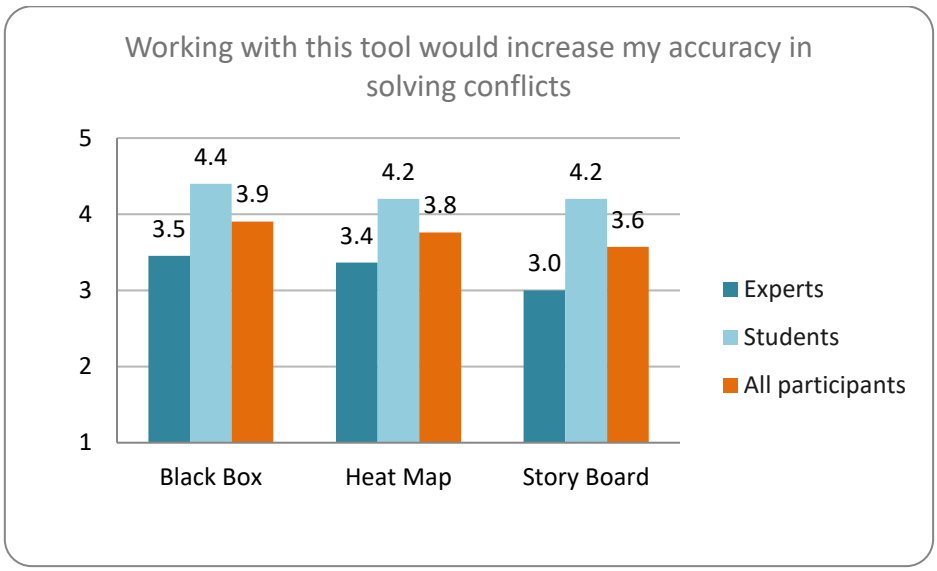


Figure 29: Post-condition questionnaire result 'Q: Working with this tool would increase my accuracy in solving conflicts.' (5-point Likert on level of agreement) N=21.

The results reveal that few experts considered that the BB and the HM solutions could potentially support them in solving conflicts with a higher accuracy. Once again, students had a more positive

EUROPEAN PARTNERSHIP

Co-funded by the European Union

opinion and considered that all the solutions, in general, would increase their accuracy in conflict solving with no significant.

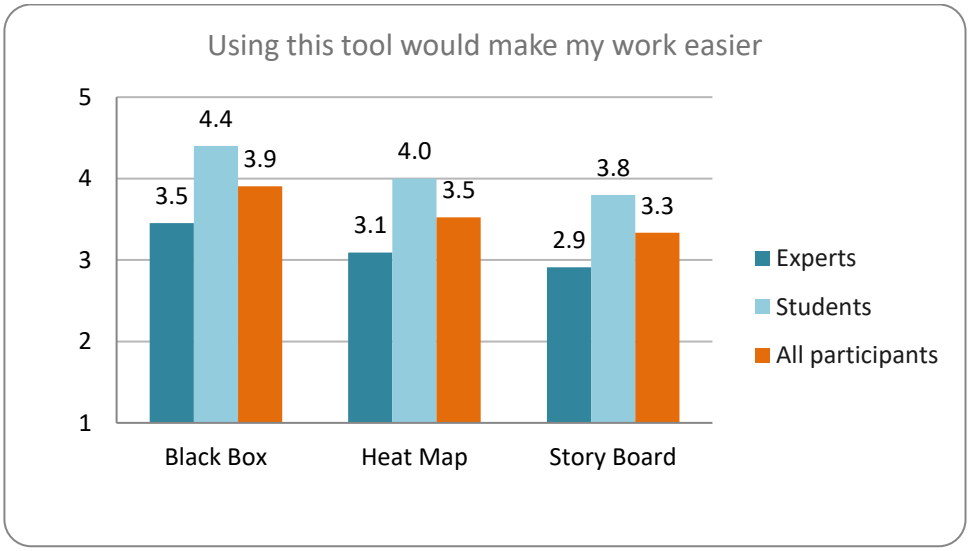


Figure 30: Post-condition questionnaire result 'Q: Using this tool would make my work easier.' (5-point Likert on level of agreement) N=21.

Similarly, to previous results, some fewer experts considered that the BB could make their work easier compared to student. Students had a more positive opinion and considered that generally all the solutions would increase their accuracy in conflict solving with no significant, with a preference towards BB solution.

Finally, the 4 items composing the work performance index has been merged after a positive significant correlation between the items has been found.

The post-hoc comparison done after the ANOVA shows how students reported a higher improved perceived work performance independently on the conditions ($p < 0.001$).

Between the two groups, students reported a significantly more improved work performance than the experts ($p = 0.032$). The same result has been found for the heat map condition ($p = 0.048$) and the storyboard condition ($p = 0.046$).

3.3.3 Correlation between acceptance and human performance (VO_CDR_3)

A correlation matrix to understand if the items of the sub-constructs measuring the human performance and the acceptance was performed between all the sample. A significant correlation between the items has been found, therefore, the items were merged to assess the correlation between the acceptance and the human performance. After performing a correlation matrix between the acceptance and the human performance, a significant correlation has been found: the human performance while interacting with the XAI tools is correlated to the acceptance of the solutions provided by the Artificial Intelligence ($p < 0.001$).

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Therefore, we tried to assess the correlation between the acceptance and the human performance splitting the sample in experts and students. In both the experimental groups a significant correlation ($p < 0.001$ in both cases) between the acceptance and the human performance has been found.

3.3.4 System performance (VO_CDR_4)

3.3.4.1 Safety

ATCOs mentioned HM solution current design can impact safety by cluttering and masking important information on the radar. The SB concept in terms of implementation was related to the fact that the amount of information is not calibrated for the type of En route task, so the time the information takes to be analysed could create cognitive tunnelling situations.

More generally in terms of XAI applications for conflict detection and resolution tools, there could be a higher risk for ATCOS to implement suggestions without checking what has been (once the ATCO trust the tools). On the long run relying on the tool could lowering ATCO the vigilance and loose skills overtime. But of course, these tool implementations would have to be followed by new training requirements to mitigate the negative effect on performance that were just mentioned.

Most of the ATCOs mentioned that they find the solutions proposed by the system were good but if they are not matching their solution, it forces the ATCO to think twice or ultimately doubt his own solution. One ATCO complemented that he would be reluctant to accept a solution that is not his own simply because he might find himself in a situation that he does not feel that he can rapidly recover, because at that point he might be 'out of the loop'.

The participants that mentioned more frequently that the proposed solution was making them doubt their own solutions were students when the solutions was not matching their own. This could point us in the direction of the importance of the solutions conformance to the ATCOs strategies and the impact on their acceptance and ultimately, safety.

The fact that the proposed solution is not matching the solution of the ATCO could ultimately make him loose more time analysing or worse, make them fall behind what is going on in their sector.

3.3.4.2 Expected Impact on ATM system

AI support and types of conflicts

Most participants felt that the AI solutions proposed were not useful for conflicts with two aircraft. They thought that the BB solution could be useful in conflicts involving three or more conflicts.

On the other hand, the visualisation conditions with more 'explainability' embedded, HM and SB correspondingly, based on most debriefings were considered less useful for more complex operational scenarios, in short, less operationally acceptable.

ATCOs felt that in more complex scenarios or when they are experiencing more workload, they could be more willing to accept the solutions proposed by the tool.

In general, students and experts faced the AI decision support in conflict resolution in two different ways based on their feedback in debriefings. The experts seem to tend to consistently compare the AI

EUROPEAN PARTNERSHIP

Co-funded by the European Union

solution with their own, assuming their solution is the best to be surpassed by the AI proposal. On the other hand, students put on the same level both their own solution and the AI proposal, being more open to accept a proposal that they did not come up with (higher level of trust).

XAI application in ATM

Most ATCOs mentioned that if they would need more time to analyse and double check the proposals from the solution with explainable AI solutions they tested and that could ultimately translate in an increased workload during operations and/or possible loss of situational awareness due cognitive tunnelling while using the tools. This seem to point out that higher explainability could be more useful for less timely critical or tasks or operational phases in which the ATCOs are subject to lower risk of cognitive workload, like planning tasks.

Training

Some ATCOs mentioned that it would be interesting to explore and better understand the advantages of the AI solutions for training. The focus could be on understanding how experts (maybe with different approaches or goals) would solve or work in certain scenarios. To make them visualize trajectories and different approaches based on different parameters could be very useful is to discuss and debrief.

During the debriefings and final discussion there was not a univocal opinion on which solutions that would be preferred for training uses. Each ATCO seemed to have their own preference. What emerged was that all solutions and AI in general were perceived as having potential for training.

The higher the visual XAI the better for training, because they elicit better the reasons behind the proposed conflict detection and on the resolution itself. Participants even mentioned that they could see the solutions with higher XAI visualisation to have potential to be used during trainings with AI tools and once the trainees see how the ML algorithm works and to build trust, with the support of these solutions, they can start using the tools with less visual XAI for the actual operations.

Two ATCOs alerted to the fact that using AI tools to learn in conflict solving scenarios too early in the training process could have drawbacks, since ATCOs could end up mimicking the AI tools work strategy before developing their own.

Personalization of ML algorithms and ATCO strategies

During the debriefings some ATCOs voiced their interest in the use of applications of AI and ML algorithms to learn strategies from them.

ATCO 18- 'I think it would be interesting if, when you arrive at the position, you have a sort of profile of how you control it and the AI adapts to each person, which would be amazing, but I guess it would be something for the future.'

Kirwan & Flynn [23] studied controllers across seven nations, and found general agreement across controllers on the factors, rules, and principles they used to devise en route resolution strategies. Strategies were defined by four main dimensions:

- Formal rules—such as Letters of Agreement, or the semi-circular rule;

EUROPEAN PARTNERSHIP

Co-funded by the European Union

- Principles—such as 'minimise number of aircraft to move,' or 'solve easy conflicts first';
- Contextual factors—such as aircraft type, destination, distance to go; and
- 'No-No's'— control strategies, eleven in total, that controllers will never use in conflict resolution. Examples include "never use speed as a resolution mechanism," or "never leave conflict aircraft not locked on heading."

Westin [39] reviewed the MUFASA project's exploration of controller resolution strategies both within- and across controllers, for reasons of developing advisory automation. Using a classification framework (i.e., resolution type/direction/degree etc), it was shown that intra-controller agreement (i.e. consistency) was higher than inter-controller agreement, and that inter-controller agreement was lower for specific manoeuvre choices.

Having a ML algorithm that learns 'Principles' and 'Control strategies' context based on a single ATCO or to a wider category based on ATCO control strategies type would be a way forward to improve AI support based on ATCOs feedback.

3.3.5 Summary of the results and conclusions

Here below a summary of the results presented above. From the outcomes we can see that the level of expertise played an important role in terms of visualization tools acceptance.

Acceptance. During the debriefings, when asked about their preference 11/11 ATCOs reported that they preferred the Black box (BB) solution, even if one of them also liked the concept of the Heat map (HM). The main reasons for the BB preference were that it was more straight forward, easy to understand and mainly it allowed them to make their decision in less time compared to the heat map (HM) or the storyboard (SB) solution. The students that preferred the HM mentioned the fact that it was visually appealing, interesting and the use of colour was appreciated.

The BB condition resulted being significantly more acceptable than both the HM ($p = 0.033$) and the SB ($p < 0.001$). No significative differences in the interaction between the condition and the complexity of the scenarios has been found.
In this study we obtained contrasting results in terms of level of acceptability between the self-reported results coming from the questionnaire and debriefing results and neurophysiological measurements (approach-withdrawal index). Although in experiments we strive to obtain common findings between the measurements and techniques we must also recognise that the explanation for discrepancy between these results can also be attributed to the fact that there is a possibility that they can measure separate constructs. This is a challenge that has been identified and studied in literature about the correlation between implicit and explicit measures [20]. Therefore, we must consider the findings complementary and accept some degree of uncertainty about the obtained results. One possible explanation for the mismatch in the outcomes could be related to the limitations of the experiment that was focusing on a single operational task (conflict resolution) and low fidelity setting. Therefore, further research is needed to better understand this mismatch in the results in Air Traffic Control.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

Level of understanding. A post-hoc comparison between the expertise level of the participants showed that there is a significant difference in understanding between students and experts, resulting in the student group having a higher understanding of the solution ($p = 0.018$).

Level of agreement. In general, experts were accepting/agreeing with the proposed AI resolution less frequently. These results might be explained by the fact that students were less critical of the HM visualization modality and some of them even mentioned that it was the resolution visualization condition they preferred because it was visually appealing.

Approach-withdrawal (Acceptability). Students exhibited a higher acceptability of the HM solution, with respect to the other two solutions. Experts experienced a higher acceptance of the storyboard solution that was significantly higher than the BB solution.

Workload. In terms of The HM showed the lowest level of workload with respect to the other solutions, both for experts and students. This behaviour appeared just during the 1st repetition, hypothesising a habituation effect for the black box and storyboard conditions.

Stress (Mental). Students exhibited the higher stress during the BB condition, and a decreasing during the HM and even more during the storyboard condition. Experts did not exhibit any significant change in stress among the conditions.

Arousal (Emotional Stress). Experts exhibited a lower arousal in the HM solution, with respect to the BB solution.

Trust. Students had a significantly higher trust in the presented resolution advisory than the experts ($p = 0.009$). Some ATCOs, especially professional ATCOs, mentioned that they felt they would need more information and training on how the ML algorithm to trust it. ATCOs consider that trust is very important of operational performance, but it must be acquired before or after operational usage, either in training, with briefing or even during debriefings.

Safety. ATCOs felt that in more complex scenarios or when they are experiencing more workload, they could be more willing to accept the solutions proposed by the tool. The participants that mentioned more frequently that the proposed solution was making them doubt their own solutions. This could point us in the direction of the importance of the solutions conformance to the ATCOs strategies and safety.

Performance improvements. In general, experts were less optimistic about the conflict resolution visualisation in terms of performance improvement. Higher explainability could be more useful for less timely critical or tasks or operational phases in which the ATCOs are subject to lower risk of cognitive workload, like planning tasks.

Personalization of ML algorithms and ATCO strategies. Having a ML algorithm that learns 'Principles' and 'Control strategies' context based on a single ATCO or to a wider category based on ATCO control strategies type would be a way forward to improve AI support based on ATCOs feedback.

Training. At the same time debriefings most participants admitted that some of the solutions (HM and SB) that were more complex could be an added value for training. The higher the visual XAI the better for training, because they elicit better the reasons behind the proposed conflict detection and on the resolution itself. Participants even mentioned that they could see the solutions with higher XAI

EUROPEAN PARTNERSHIP

Co-funded by the European Union

visualisation to have potential to be used during trainings with AI tools and once the trainees see how the ML algorithm works and to build trust

Limitations of the experiment:

Repetitions. They have been discovered different behaviours among repetitions, especially regarding workload and arousal. This could be due to the different engagement requested by the users (two or three aircrafts crossing), or by habituation effect of the proposed solution. This aspect would need to be further investigated.

Sample size. Most of the significant results were not strictly significant results, but just trends. This is due to the different behaviours showed by the two groups (i.e., students and experts), inducing a decreasing in the overall statistical power.

Low fidelity simulation setting. The low fidelity simulations might have impacted the outcomes since this experiment was focusing on single conflict scenarios and ATCOs were not using the tools in an environment closer to real operations. Some ATCOs mentioned that it was hard to recall the level, the speed, the trajectory of surrounding traffic and type of aircraft to properly assess the conflict. But the setting helped us focusing on the actual solutions and in detailed HMI aspects.

3.3.6 Lessons learned from the XAI applied to conflict resolution visualisation tools

1. **XAI / Transparency might have negative effects on performance and acceptability in conflict resolution tasks.**

Most ATCOs mentioned that if they would need more time to analyse and double check the proposals with higher XAI and that could ultimately translate in an increased workload during operations and/or in the worst case causing cognitive tunnelling while using the tools. Therefore, we can say that higher levels of transparency might have the opposite effect and lower controllers' acceptance of the system.

The obtained outcomes are tightly related to the type of task and the complex environment in which ATCOs are working, En route and TMA conflict solving is a very dynamic task and can quickly shift, and due to the nature of the task the human must stay ahead of possible conflicts that might be evolving.

2. **XAI/Transparency should be applied in operational phases that are not so timely constrained**

During debriefings ATCOs, especially experts, mentioned that if they are to be supported in conflict resolution and decision-making in a tactical role, they would need less information because they are time constrained to decide while maintaining situational awareness of their sector. The transparency that is conveyed by the visual features that were explored (apart from the BB solution) seemed to be hard to cope in terms of available time to decide while staying ahead.

Something that was also discussed was that if there is no time available to analyse the proposal made by the system, the ATCO could be pushed to accept a solution that is not correct, and this could have grave impacts for safety. Therefore, we can say that the potential of XAI/ transparency should be

Co-funded by the European Union

explored in less time critical phases of operation and for planning tasks. It would be interesting to explore transparency for a more strategical role in a sector, for instance a planner, but this is a recommendation for future activities and exploration.

In the end, for time pressured task, we would recommend having visuals that ease the creation of a mental model to solve the problem, instead of having a tool that gives some advice, that seem to be compared with the solution the operators will take, and in the end take more workload.

This doesn't exclude adding transparency to the system that will be used in operational phase that are timely constrained. Transparency will still be useful in post-operation to understand unexpected behaviour, in integration or tool development to verify the behaviour of the AI system, or in training with the operators to gain trust in the system.

3. **The parameters that are used to train the algorithms should be carefully selected because they can introduce bias in the AI /proposals.**

Usually, Air Traffic Controllers take decisions about resolutions based on the parameters they have available when it is time to take a specific decision and convert it in an action. Therefore, AI tools should consider not only all the parameters that need to be considered to solve a conflict, but also filtering just the important ones for the specific resolution. For example, the parameters defining an optimal resolution could not be the best parameters to take into consideration for that specific conflict, making the proposal optimal in abstract contexts, but less optimal in an operational environment: any multi-criteria optimising system will introduce some bias if the criteria are truly independent. In the CD&R use case, avoiding conflicts was the priority (huge bias, but a good one), then both other criteria, 'number of heading changes' and 'length of the trajectories' had the same importance in choosing the solutions (bias that can influence acceptance). This led some simple solutions with two aircrafts to change only one heading, while some operators would like to change to equally. Proposing different solution considering different parameters could be a solution, conforming the behaviour of the system to the operator behaviour could also be one

4. **Conformal AI solutions have the potential to achieve higher acceptance from ATCOs.**

The debriefings carried out post-experiment shed some light into the reasons why ATCOs accepted or refused the proposed resolutions and we found out that even if at times they would say that the resolution seemed acceptable, they refuse because it was not quite how they would have solved it.

Even if the main focus of the experiment was not to test the impact of conformance impact on acceptance, our outcomes (mainly from qualitative results) seem to confirm something that was already mentioned in literature (Westin, Borst Hillburn, 2016) [42] which is that more conformal decision aids, meaning aids that are closer to individual problem-solving styles, can improve acceptance. We also have noticed a trend that expert ATCOs brought up more often the importance of having more conformal proposed solutions to improve their acceptance of those proposals. This was also suggested in Westin et al. (2016) [42] study, conformance may only be relevant for expert users who hold consistent and well-developed decision-making strategies.

On a final note, participants voiced their interest in the use of applications of AI and ML algorithms to learn strategies from them.

5. **Transparency could support humans in building Trust in AI tools.**

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Trust in the solutions or tools that involve AI is a requirement to use it in operations, there should be no surprises. ATCOs while dealing with these tools and they should know how the tools work, how the ML algorithms are learning, and which type of variables are used while learning and they should know the limitations of those same tools. Our results highlighted those participants, during the CD&R visualisation tools experiment, agreed that trust in these tools must be acquired before or right after operational usage, meaning as training, with briefing or even during debriefing tools. Therefore, XAI can potentially be more important during those phases and not during the operational use of AI tools.

6. **Less trained ATCOs might be more willing to adopt new tools and innovative HMIs.**

Expert ATCOs are biased toward using the tool they are used to have. One ATCO expert in particular mentioned his experience with new tools that young adopted with time, and they didn't because they rather use the one, they had before, even if they saw it helped the younger ATCO solve faster the conflicts. Working several years with some tools require a lot of time to work with new tools and accept them. Training in a validation process will hardly correct this bias

7. **Using optimal tools and explanations of the tools during training could be beneficial**

From the discussions we can hypothesise that using Explainable AI for training purposes can be helpful in supporting in creating adequate and more complete mental model of the conflict, having the opportunity to be trained both in elaborating a functional solution and comparing it to an optimal one.

3.4 Demonstration activity in Virtual Reality environment

One proof of concept using Virtual reality was developed adapting an already existing tool, FiberClay. This POC is taking advantage of the third dimension given by the virtual reality environment to differentiate a subset of candidate solution one from another. Instead of displaying all the candidate solutions on the same level, candidate solutions are displayed stacked one on top of the other. This allows, like in the POC, to order the candidate in function of different criteria, notably the fitness function.

Different sets of candidate solutions have been tested using the virtual reality environment, from the whole set of candidate solution created by the Genetic Algorithm to only the best candidate solution of each generation. While the first was too hard to read, the last one, because the Genetic Algorithm was quickly finding the best solution (while still exploring after), was too repetitive. A middle point was found, by displaying only a subset of candidate solution described in the following. To extract those subsets, we prioritized two aspects:

1) Showing both good (i.e., no conflict) and bad (i.e., with conflict) candidate solution, to keep a diverse subset, and allow contra pervasive reasoning ("Why not questions"), like when showing all candidate solutions.
2) Avoid repetition of candidate solutions. Indeed, due to the GA process, candidate solutions are repeated a lot during convergence.

In the subset respecting those two statements, we decided to show the best 1% of the candidate solution, and the best 1% of the candidate solution with conflict, which correspond roughly to 100 candidate solution of both good and bad candidate solutions.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

In the following, we show how such dataset can be interact with in our Virtual Reality proof of concept.

Figure 31: First point of view of the dataset: classical upper point of view, with latitude (Lat) and longitude (Long)

Figure 31 shows the first point of view of the dataset, a classical upper point of view that is, until moved, the same as a classical screen-based visualisation that allows to annotate (red writing on the right image). As one can see, the conflict this dataset is about was between three avions, one coming from South-East (SE) and going to North-West (NW), one from South-West (SW) to North-East (NE), and one from West-South-West (WSW) to East-Nort-East (ENE).

The point of view can be changed, by dragging the dataset to the right, left, up, down, or any translation, but also rotated to have a better view of the depth use (see **Figure 31**). This last transformation allows to see the good and bad candidate solution in the dataset. Green candidate solution being bad candidate solution, and blue ones being good candidate solutions. From this point of view, the user can select them, or inversely to make sure previously selected candidate solutions are good. Both facets are used in the following.

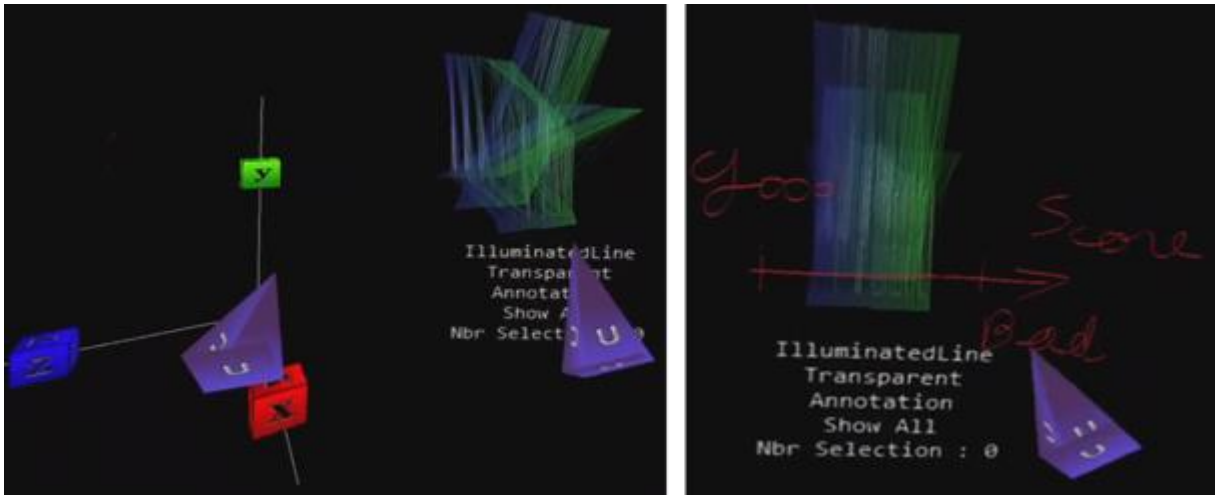


Figure 32: Rotating the dataset (left picture) allows to better view the depth axis (right picture). In the right picture, on can see the good (left of the axis) and bad (right of the axis) candidate solution

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Knowing those interactions, the user can start by exploring the dataset. A good way to explore it is to select candidate solutions and display only those, like in Figure 32.

While in this last mode, the user can focus different candidate solutions by hovering the whole dataset, see Figure 33. This exploration allows primarily to look what have been tested for one aircraft and how those actions have influenced other aircraft. In this example, the user is exploring the candidate solutions in function of the trajectory of the aircraft going from SE to NW. The user observes the candidate solutions if the aircraft goes to the right (left picture), if it doesn't change its trajectory

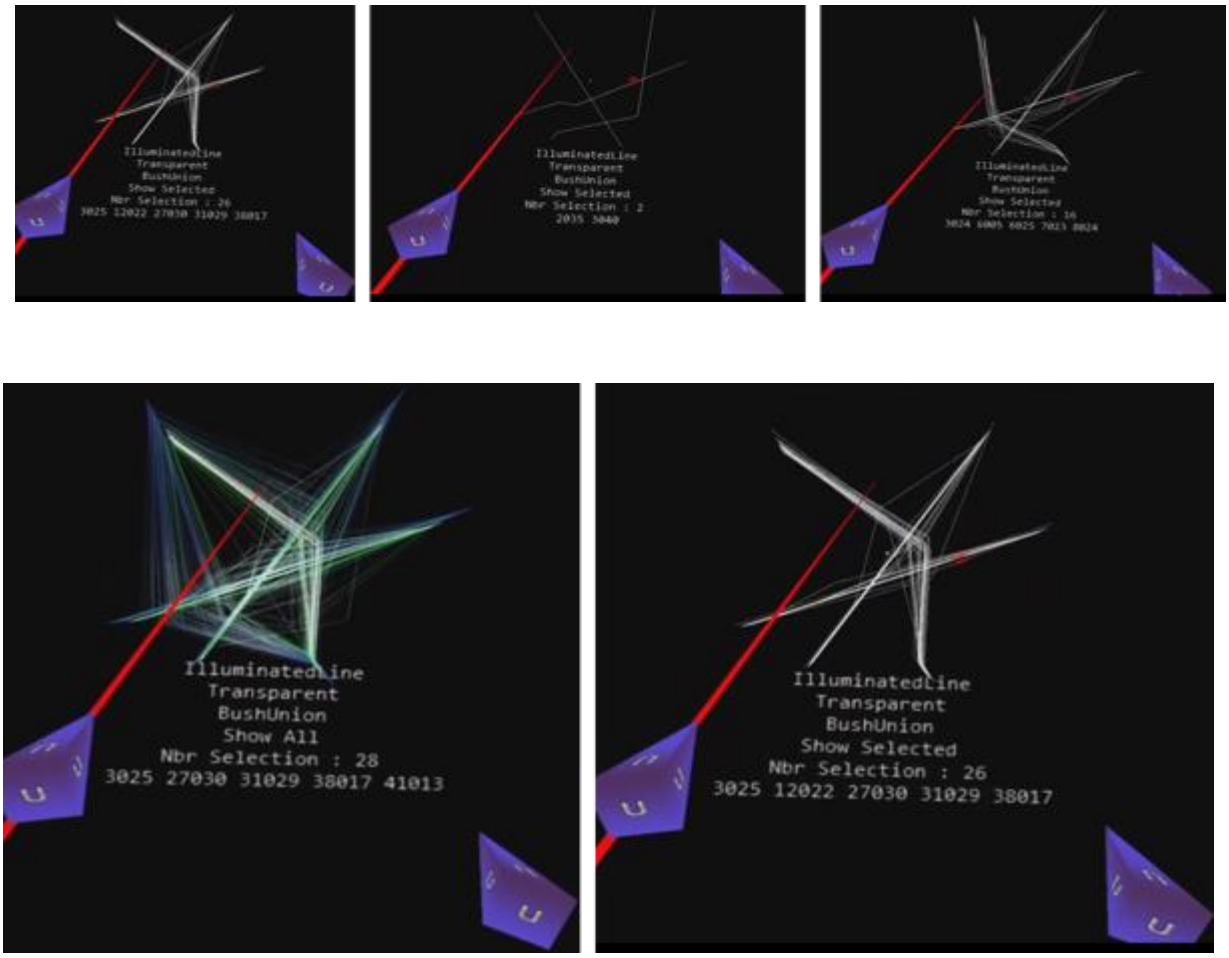


Figure 33: Hovering the whole dataset in the display only selected mode.

(middle picture), and if it goes left (right picture). During the exploration it seems that changing the trajectory of this aircraft diminish the need to change other aircraft trajectories, but to make sure, the user needs to observe the score of those candidate solutions.

To do so, the kept the last selected candidate solution from previous exploration, i.e., candidate solutions where the aircraft going from SE to NW is modified to go to the left, see Figure 34. It then looks at the score of the selected candidate solutions and observe that those selected candidate are good.

Co-funded by the European Union

Eager to understand what bad candidate solutions in the dataset are, the user de select those candidate solution, rotate the dataset to see the score of the solutions, and select the worst ranked candidate solutions (bad ones) and look at them, see Figure 6. The user can then observe that most likely, the aircraft going from SE to NW needs to go to the left, with a big enough heading change, to avoid conflicts.

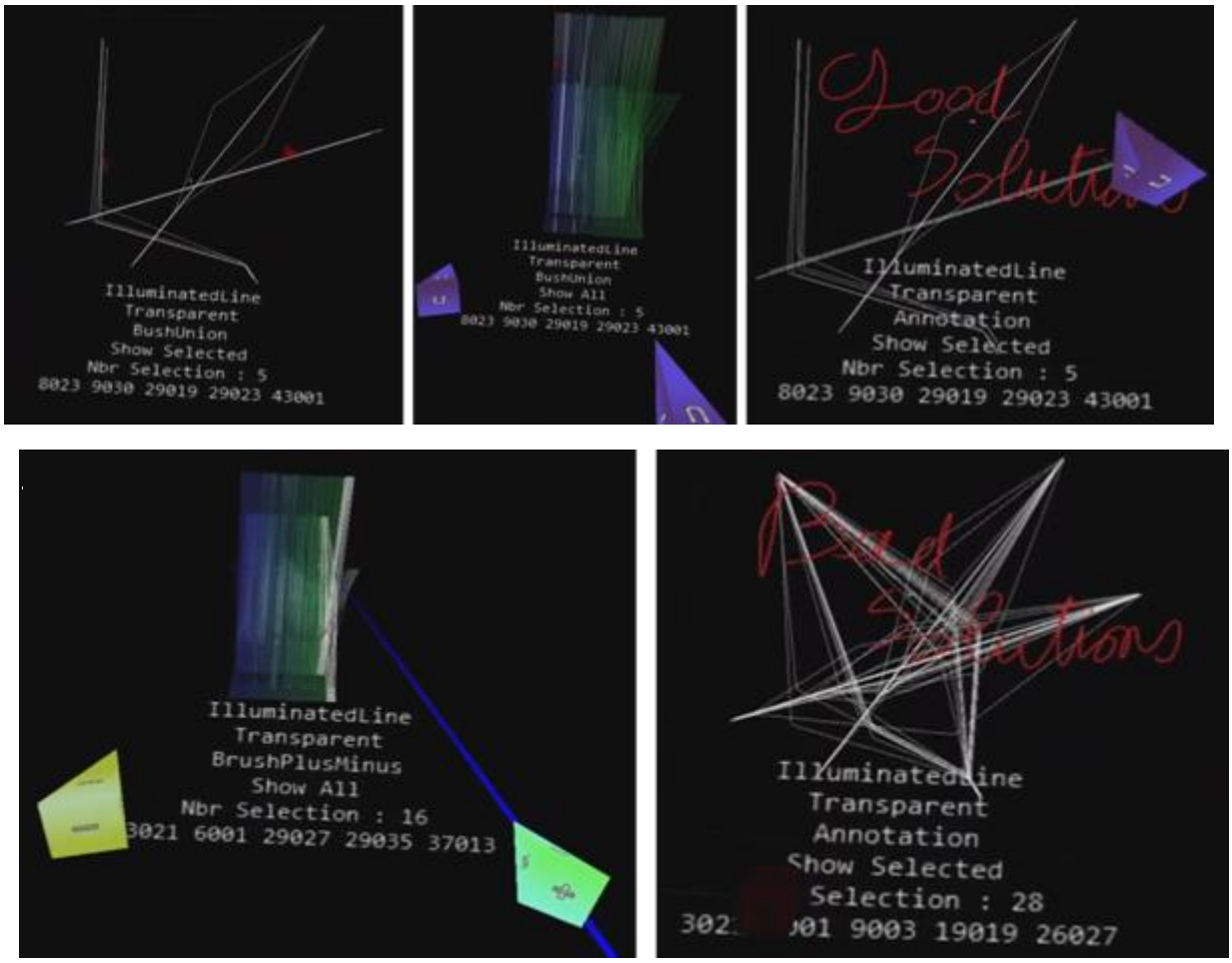


Figure 34: Visualisation mode

As for the model centric explanation reported in D4.2, the virtual reality environment can be very helpful to explain the behaviour of the GA. Using the third dimension allows various dataset presentation and filtering. This kind of interface would be more useful to the developer of the tool, or in post operation, to analyse a behaviour, and certify it. Unless a complete shift from screen-based visualisation to immersive visualisation of all the tools required on one task, adding immersive visualisation is not recommended in time critical tasks. E.g., if immersive visualisation would be added to the ATCO tools, it would require having every tool the ATCO has at his disposal in the immersive environment.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

4 Delay prediction and propagation results

4.1 Description of models and tools concept

We will briefly introduce the concept of explainable AI before discussing AI models and explainability tools.

Generally, an automated system developed with AI contains the components shown in Figure 16. Here, some data is fed into a machine learning model, and the user expects some output. This output can take the form of predicting some value or classifying some object. However, it is observed that people often have questions regarding the decision of the machine learning model, for example, why does the result appear this way? How was it calculated? etc.

Explainable artificial intelligence emerges to provide the solutions to these kinds of questions and to make automated systems transparent. Here, the main flow of operation remains unchanged, but the machine learning model becomes more transparent. This occurs when a method or tool describes the inference mechanism of a machine learning model to end users by presenting an explained output.

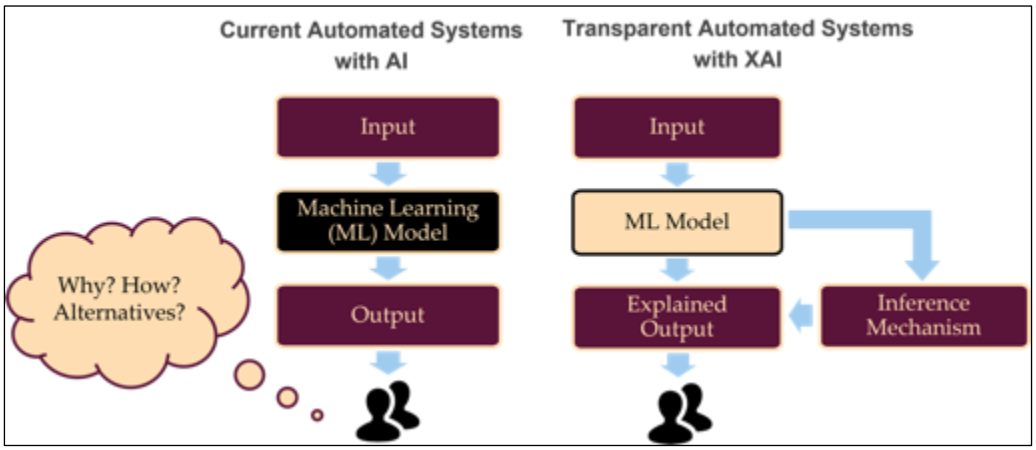


Figure 35: Intuition of Explainable AI (XAI)

In this project, based on the literature study reported in [6], Random Forest (RF), eXtreme Gradient Boosting (XGBoost) for delay prediction and Long Short-Term Memory (LSTM) for delay propagation models were adapted to perform the ATM tasks. More information regarding these methods is detailed in deliverable [8].

Random Forest (RF): The RF algorithm, which consists of a collection of randomised decision trees, is one of the well-known ensemble algorithms of machine learning utilised in numerous ATM-linked tasks [12]. To produce the random forest, N decision trees are combined, and then the predictions are made for each tree from the first step. In short, the working process of RF can be explained as follows: first, pick random K data points from the training set. Next, construct the decision trees associated with the selected data points (Subsets). Then, choose N for the size of the decision trees you wish to build. After

EUROPEAN PARTNERSHIP
Co-funded by the European Union

that, repeat steps 1 and 2. Finally, find the predictions of each decision tree for new data points, and assign the new data points to the category that wins the majority votes.

eXtreme Gradient Boosting (XGBoost): XGBoost is a scalable ML algorithm for tree boosting used in regression and classification applications [26]. An ensemble of weak prediction models is used to make the final prediction, and a weak prediction model refers to the randomised decision trees produced by RF. Two significant enhancements over the existing tree ensemble approaches have been made in XGBoost: i) optimisation of splitting the branches of a tree and ii) scalability. Because it iterates over all potential splits in the data, the current tree-boosting mechanism enables a greedy method that is computationally expensive. On the other hand, XGBoost splits the data using an approximation approach, which requires less computation [25][1]. Additionally, it supports the mechanism's scalability and can be used in a distributed architecture.

Long Short-Term Memory (LSTM): LSTM networks are a special kind of Recurrent Neural Network (RNN) proposed by [19]. It is designed to support the long-term dependency problem by introducing a memory cell and two gates into the network, consisting of five layers. The essential layers of an LSTM network are a sequence input layer and an LSTM layer. The time-series data are formed into sequences fed into the network's input layer. Figure 17 demonstrates the architecture of a simple LSTM network for time series classification of take-off delay prediction and propagation. The network starts with a sequence input layer followed by an LSTM layer. To predict class labels, the network ends with a fully connected layer, a softmax layer, and a classification output layer.

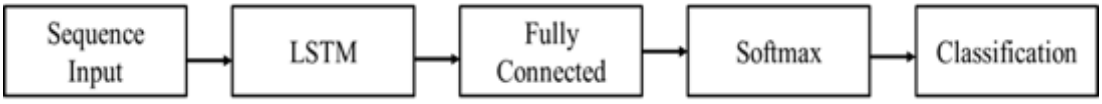


Figure 36: Block diagram of LSTM architecture for the time series classification

In terms of explainability three most common explainability tools are summarised below. These techniques were primarily applied in this project to provide an explanation for the decisions made by AI models. For more detail, we ask to see previous deliverables 4.2 [8]

Local Interpretable Model-agnostic Explanation (LIME): LIME [31] is a tool that uses an interpretable model to approximate each prediction made by any black box ML model. LIME uses a three-step process to determine the specific contributions of the chosen features: perturbing the original data points, feeding them to the black-box model, and then observing the related predictions. In the delay prediction task, each prediction of a take-off delay is shown together with a list of parameters that contributes to the delay and their weights in that prediction.

Shapley Additive Explanations (SHAP): A mathematical technique called SHAP was developed based on the Shapley Values proposed by Shapley in the cooperative game theory [38]. Shapley values are a mechanism to fairly assign impact to parameters that might not have an equal influence on the predictions. The Shapley value concept was incorporated to generate additive explanations for predictions from black-box models [39]. In delay prediction, to explain the decisions from the model (i.e., prediction), SHAP calculates each parameter's contribution to the model's prediction.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

moDel-Agnostic Language for Exploration and eXplanations (DALEX): DALEX is a python library proposed by Biecek [10] that is built upon the software for explainable machine learning presented by Biecek [10]. The main goal of the DALEX tool is to create a level of abstraction around a model that makes it easier to explore and explain the model. Explanation deals with two uncertainty levels, model level and explanation level. The underlying idea is to capture the contribution of a parameter to the model's prediction by computing the shift in the expected value of the prediction while fixing the values of other parameters. In this project for (delay prediction and propagation), DALEX works as a Breakdown plot, which detects local interactions of user-selected parameters.

4.2 Algorithmic evaluation

Regarding experimentation for both ATM tasks (i.e., delay prediction and propagation), the validation was done from two functional perspectives: i) prediction models and ii) explainability tools. Furthermore, the validation was carried out based on the dataset collected and processed by EUROCONTROL, and it uses the Enhanced Tactical Flow Management System (ETFMS) flight data with (EFD) messages for all flights during the year 2019 (i.e., May to October). In addition, several features in the dataset were subject to pre-processing steps (i.e., encoding of categorical values, missing values, etc.); further information regarding these steps was detailed in deliverable 4.2 [8]. A summary of the cleaned dataset used in the experiment is presented in Table 5.

Table 6: A summary of the experiment dataset

Data set	Description
Train Set	May – August ≈ 5903743 instances (approx. 78% of the clean dataset)
Test Set	September – October ≈ 1709841 instances (approx. 22% of the whole dataset)

4.2.1 Delay prediction

Descriptions of the validation of prediction models and explanation models are stated below.

1) Validation of Prediction Models: Both the prediction models RF and XGBoost were trained using the flight the from May-August 2019 (5.9M instances) and tested using flight data from September - October 2019 (1.7M instances). To assess the performance of the predictors, mean absolute error (MAE) values were used which computes the average difference between an actual observation and a prediction from a model:

$$MAE = \frac{1}{n}\sum_{i=1}^{i=n} |y_i - \hat{y_i}| \quad (1)$$

Co-funded by the European Union

Here, MAE was computed considering y_i, the actual delay as the baseline against $\hat{y}_i$ as the delay predicted by the models. On the whole test set, both the models performed better than the past similar work [14] in terms of MAE while predicting the take-off delay. To assess in more details, the dataset was sliced based on time remaining in minutes till the estimated of-block time (EOBT). The chunks were made as listed in Table 6 with prediction performance of exiting ETFMS, gradient boosted decision tree (GBDT), RF and XGBoost. From the table, it is observed that the RF and XGBoost performed better than the previous ones for all the intervals and notably, XGBoost surpassed the RF. Considering this outcome, the explanation models were developed to explain the predictions of XGBoost only.

Table 7: Comparison of performances on take-off delay prediction from ETFMS, GBDT, RF and xGBoost using the MAE (in minutes), lower is better and minimum values are highlighted. The MAE values for the ETFMS and GBDT are considered as reference from an experimentation performed by Dalmau et al. [12]

Time to EOBT	ETFMS [12]	GBDT [12]	RF	XGBoost
(0, 15)	10.7	8.8	7.22	7.51
(15, 30)	12.4	10.2	8.75	8.47
(30, 60)	13.3	10.5	9.05	9.05
(60, 90)	14.3	10.8	9.46	9.12
(90, 120)	14.3	11.1	9.83	9.50
(120, 180)	19.1	13.5	11.09	10.50
(180, 240)	23.0	15.4	11.58	11.46
(240, 360)	21.2	15.1	11.93	12.02

2) Validation of Explanation Model: From functional perspective, LIME and SHAP tools tries to mimic the prediction of the trained models and determines the important features to explain the prediction. To evaluate the similarity between the predictions by explanation tools and XGBoost, *local accuracy* was considered as MAE in minutes. And to assess the relevance of their chosen features based on their importance values normalized discounted cumulative gain (nDCG) was considered. nDCG compares the order of retrieved documents in information retrieval [13][40]. Results of comparison between LIME and SHAP are summarised in Table 7 for all the instances, top 100k and 67k instances in terms of most accurate predictions respectively. Finally, based on the comparison, SHAP was used to generate the visualisations to explain the predicted take-off delay prediction.

Table 8: Progression of both local accuracy in MAE and nDCG values for SHAP and LIME. nDCG is compared against the sequence of the important features from prediction model. Rows show the number of instances. for local accuracy in MAE, lower is better. For nDCG, higher is better. Best values are highlited

XAI Model	SHAP	LIME

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Metric	local accuracy	nDCG	local accuracy	nDCG
All	3.3e-6	0.806	8.62	0.882
100k	1.1e-6	0.722	4.75	0.847
67k	6.2e-7	0.717	3.13	0.800

4.2.2 Delay propagation

Below are two descriptions of the validation of propagation models and validation of explanation models.

Validation of Propagation Models: For the delay propagation task, three ML models, (i.e., RF, XGBoost and LSTM models) were chosen to perform this validation. The same dataset summarised in Table 4 was used for the propagation task; however, to conduct this validation, some preparations need to be made on the dataset to obtain the delay propagation information and match the propagation state (i.e., the delay propagation is considered between two flights of the same REG number on the same day).

The following steps are considered for dataset preparation for delay propagation:

- For specific REG (registration) number, extract all the rows
- Consider the last row for each IFPLID
- For each day, calculate the TAKE_OFF_DELAY differences between two IFPLID
- Create a new column (i.e., y_diff) and insert the TAKE_OFF_DELAY differences
- Create a new column (i.e., y_class) and insert '1' and '0' based on the following conditions
    - If y_diff>0, y_class ==1
    - Otherwise, y_class ==0
- Delete the first row in each day as that don't provide information of delay propagation
- Further, delete the two columns TAKE_OFF_DELAY, y_diff

After preparation, the dataset consisted of a total of 11528 instances. The dataset was then divided into training and test datasets, where training data contained the first 9222 instances, and the test dataset had the rest of the 2306 instances.

Validation of Delay propagation Model: For the evaluation, data are grouped into two classes, positive class indicates instances whose y_class value is 1, and negative class are those instances where y_class value is 0. Thus, all the instances in the dataset belonging to the delay propagated flights are denoted as the positive class (P), and the negative class (N) represents all the instances of the flights with no delay propagation in the dataset. Since binary classification is performed with ML models, sensitivity, specificity, and accuracy are calculated as performance metrics of the models.

True positive: an aircraft has a delay or lag in previous flights, and the test instance indicates delay propagation.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

True negative: an aircraft does not have a delay or lag in previous flights, and the test instance does not indicate delay propagation.

False positive: an aircraft does not have a delay or lag in previous flights, and the test instance indicates delay propagation.

False negative: an aircraft has a delay or lag in previous flights, and the test instance does not indicate delay propagation.

Sensitivity: measures the percentage of delayed flights that are correctly identified as a delay propagated.

$$\text{Sensitivity} = TP/(TP + FN)$$

Specificity: measures the percentage of not delayed flights correctly classified as the delay has not propagated.

$$\text{Specificity} = TN/(FP + TN)$$

Accuracy: the ratio of the number of true positive results to the number of total instances in the dataset.

$$\text{Accuracy} = (TP+TN)/(P + N)$$

Random forest, XGBoost, and LSTM are the three ML models built for the delay propagation classification. The performances of the three models are calculated on the test dataset. The total instances in the Delay propagated group (P) was 1139 and in the no-delay propagated group (N) was 1167. Table 8 shows the summary of the results.

Table 9: Summary of Delay propagation classification

	RF	XGBoost	LSTM
True positive (TP)	748	754	660
False positive (FP)	407	437	294
True negative (TN)	760	730	873
False negative (FN)	391	385	477
Sensitivity	0.66	0.66	0.65
Specificity	0.65	0.63	0.69
Accuracy	0.65	0.64	0.67

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Validation of Explanation Model: Since the visualisation generated by the expandability tools (i.e., LIME and SHAP) are same and the dataset contained same attributes for both delay prediction and delay propagation, the explanation for the delay propagation has not taken into consideration.

4.3 Description of the survey platform for user evaluation (Delay prediction)

An outline of the validation exercise is shown in Figure 18; in this validation, the total time of the exercise on the platform would be approximately 55 minutes.

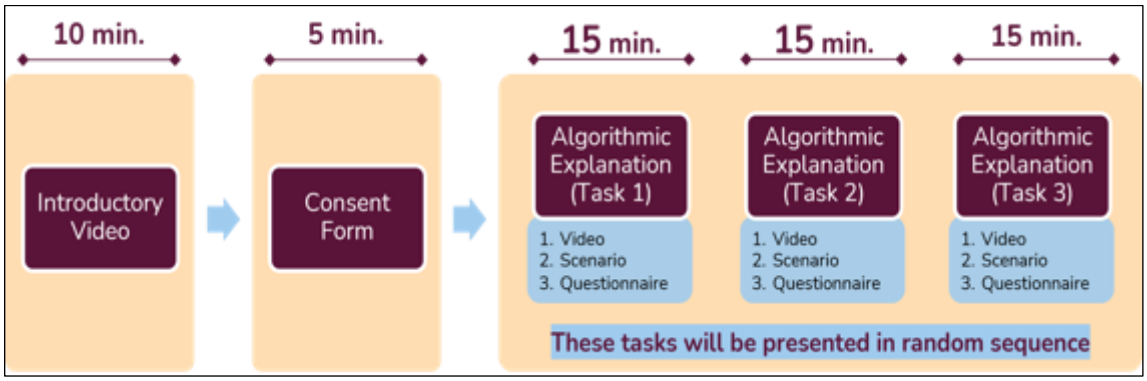


Figure 37: Validation Outline

At first, this introductory video of around 10 minutes in length. After this, at the beginning of the validation, the participant will need to fill up the consent form about participating in the validation exercise. Then, there will be three tasks for evaluating algorithmic explanations. Each of these tasks contains watching a short video on the working principle of the method, followed by a scenario and a questionnaire to gather your responses. Altogether, each task will require approximately 15 minutes, and these tasks will be automatically randomised.

Validation Platform Setup:

- Platform Development: The validation tool was developed using python 3.9 programming language, and several libraries and PostgreSQL were used to build a database. To reproduce the validation tool, the following packages presented in Table 9 must be installed.
- Platform Deployment: The tool was deployed using free Heroku Postgres.
- Data Storage: The surveys (questionnaires) data was stored on the hard disk and kept in a protected closet at MDU facilities after processing the personal data.

Table 10: Required Libraries

Library	Description
dj-database-url==1.0.0	Used to extract the Django database configuration from an environment variable.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

django==4.1	High-level Python web framework that encourages rapid development and clean, pragmatic design.
gunicorn==20.1.0	Pure-Python HTTP server used for serving Django WSGI applications on Heroku
mkl==2021.4.0	Math Kernel Library is a computing math library of highly optimized, extensively threaded routines for applications that require maximum performance.
numpy==1.22.3	Needed for several mathematical operations
pandas==1.4.3	Used as a manipulation tool for the data
plotly==5.6.0	Needed for generating the plot
psycopg2-binary==2.9.3	Used to work with Postgres databases
sqlparse==0.4.1	Provides support for parsing, splitting and formatting SQL statements.
wheel==0.37.1	Used as setuptools extension and command line tool
whitenoise==6.2.0	Allows the web app to serve its own static files.
django-crispy-forms==1.14.0	Helps to manage Django forms.

Validation Platform Parts:

The validation platform consists of three parts: First, the **home page** shown in Figure 19, which contains information about the study and an introductory video which describe the methods and instructions for each task. Next, the **List of Parameters** is illustrated in Figure 20; in this part, all the parameters and their description are presented. Finally, **Start the validation Task**; here, examining the explanation of delay prediction would begin by filling out the consent form shown in Figure 21.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

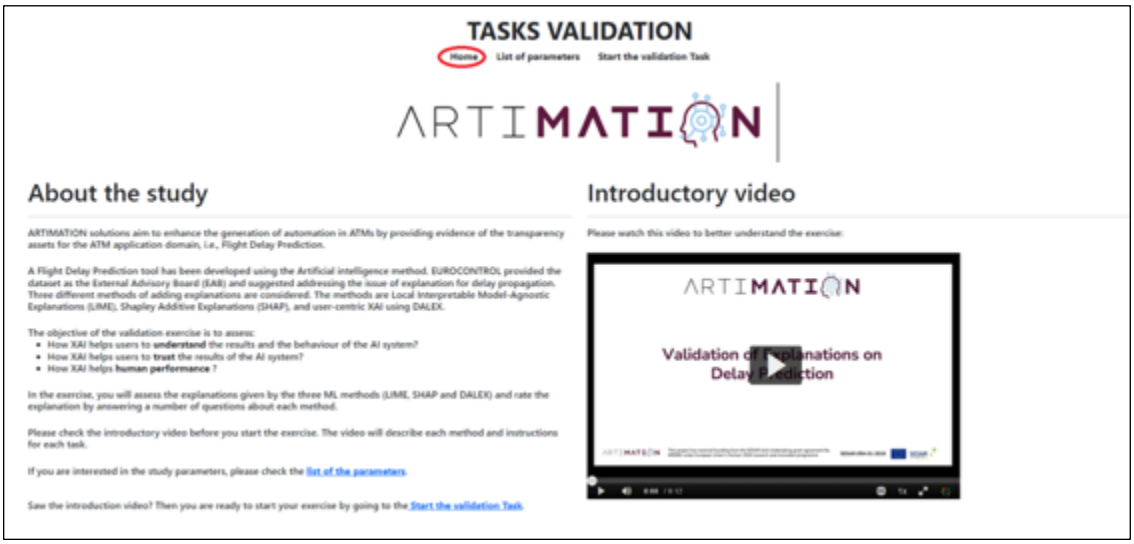
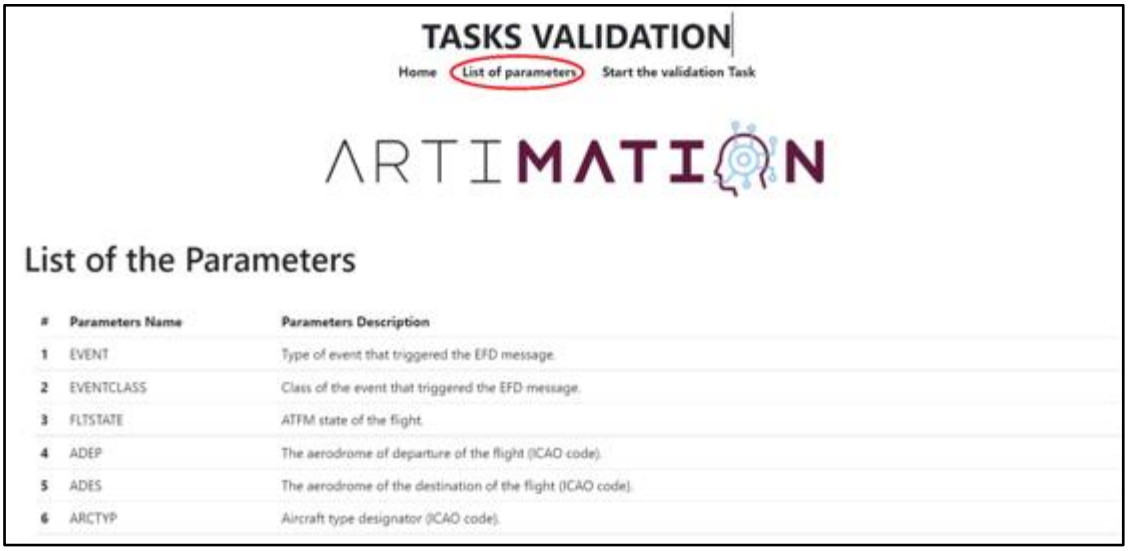


Figure 38: Home Page Of the validation platform



#	Parameters Name	Parameters Description
1	EVENT	Type of event that triggered the EFD message.
2	EVENTCLASS	Class of the event that triggered the EFD message.
3	FLTSTATE	ATFM state of the flight.
4	ADEP	The aerodrome of departure of the flight (ICAO code).
5	ADES	The aerodrome of the destination of the flight (ICAO code).
6	ARCTYP	Aircraft type designator (ICAO code).

Figure 39: List of Parameters

EUROPEAN PARTNERSHIP

Co-funded by the European Union

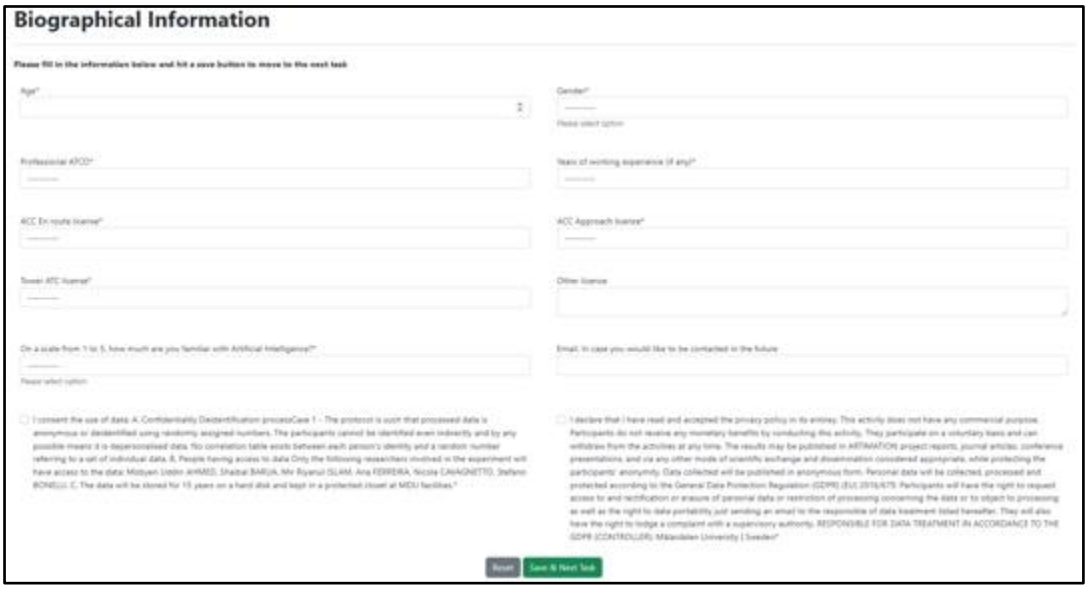


Figure 40: Start the validation Task (Biograph Information/consent form)

Workflow of the validation:

As we discussed at the beginning, the platform includes three delay prediction-based validation exercises with explanations from LIME, SHAP, and DALEX. Since the structure of the validation platform for both LIME (Task 1) and SHAP (Task 2) are the same, so we have considered only showing the LIME and DALEX workflows with platform screenshots.

For TASK 1, the participant will watch an introductory video, as shown in Figure 22, and then a scenario context of a delayed flight will be presented; an illustration of this scenario is shown in Figure 23.

- In case of explanations, the system based on LIME will present the breakdown plot shown in Figure 24, which contains important parameters contributing to the delay.
- At the end of the scenario, the user will be asked some questions to evaluate factors linked to human performance; Figure 25 presents these questions.

The Figures below present the steps for evaluating the explanation of delay prediction based on LIME

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Figure 41: Task 1(Introductory Video for LIME).

Figure 42: Task 1 (Scenario for LIME).

Figure 43: Task 1(Explanation on Delay Prediction from LIME).

Co-funded by the European Union

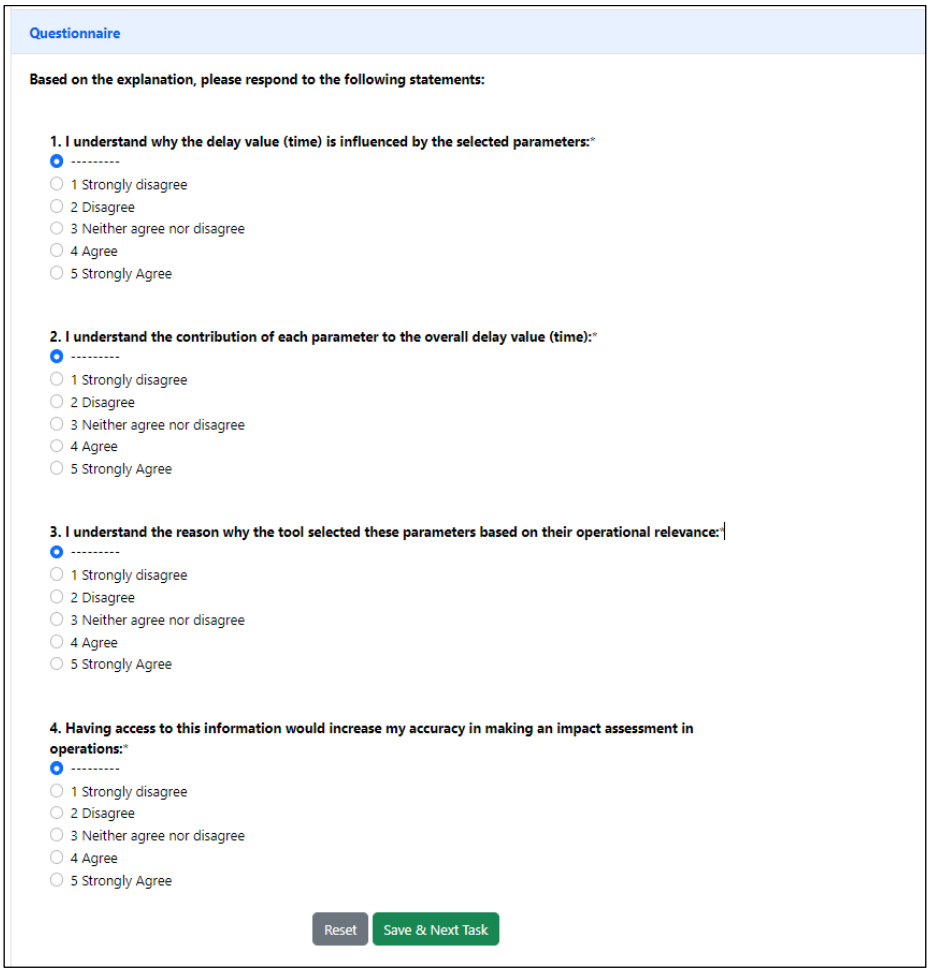


Figure 44: Task 1 (Questionnaire based on LIME Explanation).

For TASK 3, the participant will watch an introductory video, as shown in Figure 26, and then a scenario context of a delayed flight will be presented; an illustration of this scenario is shown in Figure 27.

- In case of explanations, the user will ask to select five parameters from a list of parameters, as shown in Figure 28.
- After that, the system based on the DALEX tool will present the breakdown plot shown in Figure 29, which contains the five selected parameters contributing to the delay.
- At the end of the scenario, the user will be asked some questions to evaluate factors linked to human performance; Figure 30 presents these questions.

The Figures below present the steps for evaluating the explanation of delay prediction based on DALEX.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Figure 45: Task 3 (Introductory Video for DALEX).

Figure 46: Task 3 (Scenario for DALEX).

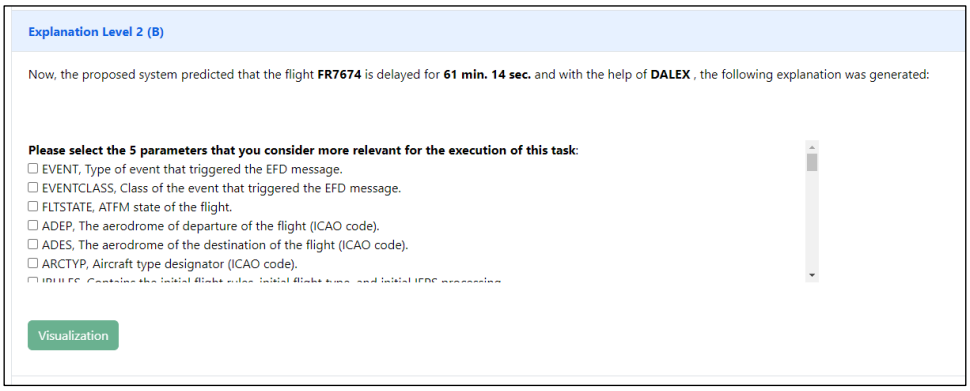


Figure 47: Task 3 (Selected Parameters for DALEX).

Co-funded by the European Union

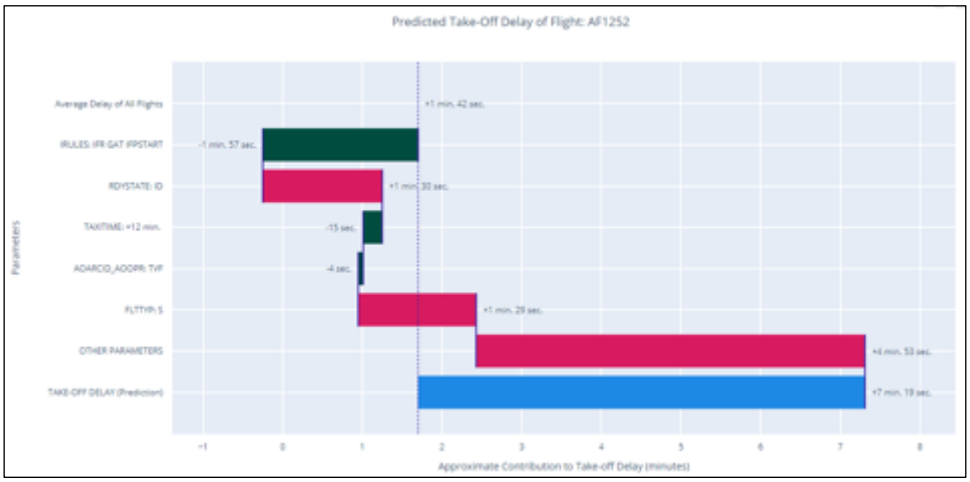


Figure 48: Task 3 (Explanation on Delay Prediction from DALEX).

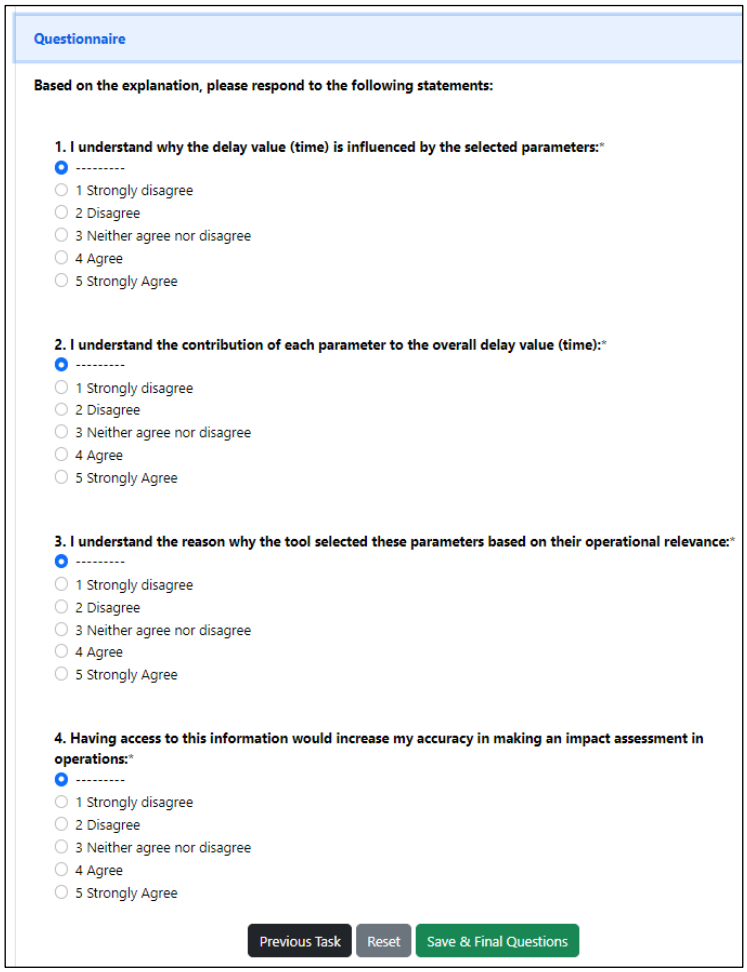


Figure 49: Task 3 (Questionnaire based on DALEX Explanation).

At the end of the validation, the user will ask to answer the final questionnaire presented in Figure 31. Finally, Acknowledgments for the participants are shown in Figure 32.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

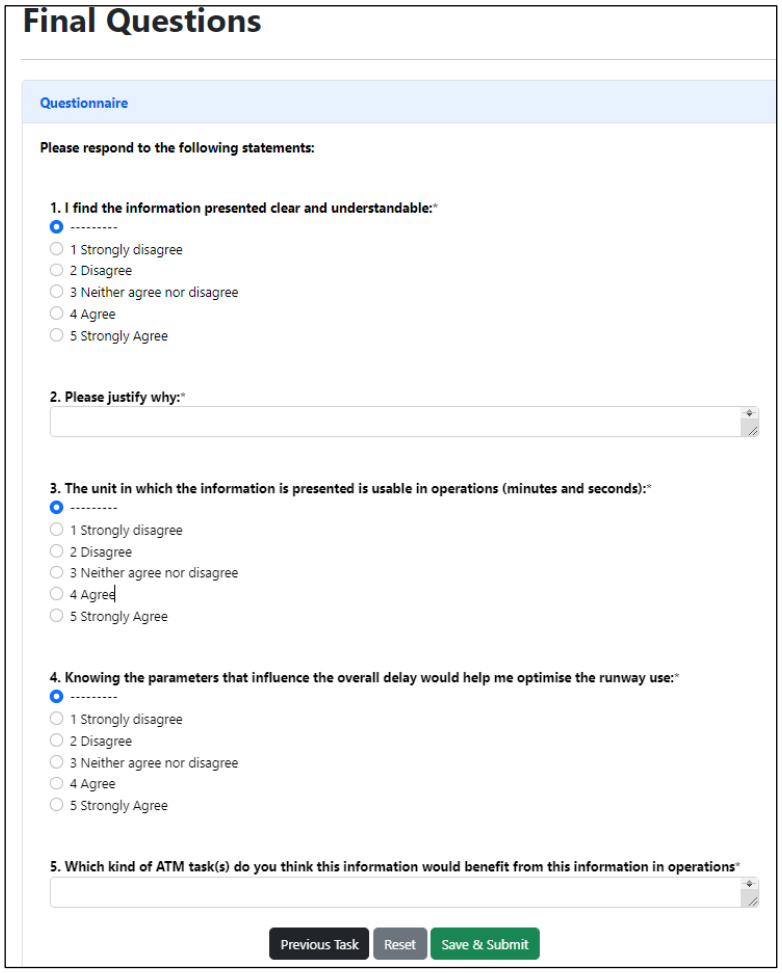


Figure 50: Final Questions

Figure 51: Acknowledgments.

Co-funded by the European Union

4.4 User validation result (Delay prediction)

The validation of the Delay Prediction use case followed the structure used to validate the Conflict Resolution use case. Therefore, two self-report questionnaire was administered: one after each condition (SHAP, LIME, DALEX), and a final questionnaire after the conclusion of all the tasks. The "post condition" self-report questionnaire aimed to assess constructs such as the understanding of the AI outcome and the impact on work performance of each tool; on the other hand, the final questionnaire aimed to assess the general usability and the impact on work performance of a Delay Prediction tool. Assuming that the estimated value of the predicted delay was correct, trust items were not considered. The Consortium tried to gather some qualitative input as well, both during the validation activity with open-ended questions, and with a qualitative interview with one representative of ANACNA (Italian Air Traffic Controllers Association), being the end-user involved in the Advisory Board of the project.

4.4.1 Post Condition assessment

LIME is a tool that uses an interpretable model to approximate each prediction made by any black box ML model.

LIME received global positive feedback from the respondents. In particular, the first understanding item (i.e., "I understand why the delay value (time) is influenced by the selected parameters") received positive feedback, showing that the algorithmic explanation provided by LIME affects positively the general understanding of the AI outcome. Conversely, the other understanding items provide more balanced feedback about the contribution and about the operational relevance of the parameters to the final delay. Positive feedback was gathered for the impact on work performance as well (i.e., "Having access to this information would increase my accuracy in making an impact assessment in operations").

SHAP is a mathematical technique based on the Shapley Values proposed by Shapley in the cooperative game theory.

SHAP gathered general positive feedback about the understanding and the impact on work performance as well. As for the LIME condition, the highest feedback about understanding regarded the influence of the selected parameters to the final delay, with less positive feedback about the understanding of the contribution of the parameters to the delay value and the reason why the parameters were selected based on the operational relevance.

DALEX is a python library proposed by Biecek [10] that is built upon the software for explainable machine learning presented by Biecek [10]

The feedback gathered from the DALEX self-report questionnaire shows a positive impact of the user-centred condition on both the understanding of the influence of the selected parameters and their contribution to the final delay. The impact on work performance received positive feedback as well.

The three methods were finally compared item by item.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

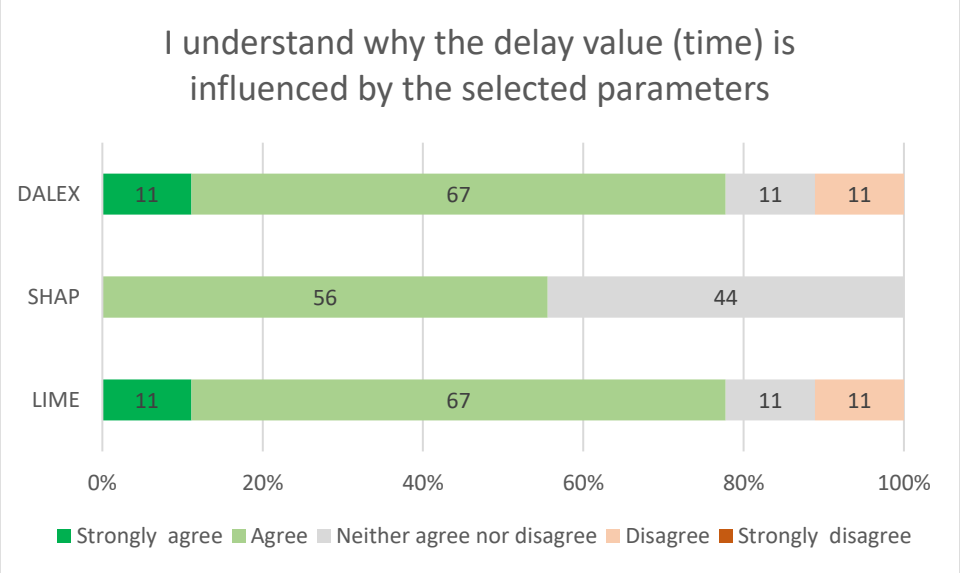


Figure 52: Method comparison for understanding

For the first understanding item, "I understand why the delay value (time) is influenced by the selected parameters", we can notice how both the LIME and DALEX conditions gathered the highest number of positive feedbacks, while SHAP received no negative feedback, resulting in a more balanced condition for the understanding of the influence of the parameters on the final delay.

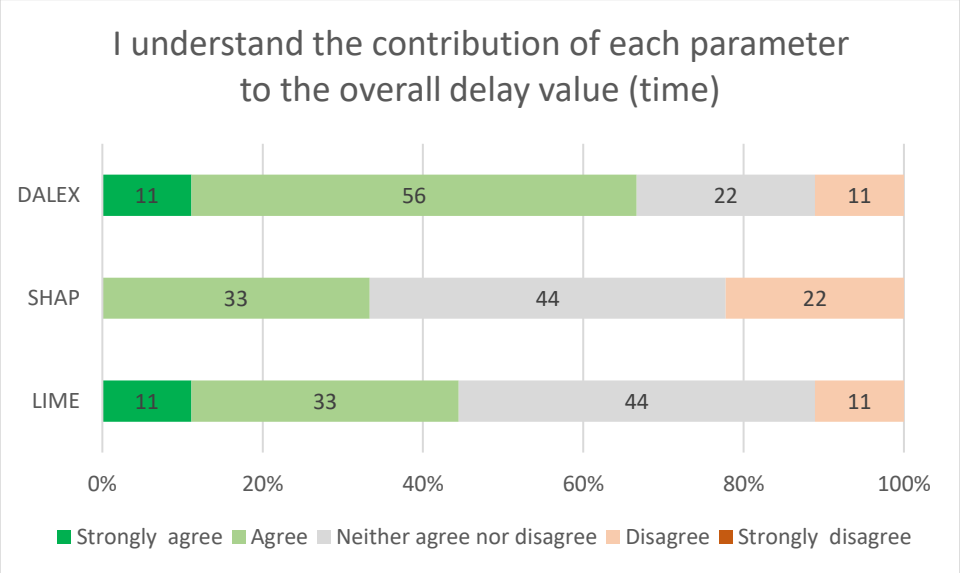


Figure 53: Method comparison for understanding (contribution of each parameter)

For the second understanding item, "I understand the contribution of each parameter to the overall delay value (time)", DALEX seems to be the best condition. Therefore, we could say that the user-centric selection of the parameters can positively influence the understanding of the contribution to the final delay value.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

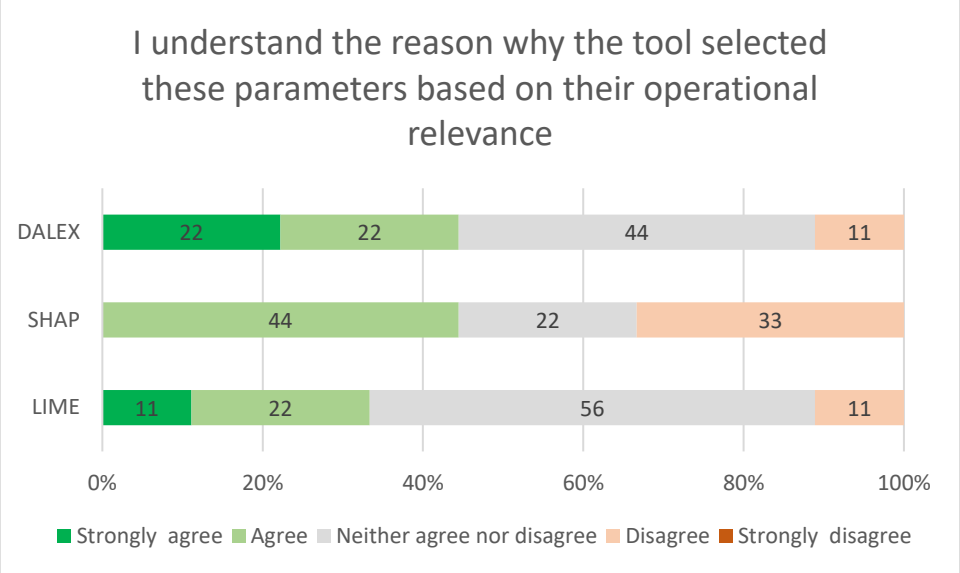


Figure 54: Method comparison for understanding (tool selection of parameters)

The last understanding item, "I understand why the tool selected these parameters based on their operational relevance", is the item with the most negative rating for all the conditions. Based on the quantitative feedback gathered, none of the conditions received at least the 50% of positive feedback about the understanding of the operational relevance of the selected parameters. This could mean that, even if the influence and the contribution on the final delay of the selected parameters for SHAP and LIME, and the contribution of the parameters selected by the users for DALEX, is high, this would not impact the operational relevance of the information received.

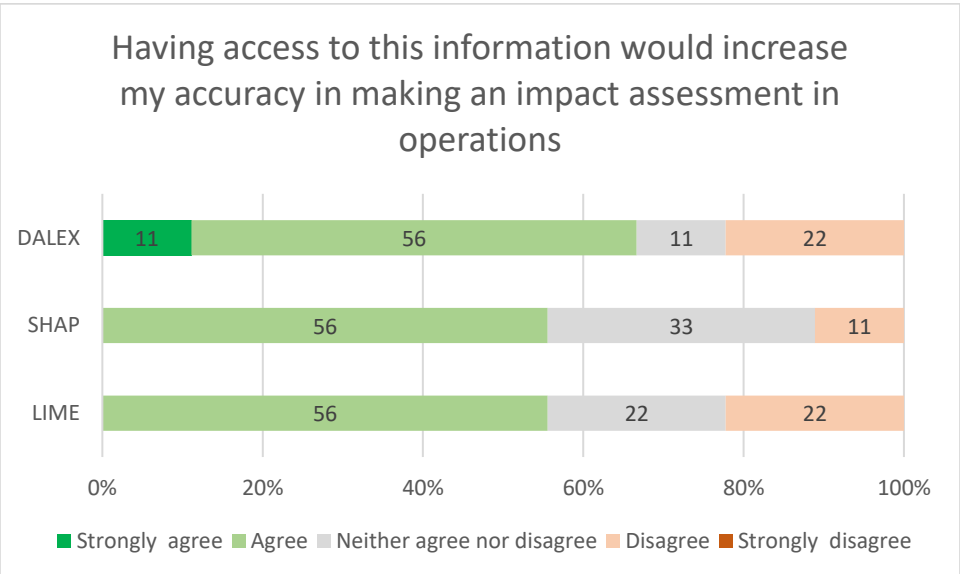


Figure 55: Method comparison on accuracy benefits for operational impact assessment

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Finally, the impact on work performance item shows us a more balanced overview of the three tools. DALEX is shaped as the more useful tool in operations between the three, confirming the slight preference expressed from the users in the last understanding item.

4.4.2 Final Questionnaire

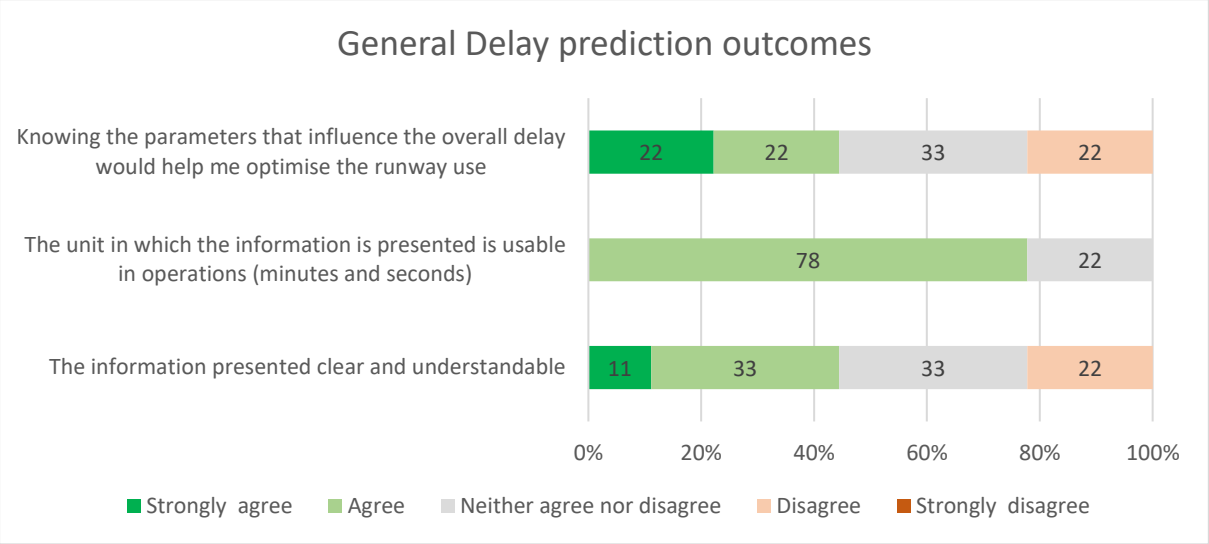


Figure 56: Overall Delay Prediction outcomes

From the final questionnaire we can notice that the two usability items ("The unit in which the information is presented is usable in operations"; "The information presented is clear and understandable") received general positive feedback, especially the first one which did not receive any negative rating.

On the other hand, the generic question about the impact on the work performance shows us that independently from the 3 different conditions, just the 22% of the participants reported negative feedback about the usefulness of a delay prediction tool in optimising the use of a runway.

4.4.3 Qualitative feedback

4.4.3.1 Open ended questions

The majority of ATCOs found the information presented clear and understandable, attributing it to the videos and narration which communicated the content intelligibly. However, although the concept was generally clear, some details and parameters remain unclear for some participants, introducing doubt in the tool's capability of calculating delays accurately. One ATCO even stated: "The basic ATC scenario presented is not supposed to create any delay for departing aircraft", showing that they would not have expected any delay for the scenario in question.

More than 20% of ATCOs did not agree with the level of clarity. Of these, some considered there was a poor relation between the explanatory video and the test itself, stating that the colours shown in the video did not match with those used in the simulation. Furthermore, the impact of the parameters on

EUROPEAN PARTNERSHIP
Co-funded by the European Union

the estimated delay and take-off time was not self-explanatory for all ATCOs. This introduced uncertainty in their capability of modifying these variables.

Aside from the purpose of the validation exercise, a final question was asked to the ATCOs, to better understand what other ATM task(s) would benefit from this information (i.e., knowledge of the parameters influencing the overall delay) in operations. In this way, one could extend the purpose of this tool to a larger range of ATM activities, including airport capacity optimisation.

Most ATCOs suggested that the information provided by the tool would allow to generate the best sequence of departures, thus optimising runway usage, increasing the airside capacity of the airport, and reducing runway occupation time. Of course, this would also require a strong coordination with the approach flow, to ensure the compliance between both arrival and departure flows. This coordination could also be used to predict potential conflicts between arrival and departure traffic. Finally, regarding human performance, this would increase the situational awareness of ATCOs.

Based on the results obtained from the runway scheduling optimisation, one could further use this information for airport strategic planning. For example, to schedule runway inspection and to predict aerodrome characteristics.

Finally, the delay prediction computed by the tool two could be useful for two more parameters, namely TSAT (Target Start-up Approval Time) and ATFCM delay. TSAT calculations could be optimised for more accurate predictions and consequently smoother departing traffic and reduced disruptions. Lastly, ATFCM could be improved using the optimised data from all previous flight phases (taxi, take-off, initial climb).

4.4.3.2 Final qualitative interview

From the qualitative assessment with the representant of Associazione Nazionale Assistenti e Controllori della Navigazione Aerea (Italian National Association of Air Traffic Controllers – ANACNA) inside the project's Advisory Board, overall positive feedback on the tool emerged. In particular, the interviewed ATCO thinks that the tool can be helpful for ground operations, which could include in the loop of decisions and awareness actors that are now excluded. This way operations could be optimised, without charging the Airlines with tasks of ground operations. There should be anyway some improvements to be performed on the tools. In particular, the DALEX condition lacks indications on what the AI is using to calculate the delay associated with each parameter, and it should be improved with techniques to reduce the variables involved in the selection of the parameters trying to develop a model to gather the most intercorrelated variables and present them as one. Finally, the main critics found by the ATCOs representant in ARTIMATION Advisory Board regard the fact that without a basic knowledge of Artificial Intelligence, it is hard to understand always what the tool is communicating to the end user.

4.5 Summary of the results and conclusions

Based on the algorithmic evaluation presented in 4.2, considering take-off time delay prediction on the datasets provided by EUROCONTROL, XGBoost using the MAE (in minutes), presented better MAE values for the ETFMS and GBDT are considered as reference from an experimentation performed by Dalmau et al [12]. For delay prediction, a comparison among the ML models shows XGBoost performed better than other models. From the algorithmic perspective, XGBoost is more scalable and better at

EUROPEAN PARTNERSHIP

Co-funded by the European Union

handling spare trees and optimizing errors than RF and GBDT. XGBoot is also a much faster algorithm for learning with large datasets compared to other ML methods.

Considering take-off time delay propagation, while comparing the three ML models, (i.e., RF, XGBoost and LSTM models), LSTM performs better than the other, although the overall accuracy is not so good. In the dataset, each aircraft mainly had three flights in a day; the delay propagation dataset indicates the delay propagated or not for a single aircraft for the last flight of each day. The data preparation procedure mentioned in 4.2.2 is followed since the propagation is time-dependent and a sequential problem, which has reduced the data size compared to the delay prediction dataset. For uniformity, the same three algorithms, i.e., RF, XGBoost and LSTM, are trained for delay propagation. The LSTM is better at solving sequence or temporal dependency; however, it requires many data than RF and XGBoost. The results may be due to insufficient samples in the dataset, and the sequences only depend on two previous flight information.

While comparing LIME and SHAP tools in terms of functional perspective, both are equally good to be applied for the XAI, however, SHAP present better results considering nDCG value. As mentioned previously, the raw dataset and attributes were for both delay prediction and delay propagation models, and LIME and SHAP provided the same visual explanations for both delay prediction and delay propagation models. Hence, from the algorithmic perspective, no information gain could be obtained validating explanations of delay propagation developed using LIME and SHAP.

While considering user evaluation to compare LIME, SHAP and DALEX, the feedback gathered from the DALEX self-report questionnaire shows a positive impact of the user-centred condition on both the understanding of the influence of the selected parameters and their contribution to the final delay.

In terms of usability and work performance, all the tools received positive feedback that the usefulness of a delay prediction tool in optimising the use of a runway.

According to the comments by the users in the open questions, the majority of ATCOs found the information presented clear and understandable, attributing it to the videos and narration which communicated the content intelligibly. However, level of clarity is an issue due to synchronization of video, text, colour, etc, and the impact of the parameters on the estimated delay and take-off time was not self-explanatory for all ATCOs.

Again, most ATCOs suggested that the information provided by the tool would allow to generate the best sequence of departures, thus optimising runway usage, increasing the airside capacity of the airport, and reducing runway occupation time.

In general, qualitative, and quantitative data show how the concept of the tool can be helpful in several ways, keeping in the loop actors nowadays excluded or included with different means of communication. Anyway, further research is needed to understand how to identify the most correct parameters to be shown event by event to the Air Traffic Controllers.

4.6 Lessons learned from the XAI applied to delay prediction

EUROPEAN PARTNERSHIP

Co-funded by the European Union

1. To gather better operational feedback from ATM experts entails that they have a minimum background knowledge on ML algorithms and AI, this way they contribute more actively.

2. A more detailed introduction on AI should be considered before user feedback collection.

3. The operational expert's involvement in validation activities should be supervised by AI experts.

4. The semantic involved in the display of ML algorithm parameters should correspond the ones used in operations.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

5 Conclusions and recommendations

Overall, all the hypothesis and objectives have been assessed in different ways: quantitative, with self-report ad-hoc questionnaires, qualitative, with debriefings and open questions, and with neurometrics (GSR; EEG). We can notice how for the Conflict Resolution use case significant differences between the experimental groups (students; experts) have been found: the students considered all the XAI solution more acceptable than the expert group. Within the experts, the black box condition was the most acceptable one. The same effect has been found in the acceptance construct, resulting in the student group having a significantly higher acceptance than the experts, and inside the expert group, considering the black box significantly the condition having a higher acceptance from the users. The student group had a significantly higher trust in the AI solution, and a significantly increased situational awareness than the experts as well. Finally, the students reported a significantly higher usability of the tools than the experts, and a significantly higher perceived work performance having the XAI tools available. This brought to a significantly higher perceived human performance of the students than the experts.

Differences between conditions have been found in the whole sample as well. In particular, the quantitative measures show how the black box resulted the most acceptable condition between the other explainability levels (Heat Map; Storyboard), even if the neurometrics assessed the Heat Map condition as the most accepted at an unaware level. The same result has been found on the acceptance construct: the black box is considered significantly the best condition in terms of acceptance in the whole sample.

Finally, a correlation between the acceptance and the human performance has been found, showing how the use of XAI assistants can improve the perceived human performance of the whole sample.

As for the delay prediction use case, the models have been built considering the work of Dalmau et al. [12] as a baseline, and the work focused on developing explanations of decisions made by the models. Here for better prediction results for delay prediction and propagation, incorporating domain knowledge and feature selection would be an important step in the ML model development pipeline, therefore further research on the topic is required. Furthermore, including contextual information such as regulations, weather information, and seasonal information, e.g., holidays, could improve the performance of the models.

The iterative process should be considered to develop a human-centric explanation using DALEX: different levels of users should be involved as the control group and test group for building the ML models and validating the explanations. To achieve high acceptance and trust of XAI, it needs to tune the models with user feedback iteratively through a lifelong learning procedure.

Finally, ARTIMATION tried to draw some recommendations as well, which will be deepened in Deliverable 7.1.

In particular, the parameters that are used to train the algorithms should be carefully selected: they can introduce biases in the AI tools and their proposals. An optimal proposal could in fact be considered less optimal, if applied in an operational environment. The selection of parameters for an optimal proposal should consider the context in which the resolution advisory should be applied. This requires further research to understand the impact of a shared situational awareness between the human and

EUROPEAN PARTNERSHIP

Co-funded by the European Union

the digital assistants, in which the system can understand the best parameters to consider for an optimal advisory.

Another consideration coming up from ARTIMATION results is a question: what is driving trust towards an AI resolution advisory? At the beginning of the project, we assumed trust was driven by transparency and explainability, but the assumption may have been partially wrong. In ARTIMATION we assessed trust only for the Conflict Resolution use case, being that a task which can have multiple good resolutions. At the same time, considering the AI outcome correct, we did not assess trust for the Delay Prediction use case, but delay predictions do not have multiple possible correct outcomes. Therefore, we think that further research is needed to understand correctly all the parameters driving trust towards AI outcomes, to provide the minimum required explainability without affecting the ATCOs' workload.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

6 References

1 Ahmed, M. U., Barua, S., Begum, S., Islam, M. R., & Weber, R. O. (2022). When a CBR in Hand is better than Twins in the Bush. Fourth Workshop on XCBR: Case-Based Reasoning for the Explanation of Intelligent Systems.

2 ARTIMATION. (n.d.). Consortium Agreement.

3 ARTIMATION. (n.d.). Grant Agreement.

4 ARTIMATION. (n.d.). Project Kick-off Meeting SJU presentation on Project Management Guidance.

5 ARTIMATION. (2021). Dissemination Plan.

6 ARTIMATION. (2021). Report on State of the Art - AI Support in ATM [Deliverable 3.1]. https://www.artimation.eu/wp-content/uploads/2022/03/D3.1-Report-on-State-of-Art-AI-support-in-ATM_v00.02.02.pdf

7 ARTIMATION. (2022). Report on lifelong ML framework with integration of causality [Deliverable 5.1].

8 ARTIMATION. (2022). Report on transparent AI models with explainability [Deliverable 4.2].

9 Ball, A. (2011, February 9). How to License Research Data | DCC. Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides/license-research-data

10 Biecek, P. (2018). DALEX: explainers for complex predictive models in R. The Journal of Machine Learning Research, 19(1), 3245-49.

11 Borghini, G., Di Flumeri, G., Aricò, P., Sciaraffa, N., Bonelli, S., Ragosta, M., Tomasello, P., Drogoul, F., Turhan, U., Acikel, B., Ozan, A., Imbert, J. P., Granger, G., Benhacene, R., & Babiloni, F. (2020). A multimodal and signals fusion approach for assessing the impact of stressful events on Air Traffic Controllers. Scientific Reports, 10(8600). https://doi.org/10.1038/s41598-020-65610-z

12 Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140.

13 Busa-Fekete, R., Szarvas, G., Elteto, T., & Kegl, B. (2012). An apple-to-apple´ comparison of learning-to-rank algorithms in terms of normalized discounted cumulative gain. ECAI 2012-20th European Conference on Artificial Intelligence: Preference Learning: Problems and Applications in AI Workshop, 242.

14 Dalmau, R., Ballerini, F., Naessens, H., Belkoura, S., & Wangnick, S. (2021). An explainable machine learning approach to improve take-off time predictions. Journal of Air Transport Management, 95.

15 Durand, N., & Gotteland, J. B. (2006). Genetic algorithm applied to air traffic management. Metaheuristics for hard optimization, 277-306.

16 European Commission. (2016, July 26). Guidelines on FAIR Data Management in Horizon 2020 (3.0). http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

17 European Commission. (2017, March 21). Guidelines to the rules on open access to scientific publications and open access to research data in Horizon 2020 (3.2). https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

18 European Commission, Executive Agency for Small and Medium-sized Enterprises, Haardt, J., Weiler, N., & Scherer, J. (2019). Making the Most of Your H2020 Project: Boosting the Impact of Your Project Through Effective Communication, Dissemination and Exploitation. Publications Office of the European Union.

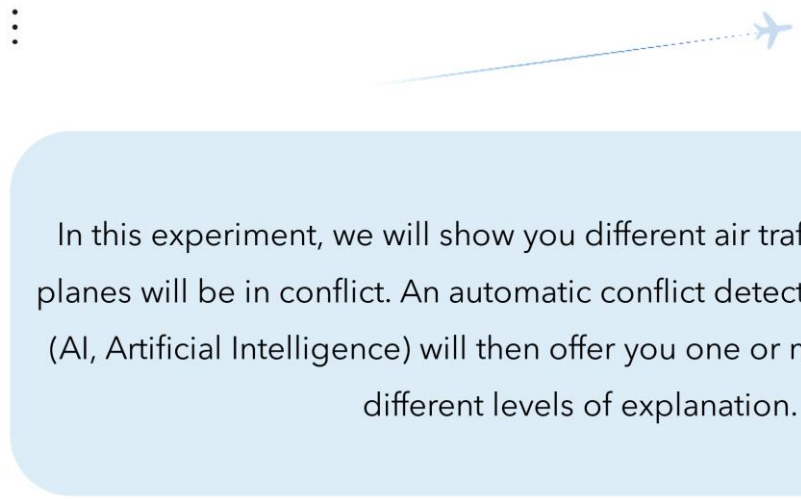
EUROPEAN PARTNERSHIP

Co-funded by the European Union

19 Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Comput, 9(8), 1735-80.
20 Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. Personality & Social Psychology Bulletin, 31(10), 1369-85. 10.1177/0146167205275613
21 Jackson, M. (2018, October 12). Checklist for a Software Management Plan (Version 0.2). https://www.software.ac.uk/software-management-plan
22 Jones, S. (2011, September 8). How to Develop a Data Management and Sharing Plan | DCC. Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides/develop-data-plan
23 Kirwan, B., Flynn, M., & Flynn, M. (2001). Identification of Air Traffic Controller Conflict Resolution Strategies for the CORA (Conflict Resolution Assistant) Project.
24 Kolodner, J. L. (1992). An introduction to case-based reasoning. Artificial Intelligence Review, 6(1), 3-34.
25 Li, P., Wu, Q., & Burges, C. (2007). Mcrank: Learning to rank using multiple classification and gradient boosting. Advances in neural information processing system, 20.
26 Open Data Research Pilot. (n.d.). How to select a repository?
27 Open Research Data Pilot. (n.d.). How to create a DMP plan. www.openaire.eu/opendatapilot-dmp
28 Open Research Data Pilot (ORDP). (n.d.). www.openaire.eu/opendatapilot
29 Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 1135-44.
30 Riche, N. H., Hurter, C., Carpendale, S., & Diakopoulos, N. (Eds.). (2018). Data-driven Storytelling. CRC Press.
31 Scheepens, R., Willems, N., Van de Wetering, H., Andrienko, G., Andrienko, N., & Van Wijk, J. J. (2011). Composite density maps for multivariate trajectories. IEEE Transactions on Visualization and Computer Graphics,17(12), 2518-2527.
32 SESAR JU. (n.d.). STELLAR Training on 20th January and 5th February.
33 SESAR JU. (2018, April 10). SESAR Human Performance Assessment Process V1 to V3- including VLDs (00.02.00 ed.).
34 SESAR JU. (2019, January 14). Communication Guidelines SESAR 2020 Projects (07.00.00).
35 SESAR JU. (2019, March). Project Handbook of SESAR 2020 Exploratory Research Call H2020-SESAR-2019-2 (ER4) (Programme Execution Guidance) (03.00.00 ed.).
36 Shapley, L. S. (1953). A Value for n-Person Games. In H. Kuhn & A. Tucker (Eds.), Contrinnutions to the Theory of Games II (pp. 307-317). Princeton University Press.
37 Strumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. The Journal of Machine Learning Research, 11, 1-18.
38 Wang, Y., Wang, L., Li, Y., He, D., Chen, W., & Liu, T. Y. (2013). A theoretical analysis of ndcg ranking measures. Proceedings of the 26th annual conference on learning theory, 8, 6.
39 Westin, C. (2017, November). Strategic Conformance: Exploring acceptance of individual-sensitive automation for air traffic control.
40 Westin, C., Borst, C., & Hilburn, B. (2016, February). Strategic Conformance: Overcoming Acceptance Issues of Decision Aiding Automation? IEEE Transactions on Human-Machine Systems, 46(1), 41-52. 10.1109/THMS.2015.2482480.
41 Whyte, A. (2014, October 31). Five steps to decide what data to keep | DCC. Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides/five-steps-decide-what-data-keep.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

7. Appendix A: Introduction Notice for CD&R



EUROPEAN PARTNERSHIP

Co-funded by the European Union

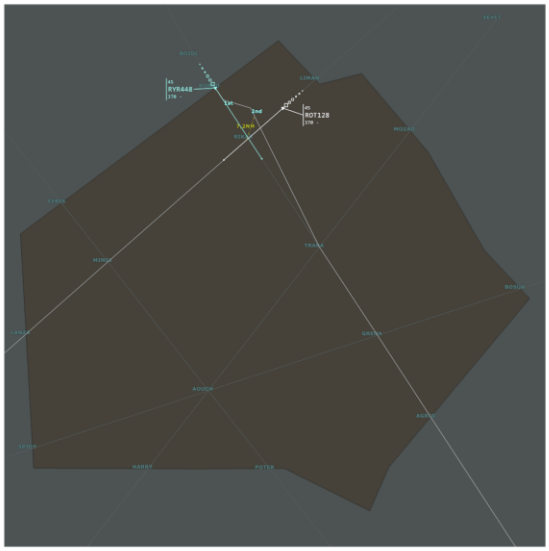
3 levels of explanation will be introduced to you

1. Black box
2. Heat map
3. Storyboard

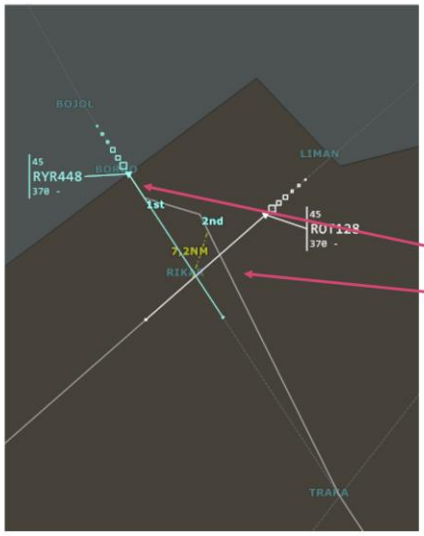
EUROPEAN PARTNERSHIP

Co-funded by the European Union

Black box: overview



Black box



In the **Blackbox** explanation, you will be presented with the trajectory of the aircraft and any trajectory changes.

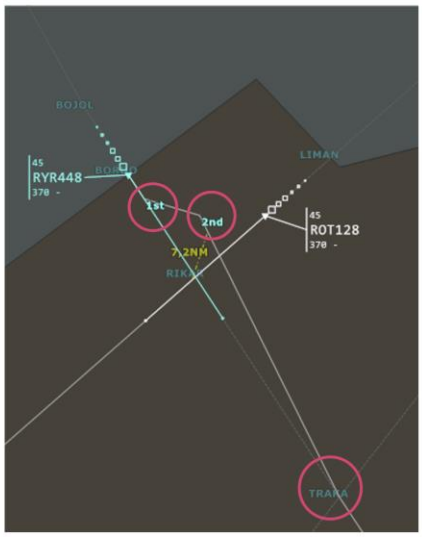
When the trajectory of an aircraft is modified, the initial trajectory of the aircraft appears **in color** (in this case, **cyan**) and the modified trajectory appears in white.

Each aircraft and its course changes have a unique associated color.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Black box

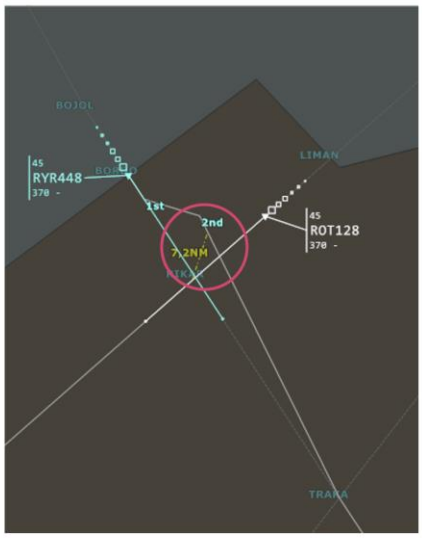


In this example, it is the trajectory of the RYR448 airplane that is modified. The trajectory of the ROT128 aircraft is not modified.

The RYR448 aircraft must first modify its trajectory at point 1 (**1st**) then at point 2 (**2nd**).

It continues its trajectory after reaching the **TRAKA** point.

Black box

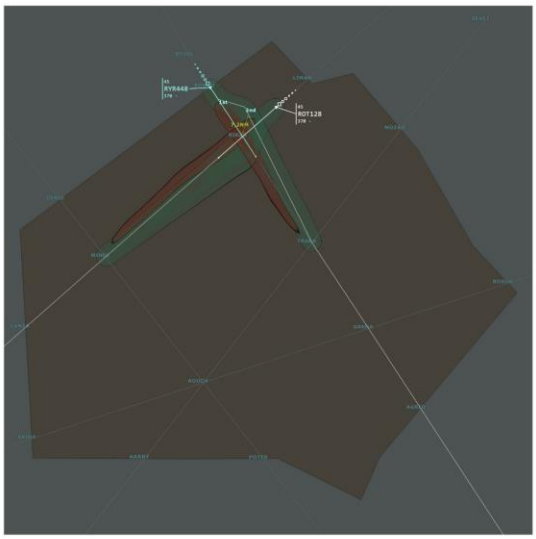


Finally, the minimum distance between the 2 planes will be shown in **yellow** (in this case, 7.2NM).

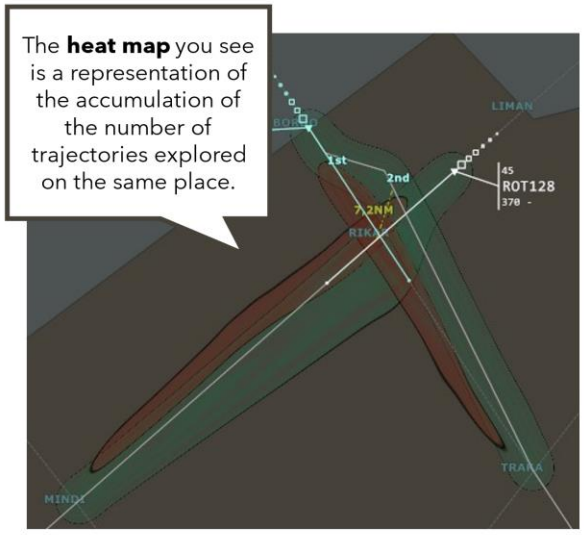
When the distance is less than 7NM, it is indicated in **orange**, and the associated trajectories is displayed in **yellow-green**:

Co-funded by the European Union

Heat map: overview



Heat map



The algorithm tests different trajectories and detects any associated conflicts.

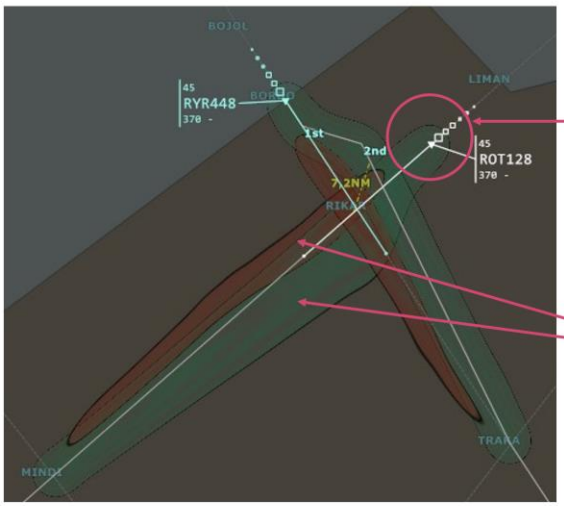
The **heat map** shows you for each aircraft:

- an envelope of "good modifications" of trajectory in green (> 7NM),
- an envelope of "bad modifications" of trajectory in red (< 7NM).

EUROPEAN PARTNERSHIP

Co-funded by the European Union

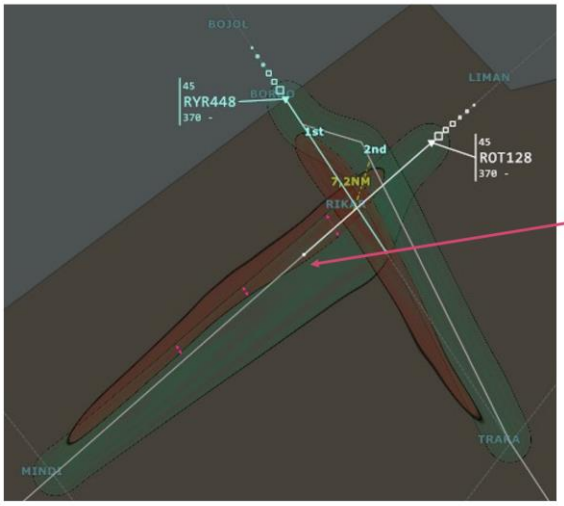
Heat map



To identify the plane corresponding to the **green envelope**, you can refer to the starting point of the zone and the current location of the plane.

The **red envelope** generally follows the path of the associated green envelope.

Heat map

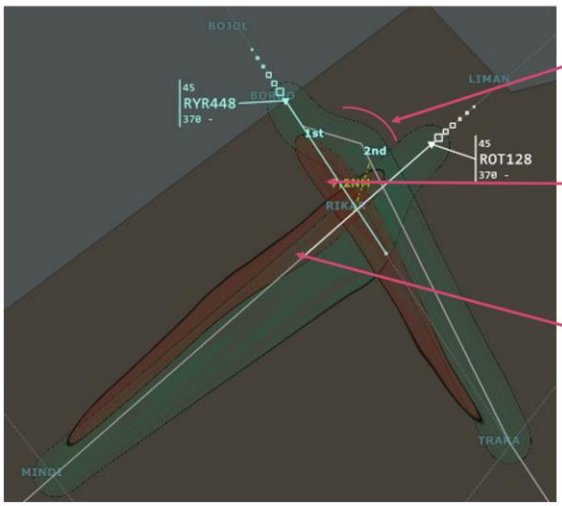


Overlap areas

When 2 zones (green/red) overlap, this means that the plane can pass through this zone (in this case, ROT128) **provided that** the second plane (RYR448) changes its trajectory sufficiently.

Co-funded by the European Union

Heat map

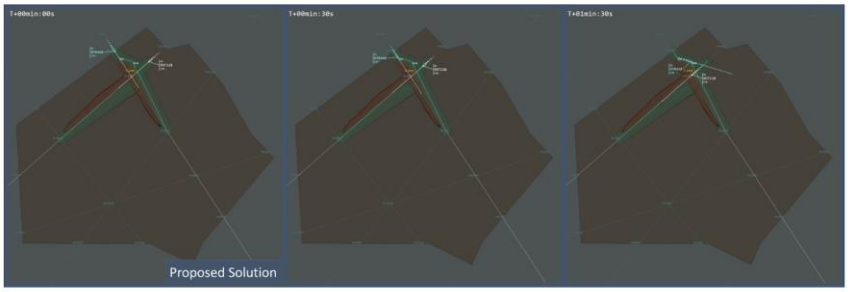


In this example, to avoid a conflict, the RYR448 aircraft must turn left…

… and it must not use this area.

ROT128 cannot use this overlap zone if RYR448 does not turn left.

Storyboard: overview → 3 parts will be presented to you

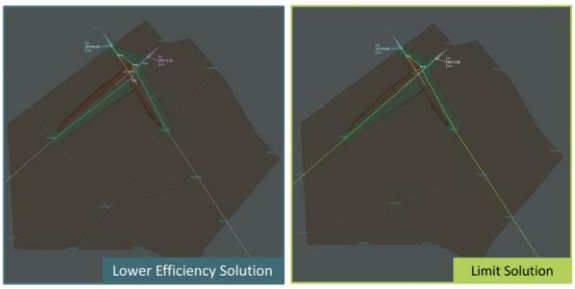


1

Timeline of the proposed solution

2

Lower Efficiency Solution



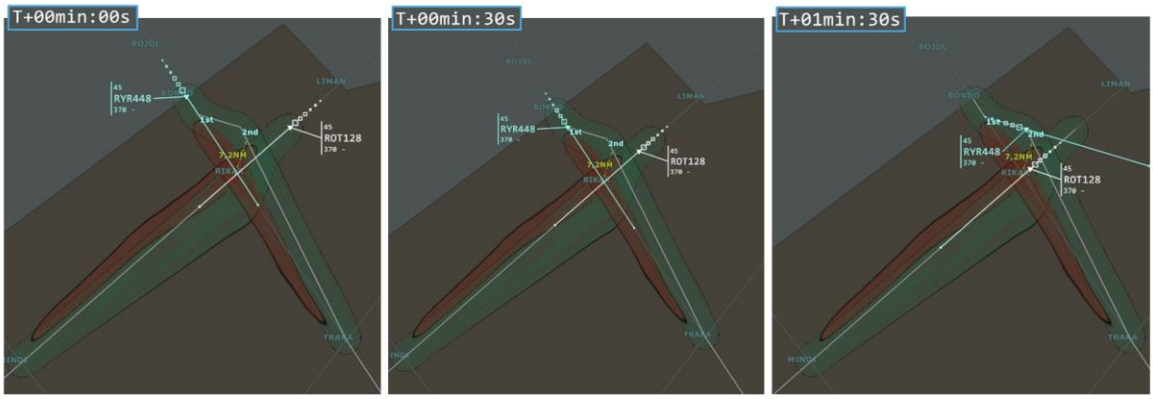
3

Limit Solution

EUROPEAN PARTNERSHIP

Co-funded by the European Union

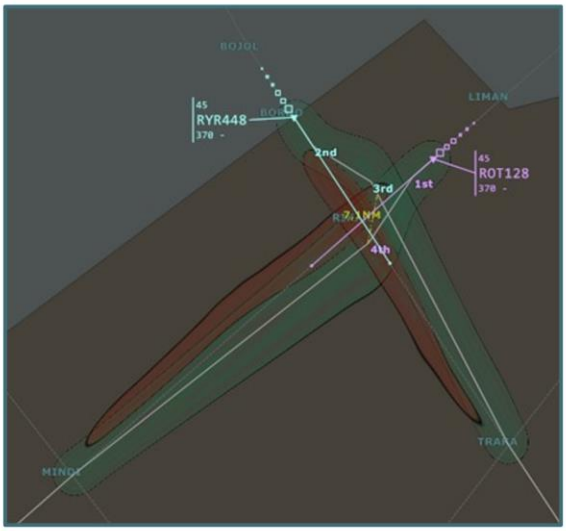
Storyboard: timeline



RYR448 is at BORDO | **RYR448 arrives at 1st** | **RYR448 arrives at 2nd**

The **timeline** presents the position of the planes over time in 3 to 4 thumbnails. Each thumbnail corresponds to a trajectory modification.

Storyboard: lower efficiency solution



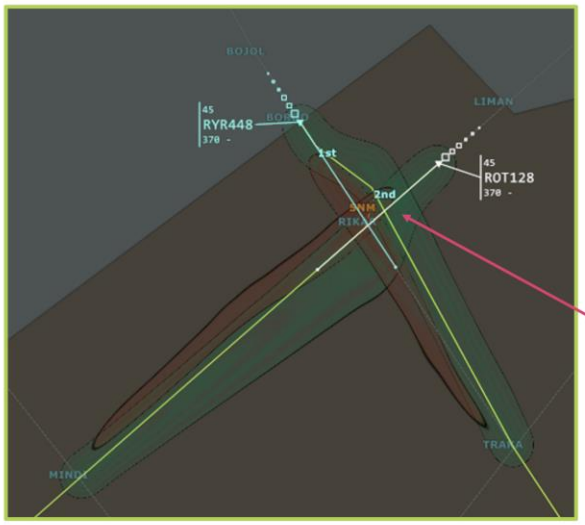
It is a less efficient solution according to the algorithm but still proposed since it can sometimes be useful to the air traffic controller.

The alternative solution is not always proposed by the algorithm.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Storyboard: limit solution



The limit case shows that the solution proposed by the algorithm is **robust**.

Indeed, if the given angle is smaller or if it is given later, the conflict will still be avoided.

The minimum distance is 5NM or more.

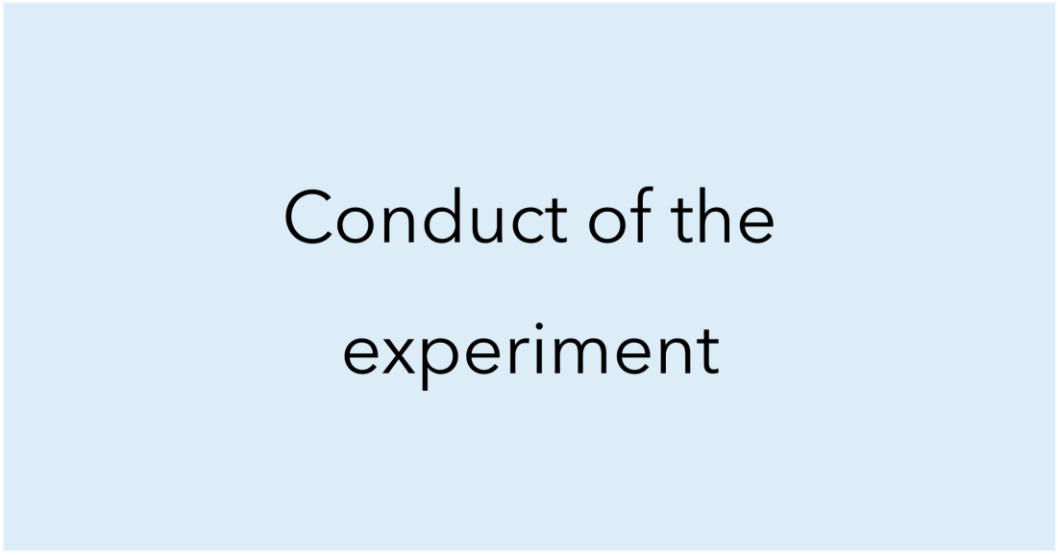
Thank you for your attention

& see you soon

EUROPEAN PARTNERSHIP

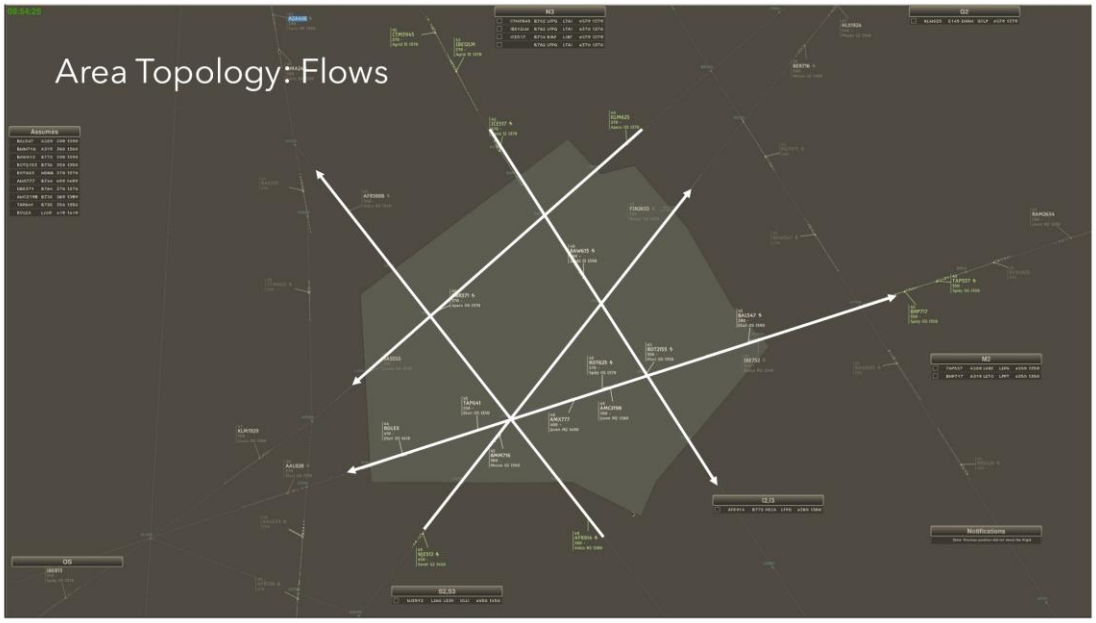
Co-funded by the European Union

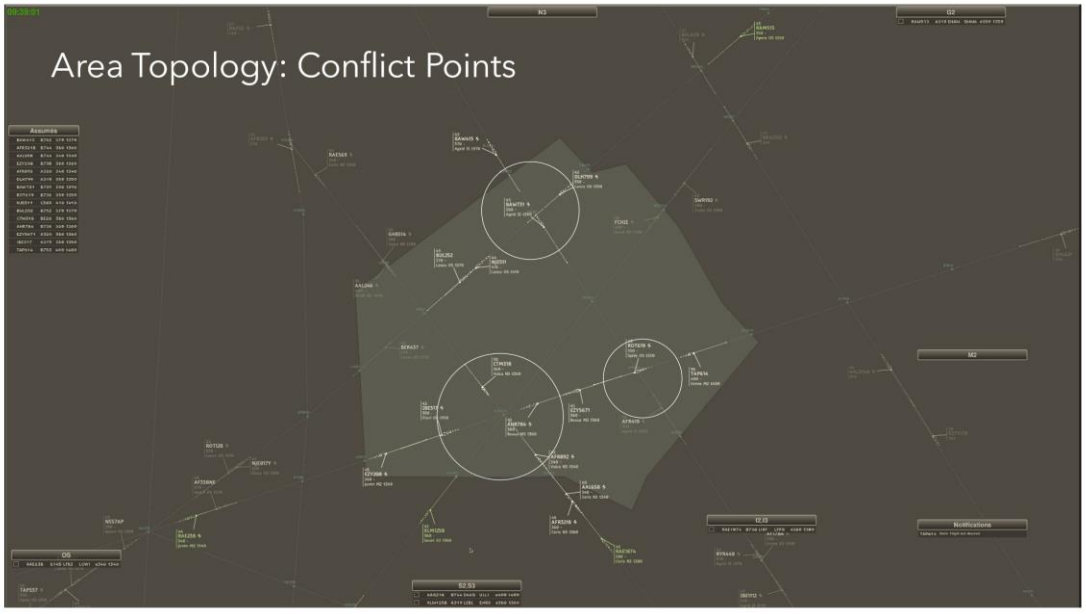
8. Appendix B: Additional Slide for CD&R on the Validation Day



You will have to resolve **conflicts** with the AI decision support. To do this, you will start by watching a video of the conflict. Then, the solution proposed by the AI, more or less explained, will be shown to you. After that, you will have to draw the solution you propose and answer a few questions.

EUROPEAN PARTNERSHIP
Co-funded by the European Union

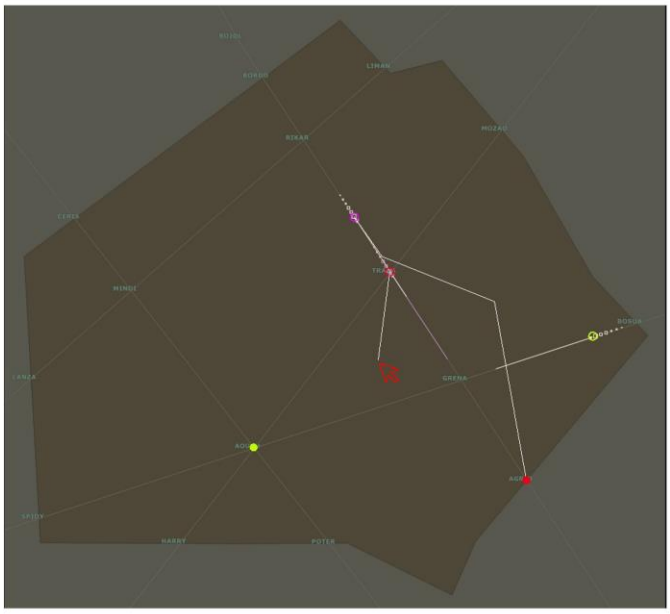




EUROPEAN PARTNERSHIP

Co-funded by the European Union

Solution drawing:



EUROPEAN PARTNERSHIP

Co-funded by the European Union

8. Appendix C. Questionnaires for Use Case 1 (Conflict Detection and Resolution)

Post Scenario Questionnaire:

I agree with the proposed solution:

Yes	No

The solution was easy to understand:

1	2	3	4	5
Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree

I understand why this solution has been generated:

Strongly disagree	1	2	3	4	5	Strongly agree

Post Condition Questionnaire:

Usability

Learn to operate the tool would be easy for me:

Strongly disagree	1	2	3	4	5	Strongly agree

I find the tool clear and understandable:

Strongly disagree	1	2	3	4	5	Strongly agree

I find the tool easy to use:

Strongly disagree	1	2	3	4	5	Strongly agree

Trust

I felt confident when using the tool:

Co-funded by the European Union

Strongly disagree	1	2	3	4	5	Strongly agree

Situational Awareness
The tool improved my Situation Awareness of the conflicts presented:

Strongly disagree	1	2	3	4	5	Strongly agree

Acceptability
I would ike to use this tool in the future:

Strongly disagree	1	2	3	4	5	Strongly agree

I like the new decision support interface:

Strongly disagree	1	2	3	4	5	Strongly agree

Impact on work performance
Using this tool in my job would allow me to solve conflicts faster:

Strongly disagree	1	2	3	4	5	Strongly agree

Using this tool in my job would increase my accuracy in solving conflicts (e.g., less mileage…):

Strongly disagree	1	2	3	4	5	Strongly agree

Using this tool would improve my work performance:

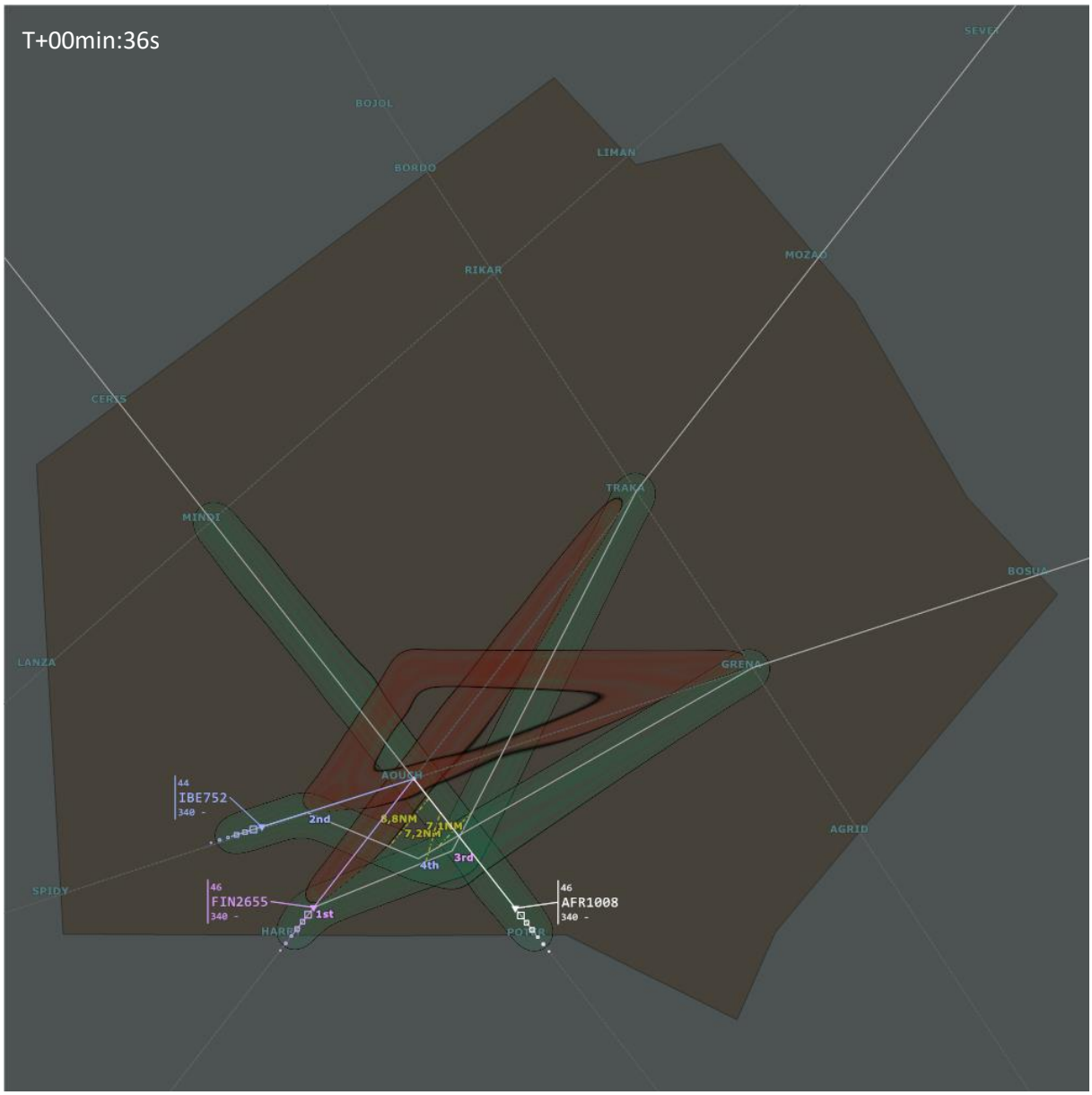
Strongly disagree	1	2	3	4	5	Strongly agree

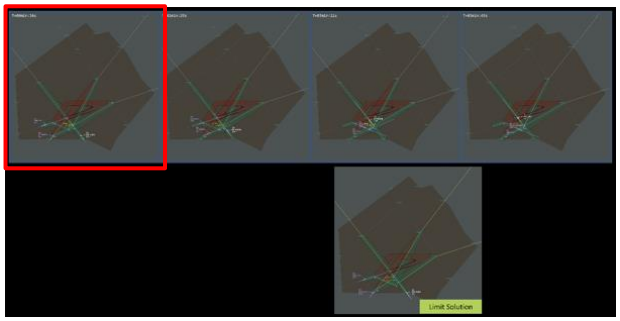
Using this tool would make my work easier:

Strongly disagree	1	2	3	4	5	Strongly agree

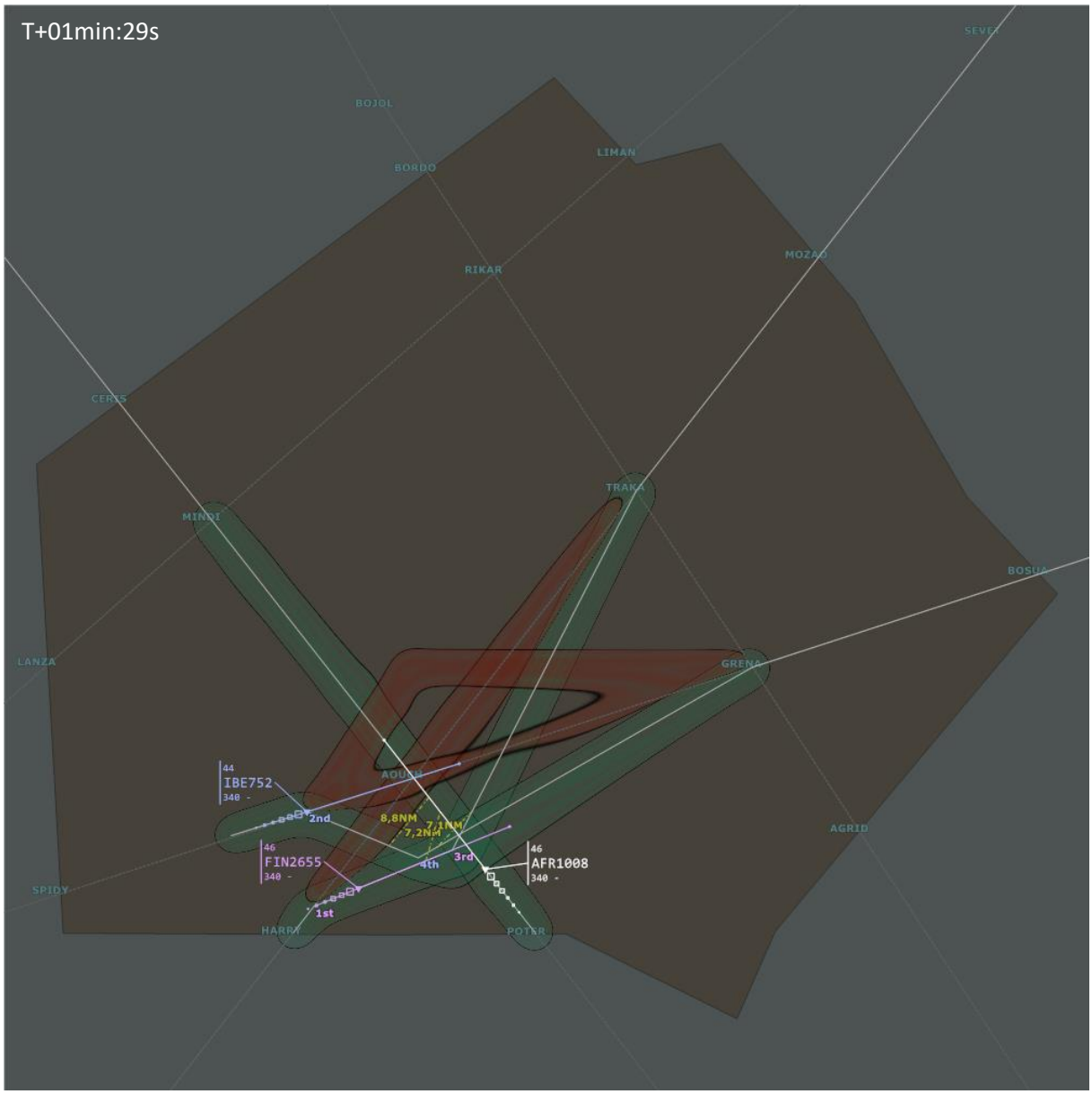
EUROPEAN PARTNERSHIP
Co-funded by the European Union

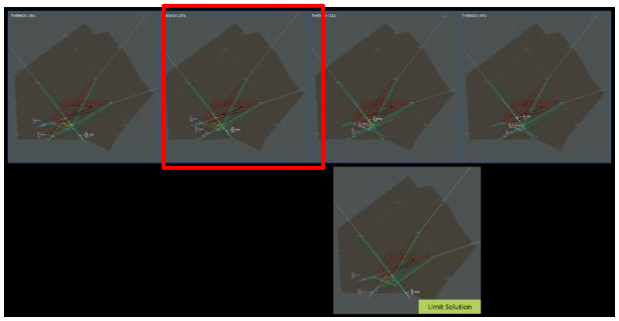
9. Appendix E. Storyboard detail



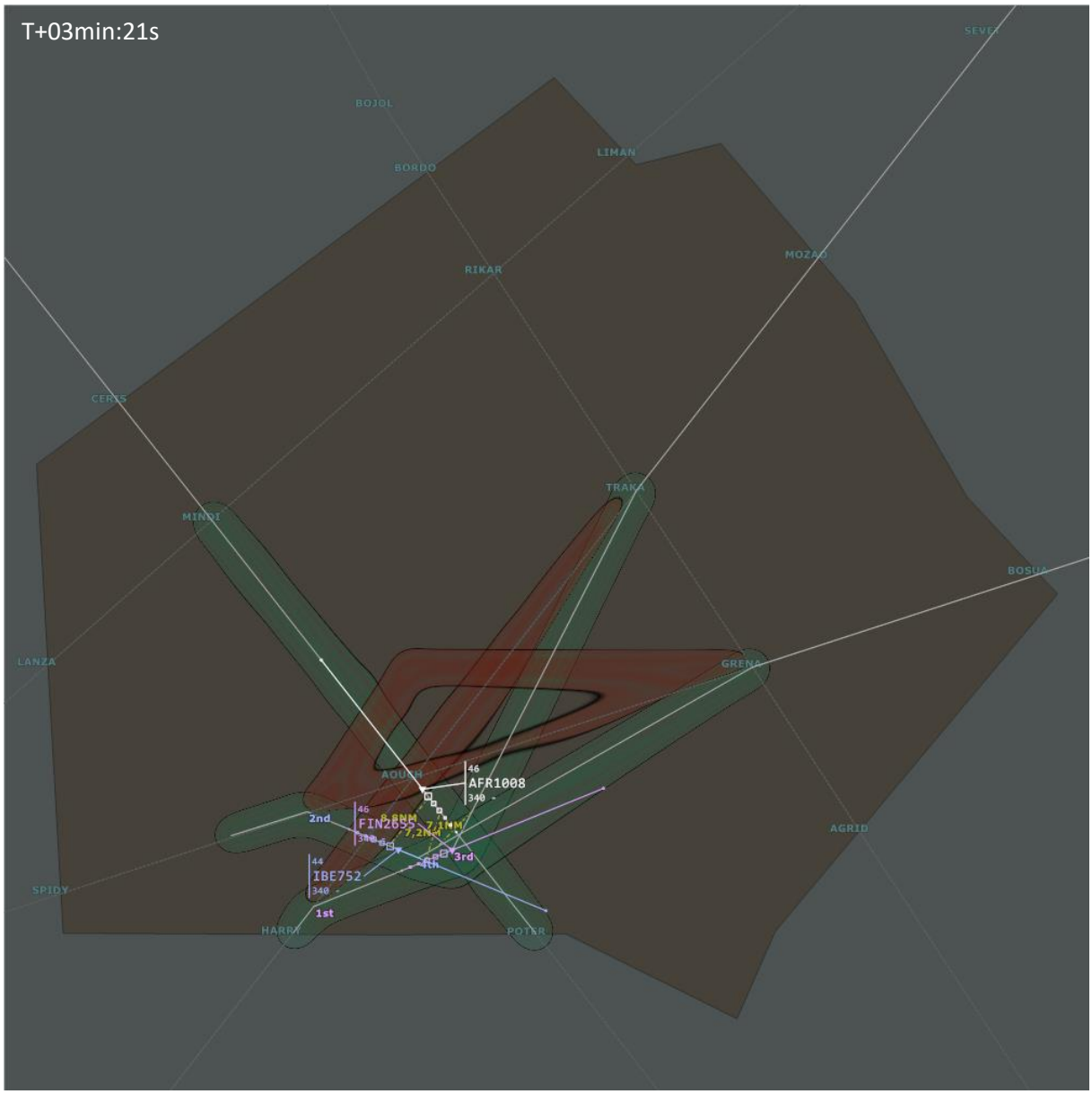


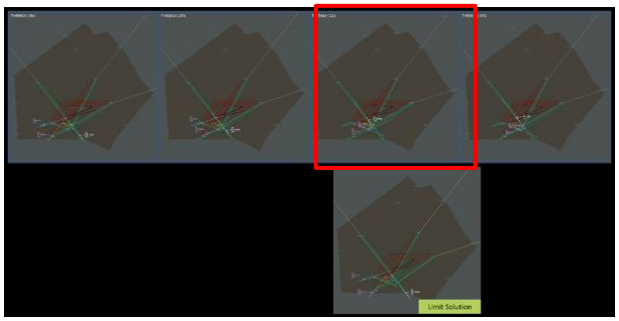
Co-funded by the European Union





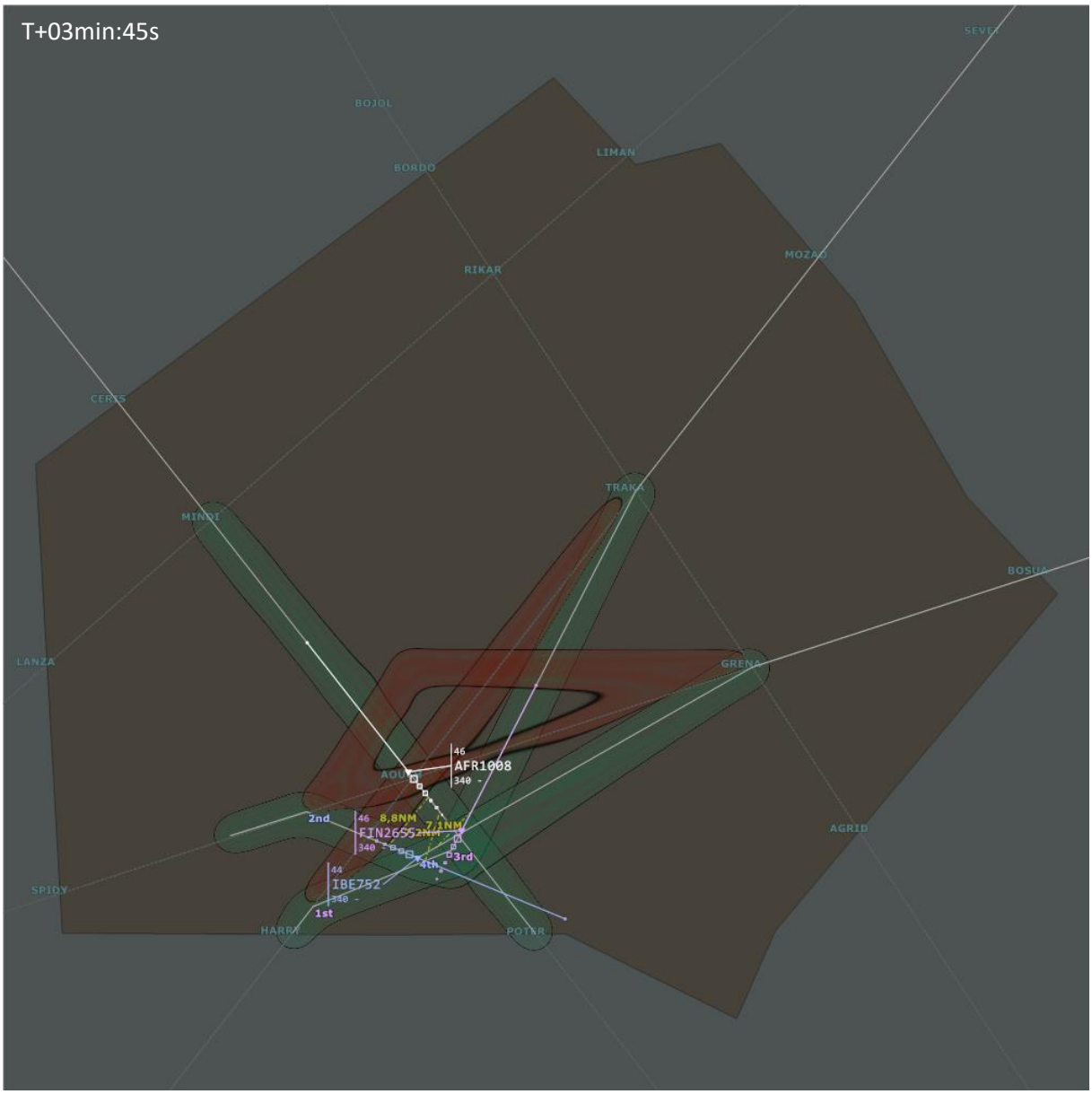
Co-funded by the European Union

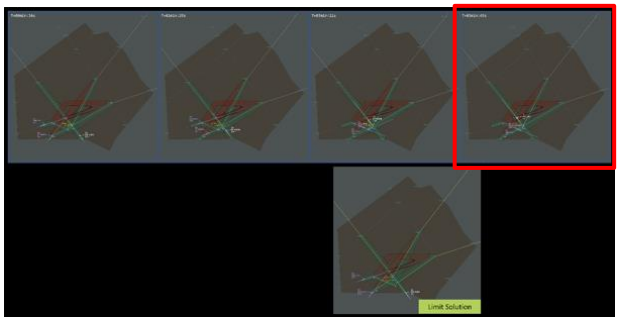




EUROPEAN PARTNERSHIP

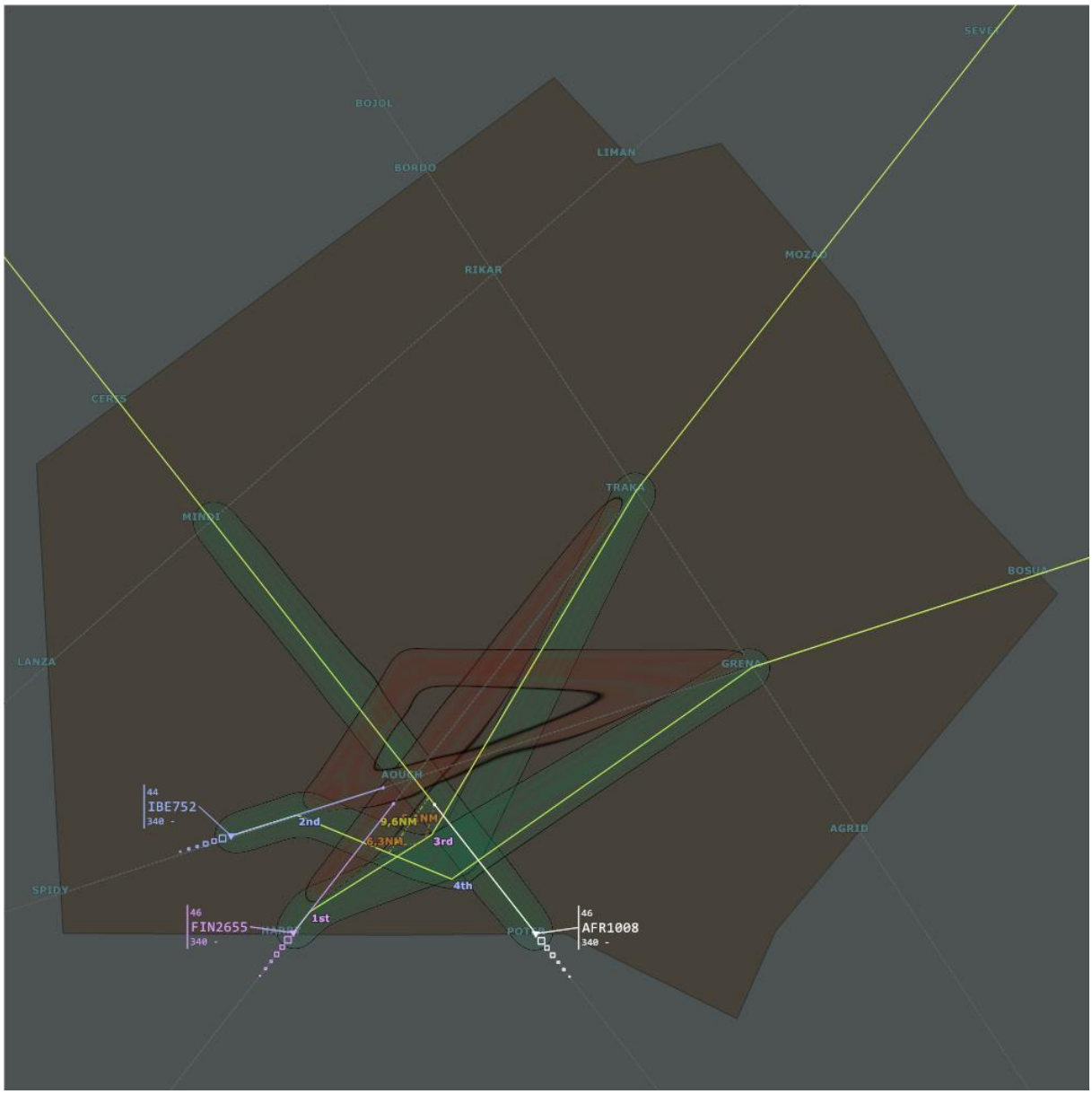
Co-funded by the European Union

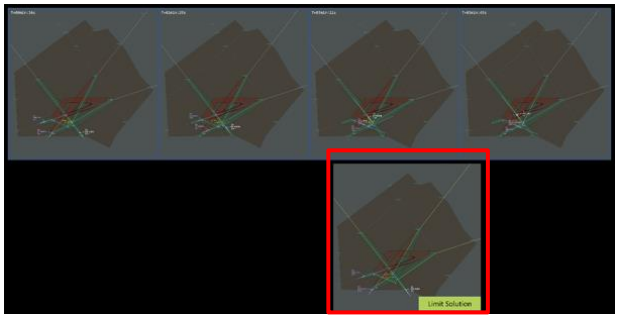




EUROPEAN PARTNERSHIP

Co-funded by the European Union





Co-funded by the European Union

10. Appendix D. Questionnaires for Use Case 2 (Delay Prediction)

Condition questionnaire (after each XAI method condition)

I understand why the end result (delay) is influenced by the selected parameters:

1	2	3	4	5
Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree

I understand how much each parameter influenced the end result (delay):

1	2	3	4	5
Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree

Usability

I find the information presented clear and understandable:

Strongly disagree	1	2	3	4	5	Strongly agree

Impact on work performance

Using this information would increase my accuracy in making an impact assessment:

Strongly disagree	1	2	3	4	5	Strongly agree

Knowing the parameters that influence the overall delay helps me optimise the runway use (o:

Strongly disagree	1	2	3	4	5	Strongly agree

The unit in which the information is presented is usable in operations (minutes):

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Strongly disagree	1	2	3	4	5	Strongly agree

Which kind of ATM task(s) do you think this information would benefit from this information in operations:

User-centred approach task (select the parameters that are more relevant to consider, that have more impact in the result)

Question about roles? might be too much for an online survey

Post-exercise questionnaire:

1. Usability

I find the information presented clear and understandable:

Strongly disagree	1	2	3	4	5	Strongly agree

Please justify why:
(Open answer box - mandatory)

2. Work performance

The unit in which the information is presented is usable in operations (minutes):

Strongly disagree	1	2	3	4	5	Strongly agree

3. Knowing the parameters that influence the overall delay helps me optimise the runway use

Strongly disagree	1	2	3	4	5	Strongly agree

4. Which kind of ATM task(s) do you think this information would benefit from this information in operations:
(open answer box - mandatory)

EUROPEAN PARTNERSHIP

Co-funded by the European Union