

Field Simulation Report

Deliverable ID:	D6.2
Dissemination Level:	Public
Project Acronym:	MAHALO
Grant:	892970
Call:	H2020-SESAR-2019-2
Topic:	SESAR-ER4-01-2019
Consortium Coordinator:	DBL
Edition date:	09 September 2022
Edition:	00.02.00
Template Edition:	02.00.05

Authoring & Approval

Authors of the document

Name / Beneficiary	Position / Title	Date
Brian Hilburn (CHPR)	WP2 leader	02/05/2022
Carl Westin (LiU)	WP6 leader	18/05/2022
Clark Borst (TU Delft)	WP4 leader	25/05/2022
Tiago Nunes (TU Delft)	WP5 Member	25/05/2022
Brian Hilburn (CHPR)	WP2 leader	02/09/2022

Reviewers internal to the project

Name / Beneficiary	Position / Title	Date
Brian Hilburn (CHPR)	WP2 leader	02/05/2022
Carl Westin (LiU)	WP6 leader	28/05/2022
Clark Borst (TU Delft)	WP4 leader	25/05/2022
Tiago Nunes (TU Delft)	WP5 Member	25/05/2022
Matteo Cocchioni (DBL)	Project Member	01/06/2022
Supathida Boonsong (LFV)	Project member	01/06/2022
Stefano Bonelli (DBL)	Project Coordinator	03/06/2022
Matteo Cocchioni (DBL)	Project Member	09/09/2022

Reviewers external to the project

Name / Beneficiary	Position / Title	Date
--------------------	------------------	------

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

Name / Beneficiary	Position / Title	Date
Stefano Bonelli (DBL)	Project Coordinator	06/06/2022
Stefano Bonelli (DBL)	Project Coordinator	09/09/2022

Rejected By – Representatives of beneficiaries involved in the project

Name and/or Beneficiary	Position / Title	Date
-------------------------	------------------	------

Document History

Edition	Date	Status	Name / Beneficiary	Justification
00.00.01	02/05/2022	Draft	Brian Hilburn	Draft Creation
00.00.02	28/05/2022	Draft	Carl Westin	Draft consolidation
00.00.03	25/05/2022	Draft	Clark Borst	Draft consolidation and revision
00.00.04	25/05/2022	Draft	Tiago Nunes	Draft consolidation and revision
00.00.05	01/06/2022	Draft	Matteo Cocchioni	Draft revision and consolidation
00.00.06	02/06/2022	Advanced Draft	Carl Westin	Advanced draft
00.00.07	03/06/2022	Final Draft	Brian Hilburn	Final Draft
00.01.00	06/06/2022	Final Release	Stefano Bonelli	Document Approved for Submission
00.01.01	02/09/2022	Final Draft	Brian Hilburn	Deliverable reopened for revision
00.02.00	09/09/2022	Final Release	Stefano Bonelli	Document Approved for Resubmission

Copyright Statement © 2022 – MAHALO Consortium. All rights reserved. Licensed to SESAR3 Joint Undertaking under conditions.

MAHALO

MODERN ATM VIA HUMAN / AUTOMATION LEARNING OPTIMISATION

This deliverable is part of a project that has received funding from the SESAR Joint Undertaking under grant agreement No 892970 under European Union's Horizon 2020 research and innovation programme.



Abstract

This document is the field simulation report, deliverable D6 .1 of the MAHALO project.

The MAHALO project set out to demonstrate a machine learning (ML) capability for en route air traffic control conflict detection and resolution (CD&R). Two constructs were hypothesized as critical to the interaction between aircraft controller and ML system: conformance was defined as the similarity between human and machine resolution strategy; transparency was defined as the degree to which the system made clear its underlying rationale. In WP6, MAHALO conducted field simulations to evaluate the impact of conformance and transparency manipulations on controller acceptance, agreement, workload, and general subjective feedback, among other measures.

MAHALO conducted two field simulations at two sites between December 2021 and April 2022. In total, 36 participants took part (final n=34 after data from two participants was discarded). Each simulation consisted of two phases. First was a training pre-test in which controllers interacted with scripted traffic scenarios that presented two-aircraft closing conflicts, and which recorded controllers' resolution strategies. Second was a main experiment phase, in which the same controllers interacted with ML solved analogues of the pre-test scenarios. ML solutions were developed during an interim training phase, in which several ML models were trained or synthetically generated to output conflict solution advisories.

For the main experiment phase at each site, conformance (3) and transparency (3) were manipulated within participant. Conformance was implemented as either a personal model, a group model, or an optimal model. ML was used to build the group and optimal models, whereas a synthetic approach was used to construct personal models for each participant. Transparency of proposed advisories was defined as either a baseline vector solution display, a prototype Situation Space Diagram (SSD) representation, or a text-based condition that combined SSD with a contextual explanation of the systems rationale (e.g. about target Closest Point of Approach, CPA).

Results showed that differences between simulation sites and test scenario sets can play a large role as extraneous variables. As a result, analysis set a shift from a pooled data approach to separate analyses by both simulation and scenario (with the inevitable challenges in data interpretation that this invites).

Having said that, several results achieved inferential statistical significance, and many more data trends were also suggested. Main effects of conformance were found on controller agreement ratings, with large differences between simulation sites. In terms of workload, a significant main effect of conformance was found for one of the simulation site / scenario combinations, in which the personal conformal model was associated with significantly lower workload than the optimal model.

Discussion centres on results of the field simulations, and some of the subtleties behind the methods and obtained results, as well as challenges to both conducting field simulation in ML, and in developing such systems generally.

Table of Contents

Abstract	4
1. Introduction.....	11
1.1 The MAHALO Project	11
1.1.1 WP6 Simulation Activities	11
1.1.2. Deliverables.....	12
1.2 Aims of the MAHALO field simulation	12
1.3 Research objective.....	12
2. Methods	14
2.1 Machine Learning models	14
2.1.1 SL model.....	15
2.1.2. RL model	16
2.2 Simulator design.....	16
2.3 Experimental Design	17
2.3.1 Independent variables	18
2.3.2 Dependent variables	19
2.3.3 Research questions and hypotheses.....	19
2.4 Materials.....	20
2.5 Procedures	20
2.5.1 Training scenario creation	20
2.5.2 Training pre-test procedures.....	22
2.5.3 ML training and test scenario creation	24
2.5.4 Main experiment procedures.....	31
2.5.5 Data analysis procedures	35
3. Methods	37
3.1. Conformance and Transparency effects.....	38
3.1.1 Acceptance of advisories.....	38
3.1.2 Agreement with advisories	42
3.1.3 Self-reported workload	46
3.1.4 Delta CPA.....	49
3.1.5 Survey results	50
3.2.4 Debrief Sessions: General Impressions	64
4. Discussion	66
4.1 Pooled data vs fine-grained analysis.....	66
4.2 Challenges in comparing personal and group models.....	66
4.3 The CD&R context	67
4.4 The benefits of transparency.....	67
4.5 On personalization and optimal systems	67
4.6 Challenges in training ML.....	68

4.7 Experimental control vs operational realism..... 68

4.8 Conformance and the potential importance of personalization..... 68

4.9 Noise in decision making..... 68

5. Conclusions..... 70

5.1 Addressing the research hypotheses 70

5.1.1 Relationship between conformance and acceptance / agreement 70

5.1.2 Relationship between transparency and acceptance / agreement 71

5.1.3 Relationship between workload and transparency / conformance 71

5.1.4 Interactive effects of Transparency and Conformance on acceptance and agreement..... 71

REFERENCES 73

ANNEXES 74

ANNEX A: TRAINING PRE-TEST MATERIALS..... 74

ANNEX B: MAIN EXPERIMENT MATERIALS..... 92

ANNEX C: TRAINING PRE-TEST RESULTS.....141

List of Tables

Table 1. Scenarios for the Training pre-test..... 22

Table 2. Parameters used for personal models..... 25

Table 3. Group and Optimal advisories in SIM2A and SIM2B 28

Table 4. Acceptance by Conformance and Transparency (absolute count), pooled data. 38

Table 5. Acceptance (fine-grained) by Conformance model, pooled data. 39

Table 6. Acceptance (fine-grained) by Transparency level, pooled data. 39

Table 7. Agreement ratings by transparency and conformance, pooled data. 43

Table 8. Controller agreement with advisories, SIM2A. 44

Table 9. Controller agreement with advisories, SIM2B..... 44

Table 10. Repeated measures ANOVA of normalised agreement 45

Table 11. Workload ratings by transparency and conformance, pooled data..... 47

Table 12. Workload ratings, SIM2A..... 48

Table 13. Workload ratings, SIM2B..... 48

Table 14. Repeated measures ANOVA of workload ratings..... 48

List of Figures

Figure 1. The MAHALO technical work package flow	11
Figure 2. An overview of the MAHALO hybrid ML system.....	14
Figure 3. Solution Space Diagram (SSD), cropped.....	15
Figure 4. Simulator UI (colours inverted for visibility).	17
Figure 5. Experimental matrix, Conformance (3) x Transparency (3).	18
Figure 6. Simulation airspace.	21
Figure 7. ML training- and test scenario creation.	24
Figure 8. SIM2A models.....	29
Figure 9. SIM2B models.....	29
Figure 10. Advisory conformance, SIM2A	30
Figure 11. Examples of test scenarios A and B, as used in the main experiment.	31
Figure 12. MAHALO transparency conditions.	32
Figure 13. Advisory/proposal interaction in Diagram and Text transparency conditions.	33
Figure 14. Proposal/advisory interaction.	33
Figure 15. Proposal rejection (only in SIM2B).	34
Figure 16. Scenario timing, main experiment.	34
Figure 17. Grouping of participants in SIM 2A based on individual separation margin preference in training pre-test.	37
Figure 18. Grouping of participants in SIM 2B based on individual separation margin preference in training pre-test.	38
Figure 19. Full acceptance, by Conformance and Transparency, pooled data.	39
Figure 20. Controller acceptance of advisories, SIM2A.	40
Figure 21. Controller acceptance of advisories, SIM2B.....	40
Figure 22. Acceptance of advisories in scenario A, SIM2A.....	41
Figure 23. Acceptance of advisories in scenario B, SIM2A.....	41
Figure 24. Acceptance of advisories in scenario A, SIM2B.....	42
Figure 25. Acceptance of advisories in scenario B, SIM2B.....	42

Figure 26. Agreement ratings by transparency and conformance, pooled data.	43
Figure 27. Controller agreement with advisories, SIM2A (left) and SIM 2B (right).	44
Figure 28. Agreement rating by separation margin, SIM2B Scenarios A (left) and B (right).	46
Figure 29. Agreement rating by separation margin, SIM2A Scenarios A (left) and B (right).	46
Figure 30. Normalised workload ratings by transparency and conformance, pooled data.....	47
Figure 31. Normalised workload ratings by transparency and conformance, SIM2A (left) and SIM2B (right).....	47
Figure 32. Workload by conformance and transparency, for SIM2B Scenario B.	49
Figure 33. Delta CPA by participant group, SIM2A.....	50
Figure 34. Delta CPA by participant group, SIM2B.....	50
Figure 35. Reported similarity with own solution strategy across conformance and transparency conditions in scenario A and Scenario B, SIM2A (n=18).....	51
Figure 36. Reported similarity with own solution strategy across conformance and transparency conditions in scenario A and Scenario B, SIM2B (n=14).....	51
Figure 37. Reported similarity with own solution strategy, SIM2A.	52
Figure 38. Reported similarity with own solution strategy, SIM2B.....	52
Figure 39. Reported understanding of advisory across conformance and transparency conditions in scenario A and Scenario B, SIM2A (n=18).	53
Figure 40. Reported understanding of advisory across conformance and transparency conditions in scenario A and Scenario B, SIM2B (n=14).	54
Figure 41. Reported understanding of solution, SIM2A.....	54
Figure 42. Reported understanding of solution, SIM2B.....	55
Figure 43. Post-session questionnaire item 1: Solution accuracy.....	55
Figure 44. Post-session questionnaire item 2: Solution safety.	56
Figure 45. Post-session questionnaire item 3: Solution efficiency.	56
Figure 46. Post-session questionnaire item 4: General agreement.....	57
Figure 47. Post-session questionnaire item 5: Solution difference.	57
Figure 48. Post-session questionnaire item 6: Solution superiority.	57
Figure 49. Post-session questionnaire item 7: Lower workload.	58
Figure 50. Post-session questionnaire item 8: Trust.....	58

Figure 51. Post-session questionnaire item 9: Solutions too early. 59

Figure 52. Post-session questionnaire item 10: Solutions too late. 59

Figure 53. Post-session questionnaire item 11: Quicker resolutions. 60

Figure 54. Post-session questionnaire item 12: Ease of use. 60

Figure 55. Post-session questionnaire item 13: Understandable format. 60

Figure 56. Exit questionnaire item 1: Accepted without agreeing. 61

Figure 57. Exit questionnaire item 2: Accepted without inspecting. 61

Figure 58. Exit questionnaire item 3: Computers will do more. 62

Figure 59. Exit questionnaire item 4: Computers will equal me. 62

Figure 60. Exit questionnaire item 5: Less rewarding job. 63

Figure 61. Exit questionnaire item 6: More than one solution. 63

Figure 62. Exit questionnaire item 7: Controllers’ acceptance. 64

1. Introduction¹

1.1 The MAHALO Project

The MAHALO project has had two high-level goals: First, to develop and demonstrate a hybrid machine learning capability for detecting and resolving en-route air traffic control conflicts; Second, to assess the impact of such a capability in terms of human performance. In particular, MAHALO focused on two constructs thought to underlie human-AI interaction. The first of these is *conformance*, which the project has defined as the apparent strategy match between human and AI systems. The second construct, *transparency*, refers to the degree to which the system makes its internal processes apparent to the operator. The MAHALO project set out to experimentally manipulate these two constructs, and to explore their main and interactive effects on a broad number of human performance measurements, including conflict detection performance, automation acceptance, rated workload, and others.

1.1.1 WP6 Simulation Activities

MAHALO Work Package 6 (WP6) focuses on designing and conducting the project field simulations. WP6 is the culmination of the project’s five technical work packages, from conceptual definition to UI development, to ML/E-UI integration, to experimental design, to conduct and analysis of field simulations.

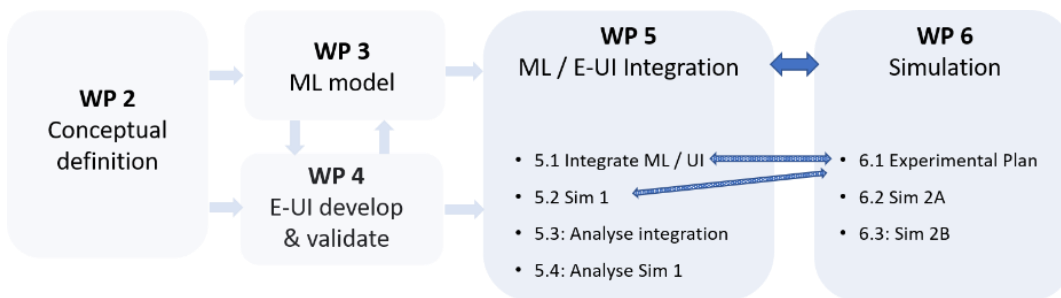


Figure 1. The MAHALO technical work package flow

¹ The opinions expressed herein reflect the authors’ views only. Under no circumstances shall the SESAR Joint Undertaking be responsible for any use that may be made of the information contained herein.

1.1.2. Deliverables

Over the project runtime, from WP2 through WP6, a series of deliverables was issued, in step with project technical progress. Earlier deliverables in this series described the staged process by which the research team:

- Conducted a state-of-the-art review (SOAR) of Machine Learning (ML) advances (D2.1);
- Developed and demonstrated a Machine Learning (ML) capability (D3.1; 3.2);
- Designed an experimental user interface and simulation capability (D4.1);
- Conducted human-in-the-loop validation trials of the user interface (D4.2);
- Integrated ML capabilities with the simulator and experimental interface (D5.1);
- Conducted a first full simulation to demonstrate the entire test platform (D5.2); and
- Specified experimental design for the final simulation sessions (Deliverable 6.1).

The culmination of this effort was the WP6 two-part field simulation, which consisted of Simulations 2A and 2B. The results of this field simulation are presented in this report.

1.2 Aims of the MAHALO field simulation

The aims of field simulation were twofold: first, as a demonstration of the final ML system using real-time controller-in-the-loop trials, and also to assess the impacts of manipulations of both the advisory system's solution *conformance*, and its *transparency*, and how these impacted controllers' objective performance and acceptance, as well as controllers' subjective feedback (e.g. workload ratings, agreement ratings, debrief feedback, etc.).

Two field simulations were conducted, one in Sweden using LFV controllers, and one conducted in Italy with ANACNA controllers. Hereafter, these two field simulations are referred to as Simulations 2A and 2B, respectively. Simulations 2A and 2B each consisted of two phases: a Training pre-test (also called Conformance pre-test), in which controllers interacted manually with en-route air traffic scenarios, and a Main experiment, in which controllers supervised (and, if desired, intervened in the behaviour of) a ML advisory system designed to resolve pending en-route traffic conflicts. The previously submitted experimental plan (MAHALO deliverable D6.1) followed the structure suggested in the SJU's *SESAR 2020 Experimental Approach Guidance ER* document, to present the research questions, testable hypotheses, independent- and dependent variables, data collection procedures, data analysis and security plans, and other experimental considerations associated with simulations 2A and 2B.

1.3 Research objective

The aim of WP6 field simulation activity was to empirically address the initial high-level issue posed by the MAHALO project, namely: how changes in the conformance and transparency of ATM CD&R automation (ML) might impact human / machine system performance. Or, stated as a research *question*:

How does the strategic conformance and transparency of a machine learning decision support system for conflict detection and resolution affect air traffic controllers' understanding, trust, acceptance, and workload of its advice and performance in solving conflicts, and how do these factors (conformance and transparency) interact?

This question was expanded into seven specific and testable research hypotheses, which made specific assumptions about the statistical main- and interaction effects expected in response to experimental manipulations of system conformance and transparency. These testable hypotheses are discussed in section 2.3 of this report.

The remainder of this document will now discuss the methods and results associated with the MAHALO WP6 field simulation.

2. Methods

2.1 Machine Learning models

MAHALO deliverables D3.1 (*Machine Learning Report*) and D3.2 (*Machine Learning Demonstrator*) together described the development, tuning, and demonstration of the MAHALO Machine Learning (ML) capability. Figure 2 shows a general overview of the MAHALO hybrid ML system, which consisted of two ML approaches (see also MAHALO D3.1):

- *Supervised Learning* (SL) that attempts to mimic human learning with a neural network architecture, and which learns the controller’s actions and enables presentation of personalized conflict resolution advisories, as well as resolution advisories based on the group average;
- *Reinforcement Learning* (RL) that aims to provide optimized conflict resolution advisories, according to a series of cost functions.

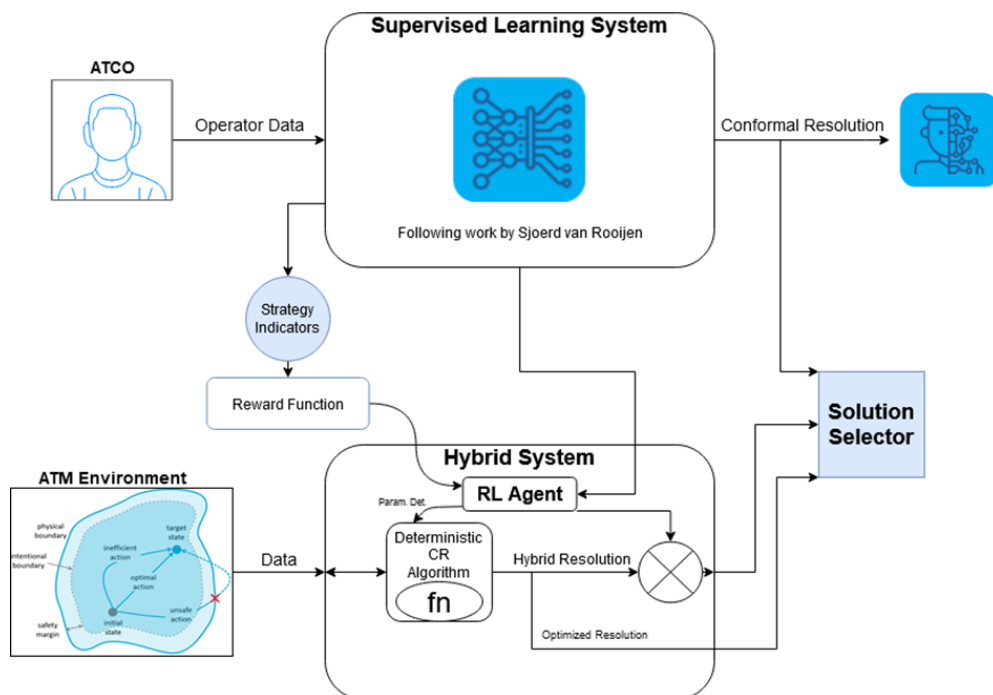


Figure 2. An overview of the MAHALO hybrid ML system.

The following sections briefly review the MAHALO SL and RL models underlying the MAHALO ML approach. Notice that SL and RL models were used in ways that enabled presentation of three different model conditions, as will be further discussed in section 3.3: a *Personalized* model, in which separate

SL models were trained on each given controller’s CD&R actions as collected during the preliminary phase of both SIM2A and SIM2B; a *Group* model, in which two SL models (one for SIM2A, and one for SIM2B) were trained, using the aggregated actions of controllers within each group; and an *Optimal* model, in which the RL model was trained using the same traffic scenarios used for the two field simulations. Notice that the optimal model did not rely on controller previous actions, but were designed to provide geometrically optimal solutions.

Given technical and time constraints, and also for reasons of experimental control, the ML models were not fully integrated in the simulation platform for real-time performance. Instead, the ML models were used in an off-line mode, and the resulting solutions were then presented to controllers. Notice that the ML models could not determine aircraft choice nor timing of the advisory, only resolution heading angle, so the research team crafted scenarios semi-manually, as discussed later.

2.1.1 SL model

Supervised learning (SL) relies on repeated exposure to training data samples that can be labelled in their output (typically in a binary way), which allows the SL agent to eventually classify new novel situations. The term “supervised” refers to the process of training the system, which is analogous to a student–teacher relationship. To take a very basic example, the teacher might present the student images of different animals and ask her to correctly identify the animal. The teacher provides feedback (was the answer correct or not?) after each trial. Over time, the student learns to recognize and classify different animals. In the real world, supervised learning is used for much more challenging applications, such as voice recognition and facial recognition.

In MAHALO, the developed SL model was based on earlier work [1] which learned to provide controllers with resolution advisories that matched their own preferred CD&R strategies.

In the case of MAHALO, the bridge to enable this was the graphical depiction of air traffic solution geometry, as presented in TU Delft’s prototype Solution Space Diagram (SSD). Because this graphical representation could be used both by the controller under manual conditions, and also processed as pixel data by a convolutional neural network (CNN) model, this provided a direct way for the SL agent to operate with the same data as the human air traffic controller.

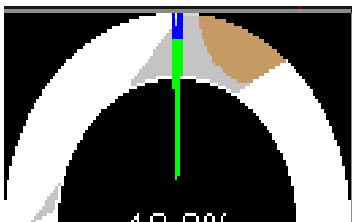


Figure 3. Solution Space Diagram (SSD), cropped.

The SL model used (128x64) pixel SSD graphical data as input, and the associated controller solution (for experimental reasons, this solution was limited to left turn, right turn, or zero heading change). The SL model output presented the direction and relative heading angle to resolve detected conflicts. As discussed later, data requirements (ML in general requires many training examples) limited the learning stability and performance of the SL model as trained for individuals (i.e., in the personal conformal condition), so for this condition the research team had to manually create solutions. Notice that for the group conformal condition, however, sufficient data were collected to allow adequate model performance, and the group condition therefore used SL output directly.

Further details of the procedures and specifications associated with the SL data generation, data inputs and outputs, parameter tuning, algorithm specification and training, can be found in MAHALO deliverable D3.1 (*Machine Learning Report*).

2.1.2. RL model

RL follows a rule-based approach, in which an algorithm arrives at a solution that maximizes reward, according to an optimization formula. The RL model used as input sector and traffic information (e.g. aircraft location, velocity, heading), as well as information on the given traffic advisory (pixel data).

MAHALO used two different RL approaches: Q Learning, and Deep Q Learning from Demonstration (DQFD). In the first instance DQFD was used. Because DQFD is more complicated to create and use, Q learning was retained as a fallback option. In the end, this fallback option was not required. Specifically, the project used Q-Learning coupled with the modified voltage potential (MVP) model of CD&R, which achieves separation assurance by representing other aircraft and destination as similar- and opposite potentials, respectively [2]. In the MAHALO RL model, the algorithm acted to optimize tuning of the MVP function, by continuously updating the parameters of MVP tuning.

Based on its reward function the RL agent iteratively evaluates performance and tunes the parameters. After running many training cycles, performance of the Q learning agent converges and stabilizes toward optimal performance. At that point, training is stopped and the agent can be integrated with the simulation. The RL model learned, based on the reward function, to choose solutions that maximized reward.

The DQFD approach, in general, presents a few challenges. The first is data requirements. Like other ML approaches, DQFD requires many training samples to converge on stabilized performance. Second, DQFD allows for many parameters (e.g., learning rate) to be set. All of these parameters have an influence on the eventual convergence and stabilization of the model, and the large number of associated degrees of freedom complicate convergence. Third, in the case of MAHALO, the team discovered during the project that the high dimensionality of the pixel data (captured in the SSD), complicated stabilisation of the RL solutions.

Further details of the RL model specification, development, data inputs and outputs, parameter tuning, and validation can be found in MAHALO deliverable D3.1 (*Machine Learning Report*).

2.2 Simulator design

The platform used for conducting field simulations was based on TU Delft's SectorX, a Java-based, medium-fidelity ATC research simulator, designed to run on portable PCs, and originally intended for conducting HITL experiments. For this purpose, SectorX was already a highly capable simulator, with flight dynamics conforming to BADA flight performance models. For purposes of the MAHALO field simulations, several SectorX capabilities were either extended or newly developed, to enable either manual control (in the training pre-test phase of each field simulation), or scripted automatic control (during the main experiment phase of each field simulation).

As configured for MAHALO field simulations, SectorX incorporated STCA, MTCD, and VERA separation calculation capabilities². As elaborated in MAHALO deliverable D4.2 (E-UI Validation Report), the following (manual / automatic) simulation capabilities also had to be validated prior to field simulations:

- Assuming, transferring and clearing aircraft to destination (exit waypoint);
- Clearing aircraft to target altitudes;
- Issuing heading and altitude clearances;
- Incorporating text-based explanations (to enable the experimental transparency manipulation, as discussed later);
- Display of resolution advisories as generated by the ML models;
- Dependent measure data outputs relating to time, control inputs, traffic state, pixel data, aircraft track and altitude deviation;
- Presentation and recording of controller post-solution (agreement) and post-scenario (workload) ratings.

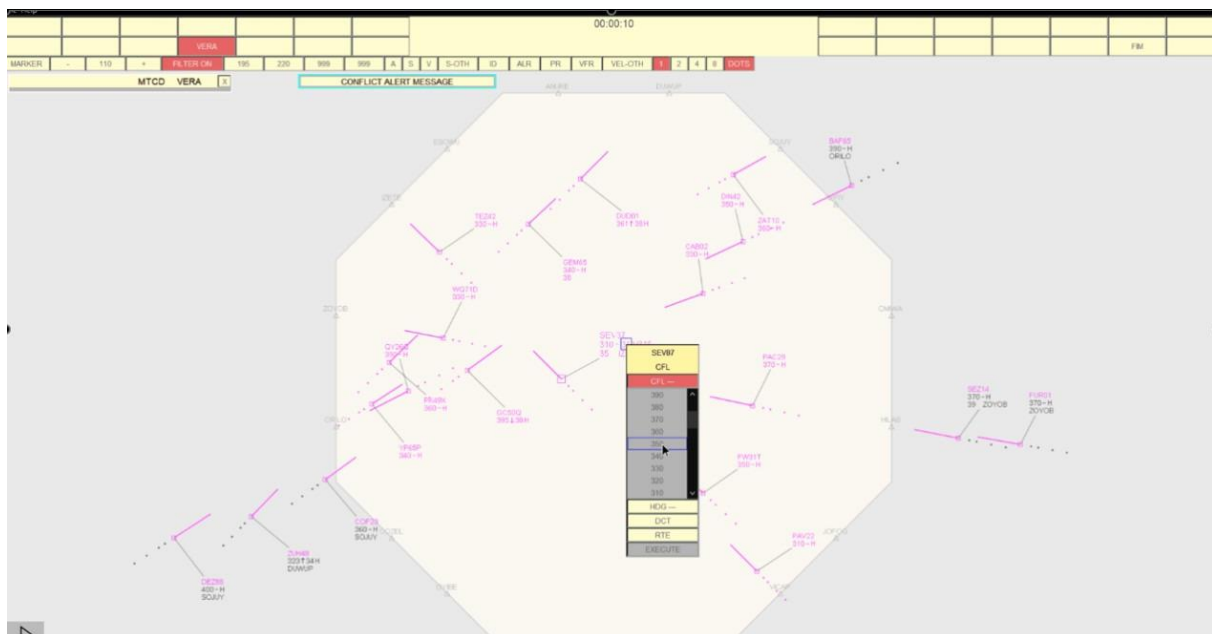


Figure 4. Simulator UI (colours inverted for visibility).

2.3 Experimental Design

As discussed in sections 2.5.2 and 2.5.4, field simulations were conducted at two different sites and were called SIM2A and SIM2B. Each of the simulations consisted of a training pre-test, and main experiment phase. As detailed in the earlier deliverable D6.1 (*Experimental Design*), a “yoked” within-participants design was used in which the same controllers took part in the initial (i.e. training pre-test) and follow-on (main experiment) phase of the field simulation.

² although neither simulation site uses VERA per se, Simulation 1 testing confirmed that controllers accustomed to e.g. the TopSky SEP tool were comfortable with the separation tool as simulated.

Simulator- and experimental design started from the following high-level operational assumptions:

- The controller and AI will work cooperatively in CD&R;
- The controller will work in single person mode;
- The future environment will be based on 4DTM, where the majority of conflicts are solved strategically;
- Traffic levels and sector sizes will be greater than today;
- The controller will maintain authority for final implementation of solutions;
- Air-Ground communication will be via CPDLC datalink, not voice.

As discussed in section 2.5.1, both the training pre-test and main experiment phases used the same en route airspace, and basic simulation procedures. The core difference between the two phases was the level of active control: In the former, controllers acted upon detected conflicts to devise and implement a solution. In the latter, controllers supervised the simulated automation, and only responded to system advisories by either accepting or modifying advisories.

The aim of the training pre-test was only to collect controllers’ behaviours during presented en route conflicts. Controller solutions, in terms of aircraft choice, solution type (heading versus altitude), solution value (degrees of heading change), and response time were all collected for reasons of training the ML. As such, there were no experimental manipulations during the training pre-test phase.

2.3.1 Independent variables

The main experiment phase experimentally manipulated both transparency and conformance as within-participant independent variables. Transparency was varied between the three sessions of the main experiment, whereas conformance was varied within each session. Transparency was defined as either vector, diagram, or text (see also section 2.5.4). Conformance was defined as one of three levels: personal, group, or optimal. As shown in figure 5 the resulting 3 x 3 matrix of independent variable levels represented nine experimental conditions. For each participant the order of these conditions was based on a Latin square procedure, that ensures that experimental condition appears only once in each row and each column.

		Transparency		
		Vector	Diagram	Diagram &Text
Conformance	Personal			
	Group			
	Optimal			

Figure 5. Experimental matrix, Conformance (3) x Transparency (3).

Conformance is a multi-dimensional concept in the case of CD&R. Clearly a heading and an altitude solution are different and thus “nonconformal.” But how conformal is a 10 degree versus 15 degree heading change? Or turning an aircraft right as opposed to left? Whether a solution conforms to a controller’s strategy will depend on, for example:

- Response time / latency of conflict detection and resolution
- Aircraft choice

- Resolution type (heading, altitude, speed)
- Resolution direction (e.g. right versus left?)
- Resolution value (e.g., 10 degrees? 15 degrees?)
- Spatial relationship (e.g., turn A behind B, or B behind A?)
- Separation margin (e.g. 6 nm or 10 nm?)

See section 2.5.3 for details on how conformance was defined in creating the personal test scenarios.

2.3.2 Dependent variables

Field simulation sessions collected two main types of data, objective performance data as recorded by SectorX, and subjective feedback data as provided during questionnaires and on-screen ratings. Objective data consisted of both event- and state data, including simulator air traffic events and also controller/ML inputs. The following dependent variables were defined:

- Acceptance – whether the controller chooses to implement a given solution. After preliminary analysis this was converted from a binary to a five point measure, to permit finer-grained analysis (see also section 2.5.5);
- Response time – from onset of resolution advisory to response (execute button pressed) in seconds (s)
- Agreement – the self-reported extent to which the controller agreed with the solution. Was presented via on-screen prompt, with a 0-100 scale;
- Workload rating – also self-reported on a 0-100 scale;
- Understanding advisory – self reported on a 1– 6 point Likert scale (1 = disagree highly, 6 = agree highly).
- Similarity of advisory with own solution strategy – self reported on a 1– 6 point Likert scale (1 = disagree highly, 6 = agree highly).
- Delta closest point of approach (CPA) – difference in nautical miles (nm) between proposed CPA of advisory and achieved CPA in solution (as modified by participant)

Other questionnaire items, mainly during the main experiment phase, collected information on controller attitudes toward automation, control strategies, and trust in automated systems. These test materials can be found in Annexes A and B for the training pre-test and main experiments, respectively.

2.3.3 Research questions and hypotheses

Field simulation addressed several broad research questions, each stated as a testable hypothesis regarding main- and interaction effects, as follows:

Hypothesized main effects³

- Hypothesis 1: controller acceptance of, and agreement with advisories will be higher if those advisories are based on solutions that conform to the controller’s preferred solution (personal models as derived from their performance in the training pre-test);

³ It was also hypothesized that trust would be impacted by experimental manipulations of conformance and transparency, but the final experimental design did not allow a fine-grained analysis of this effect. Questionnaire response regarding trust were, however, collected.

- Hypothesis 2: controller acceptance of advisories will be higher if those advisories are presented in a high transparency display format;
- Hypothesis 3: Both transparency and conformance manipulations will be associated with a change in reported workload.

Hypothesized interaction effects

- Hypothesis 4a: Under low transparency, personal conformal advisories will be more accepted/agreed upon than will optimized advisories;
- Hypothesis 4b: Under high transparency, this trend will be less pronounced, and the difference in acceptance/agreement between personal conformal and optimal advisories will be smaller.

2.4 Materials

SectorX was run on a Windows laptop connected to an external 28” display with a resolution of 1920 x 1080. Participants interacted with the simulation via mouse and keyboard. An auxiliary laptop was used to collect post-solution agreement ratings on conformance (agreement with own solution strategy) and understanding items.

A *Tobii Pro Glasses II* eye tracker was used during training SIM2B pre-test and main experiment sessions⁴. The Tobii uses an eyeglass-mounted system and the infrared corneal reflection technique, to record eye point of gaze (EPOG) at 50Hz. Post processing of EPOG data enables calculation of such additional measures as fixation frequency and duration, blink rate, and pupil diameter. The system also provides a red EPOG indicator on the recording video, which allows for realtime experimenter monitoring of EPOG and signal quality. Eye tracking data have not yet been analyzed, and eye tracking results are not presented in this report.

Briefing guides, training scripts, and survey instruments were produced in advance of field simulations. These can be found in Annexes A (for the training pre-test) and B (for the main experiment).

2.5 Procedures

Insights gained from field SIM2A led to a few minor tweaks and refinements in the later SIM2B procedures. The following Procedures sections 2.5.1 through 2.5.5 will discuss the common methods across the two simulations. Where relevant, differences between 2A and 2B procedures will be noted.

2.5.1 Training scenario creation

MAHALO used two broad categories of scenarios as follows:

- Training scenarios—were those used in the training pre-test. Six traffic scenarios were used. These scenarios were modeled on current day ATM and CD&R procedures, and controllers were required to resolve conflicts manually. The training scenarios included no manipulation of conformance nor transparency.

⁴ . For logistical reasons the eye tracker was unavailable for SIM2A.

- Test scenarios-- were those used in the main experiment. Two scenarios were used, selected out of the six traffic scenarios in the training pre-test. These scenarios were adapted during the interim training phase to incorporate automated CD&R solutions and support supervisory control procedures. However, the two scenarios were identical to the ones used in the training pre-test in terms of airspace, traffic, and conflicts.

All traffic scenarios were based on a 100 x 100 nm generic enroute air traffic sector, in an approximately octagonal shape (see figure 6) that allowed creation of unrecognisable scenario variants (e.g., via mirroring and rotation). Traffic flows were free-route, with a few prominent flow directions. Enroute altitudes were within reduced vertical separation minimum (RVSM) airspace, FL290 to FL410. For reasons of experimental control (i.e., not wanting to confound certain traffic closure geometries), the usual semi-circular rule (in which cruise altitude depends on heading) was disregarded.

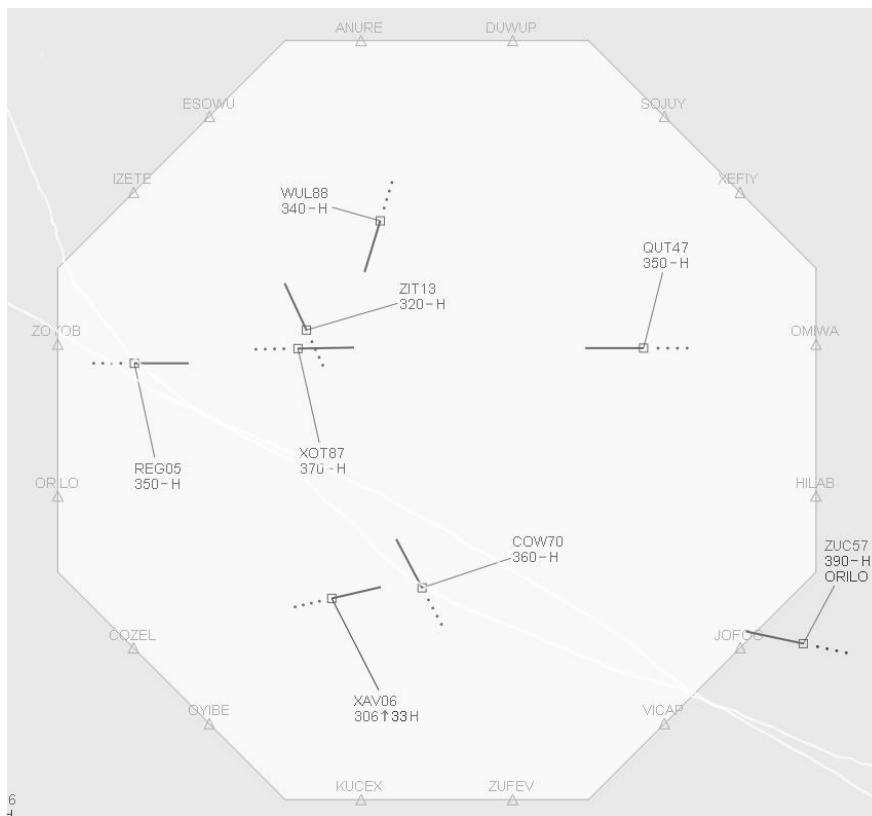


Figure 6. Simulation airspace.

For the training pre-test, six base scenarios were manually created with a maximum sector occupancy of 22 aircraft. These are shown in Table 1. Each of these base scenarios was transformed via mirroring / rotating traffic pattern, adjusting FL, and changing waypoint names and call signs, into six scenario variants, yielding an initial set of 36 total traffic scenarios for the training pre-test. These variants used rotation angles of -10, 0, 10, 20, 30, and 40 degrees from the base scenario.

Table 1. Scenarios for the Training pre-test.

Scenario	Name	Conflict type	Conflict angle (deg)	Closest point of approach	Time to CPA (s)	Callsign A/C A				Callsign A/C B				FL	A/C in scenario	Max A/C in sector	Min A/C in sector	Sector size (nm)
						A/C	Type	Speed	HDG	A/C	type 2	Speed	HDG					
1	S1_1	Crossing	090	0	327	PAJ61	A321	245	134	XT52Q	B738	245	224	370	19	17	13	100x100
2	S2_1	Crossing	068	0	310	DIN42	B788	262	245	FW31T	A346	262	313	350	22	17	14	100x100
3	S3_1	Crossing	134	0	255	JX21W	B737	255	215	NN17Z	B738	255	349	380	20	16	14	100x100
4	S4_1	Crossing	102	0	270	YAB29	A321	260	316	GL08G	B738	260	214	310	27	18	14	100x100
5	S5_1	Crossing	088	0	250	VP79D	B77L	285	113	KV17F	A359	285	200	400	20	15	10	100x100
6	S6_1	Crossing	067	0	260	PR71C	B738	240	179	VC59B	A320	240	246	330	17	15	8	100x100

Each training scenario was carefully scripted and tweaked to present a single same-altitude two-aircraft closing conflict (CPA = 0 nm). A CPA of 0 was chosen to reduce solution bias (i.e. a CPA of 4 nm would favour a small heading change for one aircraft). Conflict points-of-collision were distributed throughout the airspace, to reduce predictability. Traffic scenarios included proximate ‘noise aircraft’ that limited (but did not prevent) altitude maneuvers. This was done to encourage heading solutions. For reasons of experimental control, the eventual ML system advisories presented only heading solutions. No altitude solutions were presented. Transparency manipulations (i.e., adding diagram and text to the baseline vector display) focused exclusively on heading solutions.

Scripted conflicts

Each scenario contained a single two-aircraft closing conflict, with the conflict pair at the same flight level and on converging headings. This decision to restrict conflict types to only two aircraft was done largely to simplify creation of ML models, and to limit the amount of training data required. Trying to capture more complicated conflict patterns would have placed data requirements beyond the scope of the project.

2.5.2 Training pre-test procedures

Participants

SIM2A used 20 air traffic controllers provided by ANACNA Italy, equally distributed between Padua, Milan, Rome, and Brindisi ACCs. Age ranged from 35 to 59 ($\bar{X} = 45.5$). SIM2B used 16 air traffic controllers provided by LFV Sweden. All 16 controllers were assigned to Malmö. Age ranged from 37 to 58 ($\bar{X} = 43.9$).

Simulation sessions

The training pre-test sessions were conducted from 1-7 Dec 2021 (SIM2A) and 22 – 25 March, 2022 (SIM2B).

After they introduced for training pre-test participants the MAHALO project and the research consortium, researchers administered a consent form and demographics questionnaire. A briefing was then conducted on the simulator interface, the general format for measurement sessions, and what

was expected of participants. This was followed by a 25-minute training session, broken into three steps:

- a 10-minute training walk-through, in which the researcher instructed the participants on the use of the simulator and interface;
- a 10-minute self-training, in which participants interacted independently with training scenarios, and experimenters made themselves available to answer questions;
- a five-minute training test, in which the experimental leader queried participants during a dynamic scenario session, to check their knowledge of SectorX, including understanding of flight label elements, interface use, etc. Any questions that participants had were fully answered before completion of the training test.

Participants were then presented three simulation sessions of roughly 30 minutes, each of which included 12 short (2.5 minute) traffic scenarios. Including a 10-minute break between each of the three sessions, the entire training pre-test session lasted about three hours per participant.

During the three measurement sessions, controllers were each presented a total of 36 traffic scenarios. These 36 scenarios were created as six variants of six base traffic scenarios. Controllers interacted via mouse and keyboard with the SectorX simulation. All scenarios were displayed as a hypothetical en route sector in a roughly 100 nm square airspace. The simulation ran at 2X speed, meaning that aircraft moved two times faster than normal. Participants were made aware that this overspeed could affect aspects of the simulation including rate of climb and descent. The controller's task included ensuring separation between aircraft and transiting aircraft through their assigned exit waypoint and exit flight level as flight planned. Participants were instructed that loss of separation conflicts might occasionally occur between aircraft, and these were defined by the standard five NM and 1000-foot protected zone. Short-term conflict alert (STCA) functionality was provided, and operated with a threshold of 120 seconds simulation time (i.e., 60 seconds real-time before loss of separation).

During each scenario, a flow of en route traffic was presented. Training scenarios were scripted to include a single two aircraft conflict. These conflicts appeared at different times during each 2.5-minute scenario. Controllers were instructed to solve conflicts as they wished, although scenario design encouraged the use of heading solutions via limiting aircraft at other flight levels which discourage the use of altitude solutions.

Some simplifying assumptions were made for purposes of simulation. These included: no departure or destination aerodrome was presented: Semi-circular cruising level rules were disregarded (as previously mentioned); wind effects were absent; and issued clearances were carried out without any aircraft / pilot delay. Further, air-ground communication was carried out via datalink CPDLC, and thus no radiotelephony (RT) was required.

Annex A contains the training pre-test participant materials, which include the following:

- statement of informed consent (compliant with article 13 of the European GDPR 2016/679);
- demographics questionnaire (covering age, ATC ratings and experience);
- experiment briefing; and
- debrief questionnaire (eight items, including general open-ended questions and specific (agreement scale) items on conflict resolution strategies, and general impressions).

Data logging

As configured for the field simulations SectorX recorded both traffic (radar) events, and control actions, and wrote to timestamped state- and event XML log files. Running at 2x real-time, traffic state (callsign, heading, altitude, bank angle, speed, target altitude and heading, etc) was periodically written to file for each active aircraft. Control interactions were also recorded, including:

- mouse clicks,
- flight clearances (including agent, aircraft, clearance type, and clearance value),
- menu interactions,
- proposal interactions
- zoom / pan inputs,
- flight label clicks and selections,
- VERA interactions, and
- Post-scenario subjective responses.

All events and traffic states were stored in real- and simulation (2x) time.

2.5.3 ML training and test scenario creation

General approach

The basic sequence from training scenarios to trained models is shown in figure 7.

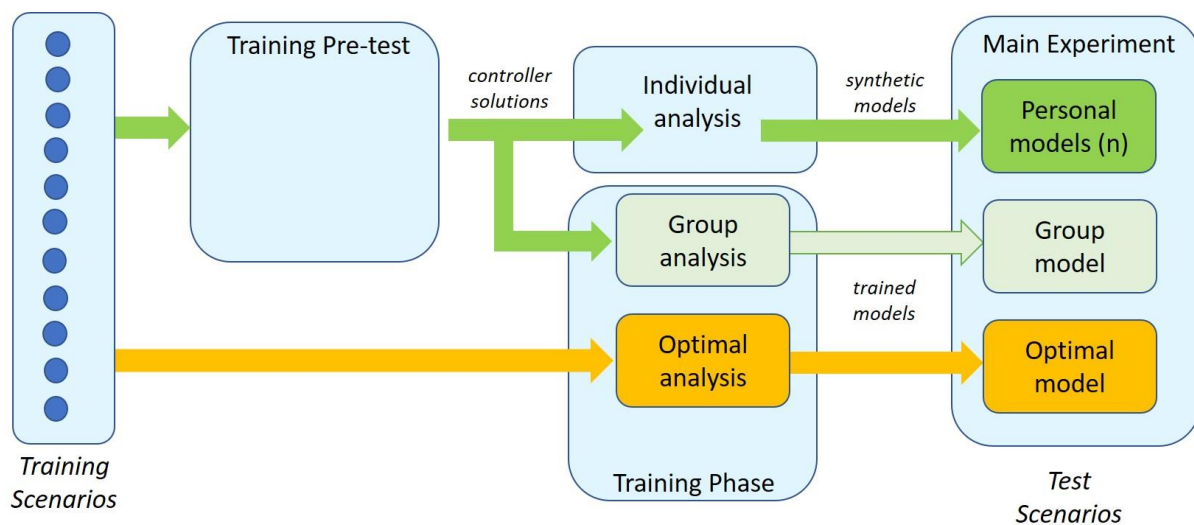


Figure 7. ML training- and test scenario creation.

Figure 7 shows the general workflow involved in going from training scenarios to ML training to test scenario creation, as used in the main experiment. Notice that the path is slightly different for each of the three main experiment conformance models. The personal and group models were both based on controller behavior during the training pre-test, in which a total of 36 training scenarios was presented. The resulting controller solutions were then used in one of two ways.

For the personal model, ML training performance was insufficient. This possibility had been anticipated, since each individual would generate only 36 training samples from the training pre-test,

a very small number in terms of machine learning data requirements. As a result, it was decided to use a synthetic advisory generation procedure. Personal models for each individual and scenario were created, by hand, following an analysis of individual participant's performance in the training pre-test. The personal models were created using the parameters in Table 2. The created personal models, as used in the main experiment test scenarios, set these parameters in line with training pre-test performance for each given controller and each given reference traffic scenario.

We determined conformal models for each of the six scenarios. Since each scenario, in the pre-conformal simulation, was repeated six times, the conformal model was based on the most frequently implemented solution across scenario repetitions. When comparing controllers solution performance, we observed differences in their solution variability – i.e. the spread of different solutions implemented across repetitions. Controllers varied in how consistent they had been in solving the same conflict. An example of a more consistent participant was one who in Scenario A solved all six repeated conflicts by first vectoring aircraft A. In five out of six scenarios, aircraft A was vectored left behind aircraft B. In five out of six scenarios this first interaction was followed by a seconds, vectoring aircraft B left in front of aircraft A. For this participant, the average CPA was 12.12 nm with a standard deviation of 2.87 nm. In contrast, an example of a participant with low consistency was one who in scenario A solved five out of six repetitions using heading (one with altitude) interacting first with either aircraft A (two times) vectored left behind B or aircraft B (three times) vectored right behind A (two times) or vectored left in front of A (one time). The average CPA for this participant was 9.79 nm with a standard deviation of 3.19 nm.

The definition of individual conformal solutions was based on a manual analysis of participant's solution across scenario repetitions, using the framework devised in[3]. With each scenario being repeated six times, the conformal advisory represented the most frequently implemented heading solution across these repetitions. Table 2 shows which solution parameters that were used to define the individual conformal models and how the data was treated. To derive the conformal heading value (e.g. 20 degrees) the participant's separation distance preference across repetitions was used. Note that this contrasts with the approach used in the MUFASA studies [2, 3] to derive directional value, where it instead was based on analysing the participant's most frequently used directional value. The reason for basing directional value on separation distance was based on feedback from controllers in the first workshop that separation distance is an important parameter to consider when solving conflicts. Furthermore, the use of separation distance to derive directional value allowed for affording text/agent-based transparency in a consistent way between individual conformal advisories, group conformal advisories, and optimal advisories. The agent-based transparency explained to the participant that the underlying reason for the directional value was to achieve a desired separation distance.

Table 2. Parameters used for personal models.

Solution parameter	Description
0. Decision time (used to derive conflict detection time and when advisory should be presented).	An average decision time was calculated based on all participants judged decision time. The average decision time was based on the time between use of VERA tool on conflict pair and first interaction, where

	VERA was used before the first interaction. Based on Sim2A, the decision time was 18 seconds.
1. Resolution time (first interaction to solve conflict) ⁵	Based on earliest time that conflicts were interacted with (note that participants frequently required several interactions to solve the conflict) across repetitions minus the average decision time for all participants across all scenarios (this was based on SIM2A time of 18 sec.). Both heading and altitude solutions were analysed when determining earliest analysis time.
2. Decision strategy (control preference or geometry preference)	Control preference and geometry preference was considered when determining in what direction to turn the aircraft selected under aircraft choice. If participant had a preference for a specific geometry, e.g. A behind B, or for vectoring behind or in front. E.g. participant 15 always (in repetitions) vectored one aircraft behind the other, although aircraft choice varied equally across repetitions (3 of each). If control or geometry preference was not observed, then heading direction was based on heading direction for earliest solution where heading was used.
3. Aircraft choice	Based on most frequently used/interacted aircraft (A or B) considering both heading and altitude solutions. Where no preference could be determined, i.e. both A and B selected equally often, the following rule applied: 1) Select aircraft that was more often selected when only considering one interaction to solve the conflict. 2) if not possible, select aircraft that was interacted with the earliest across repetitions.
4. Resolution type	Only considering heading.
5. Heading direction	Based on most frequently used direction (left or right). If no preference could be established, the turn direction was based on 1) control of geometry

⁵ Conflict detection time was approximated using the time at which VERA was activated for either of the two aircraft of the designed conflict, or STCA was triggered, whichever came first. For all uses of VERA before the conflict resolution was implemented, a decision-making time was calculated by taking the difference in the time between the resolution was implemented and the VERA tool was used on any of the aircraft in the designed conflict. A reciprocal calculation was made for STCA. An average for the decision time for VERA and STCA, for each participant, was calculated. This value was used as a proxy for the decision-making time in cases where neither VERA nor STCA could be used as an indication of conflict detection and time taken to make the first interaction to solve the conflict. The average decision-making time for VERA (time between VERA activated and first interaction made), across all participants and scenarios was 18.4 s. For STCA, the average decision-making time was 21.1 s.

- preference, or 2) direction in earliest interaction where heading was used to solve the conflict.
6. Heading value
Derived in two steps: The first step determined the participant's average achieved separation distance for solving the conflict across scenario repetitions for a specific scenario. The second step calculated the heading value required to achieve this separation based on that participant's intervention time, aircraft choice, and heading direction (as determined for above solution parameters).
7. Separation distance
Based on average separation distance across scenario repetitions where heading was used to solve the conflict.

Here, a more frequently implemented solution represents a more consistency behavior. Because conflicts also could be solved with altitude, not all participants solved all repetitions using heading. The consequence of this was that participants varied in how many times the repetitions had been solved using heading. Therefore participants varied in how consistent they had been in solving the conflicts. Consistency also varied depending on which solution parameter that was considered. For example, several participants' first interaction was equally distributed between aircraft A and B (i.e. interacting with aircraft A first in three scenario repetitions, and aircraft B first in three scenario repetitions). As such their consistency in terms of aircraft choice was low. In terms of decision strategy, however, they displayed a consistent behavior with in five or six out of six repetitions having aircraft A going behind aircraft B (achieved either by vectoring A behind B, or B in front of A).

For the group model, controller solutions were fed directly into ML training, as shown in figure 7. Again, each controller was presented 36 separate scenarios during the training pre-test phase. This yielded $36n$ (where n is the number of participants at each site) training samples. Given the difference in participant sample size between simulations 2A and 2B, the 2A group model was trained on different data sizes. In SIM2A, 720 (i.e., 36×20) samples were used to train the group model. For SIM2B, the group model was trained on only 576 (36×16) samples.

The SIM2A group model showed sufficient convergence and stability. However, the smaller 2B training sample appeared to have a limiting effect on model stability. For this reason, the eventual group model used in SIM2B combined the 2A and 2B training pre-test sample sets ($n=1296$). Notice that the detection time parameter for the group model, which the training model itself did not output, was set to 18 seconds, which represented the earliest interval seen in training pre-test performance (and was therefore equivalent to the lowest interval seen in the personal models).

For the optimal test scenarios, training scenarios were fed directly into RL training during the training phase. As shown in figure 7, generating the optimal scenarios bypassed controller interaction during the training pre-test phase.

Offline ML solution generation

As noted in section 2.1, group and optimal test scenarios incorporated ML solutions, but these solutions were scripted into the test scenarios, rather than being generated in real time. That is, during

the interim phase between training pre-test and main experiment, the ML models (for the group and optimal conditions) were trained offline using the training pre-test samples, and the generated solutions were then recorded and manually inserted for replay in the solution traffic scenarios.

For different reasons, two inherent limitations of the experimental design forced a semi-manual process for creation of the group and optimal test scenarios. First, traffic scenarios featured closing collisions of two aircraft in same altitude / no wind conditions, and derived solutions were only associated with a single aircraft. In this situation, there is no basis on which to choose one aircraft over the other to maneuver. The research team therefore chose aircraft to be maneuvered, based on the most frequent choice across controllers. Second, the timing of solution advisories is critical. If an advisory is too early, it can act as an alert and both contaminate experimental data, but also add to the controller's workload. If an advisory is too late, it is by definition useless as the controller has already devised a solution. For the personal conformal condition, determining the proper advisory presentation time was fairly straightforward. Label or VERA interaction was used as a proxy for detection time, and advisory presentation in the test scenarios was keyed to the timing of these interactions. For the group- and optimal conditions, however, the most appropriate time for advisory onset seem to be group average. For this reason, solution display onset was manually scripted according to group average, for both the group conformal and optimal conditions. Notice that one impact of this group average timing is the potential mismatch for some controllers in the solution timing, in training pre-test vs main experiment scenarios. That is, if a given controller was much slower than average to respond across training pre-test scenarios, their solutions (during group and optimal conditions) might be presented too early for that individual's preference. This issue will be discussed later in chapter 3.

Table 3. Group and Optimal advisories in SIM2A and SIM2B

<i>Advisory conformance</i>	SIM 2A		SIM 2B	
	S2	S3	S2	S3
GROUP				
Advisory time	48	51	54	58
Control action	In front	In front	In front	In front
Aircraft	A	A	A	A
Resolution direction	Right	Right	Right	Right
CPA (aim)	10.5 (corrected, target 9)	9 (corrected, target 7.5)	6.9	6.6
Heading deviation	30	20	20	15
OPTIMAL				
Advisory time	20	20	114	96
Control action	Behind	Behind	Behind	In front
Aircraft	B	A	A	B
Resolution direction	Right	Left	Left	Left
CPA (aim)	6.6 (corrected, target 6)	7.7 (corrected, target 6)	10.7-10.8 (corrected, target 10)	10.3-10.6 (corrected, target 10)
Heading deviation	17	-15	-40	-29

Note that the group model is identical across samples with respect to control action, aircraft choice, resolution direction. Advisory time varies but is still close. The major difference is the separation distance and heading deviation. In SIM2B, the separation distance is tighter and heading deviation smaller. The optimal differ between both sample and scenarios on several parameters. Noteworthy is the difference in advisory time and the separation distance between SIM2A and SIM2B.

Advisory conformance, SIM 2A personal models

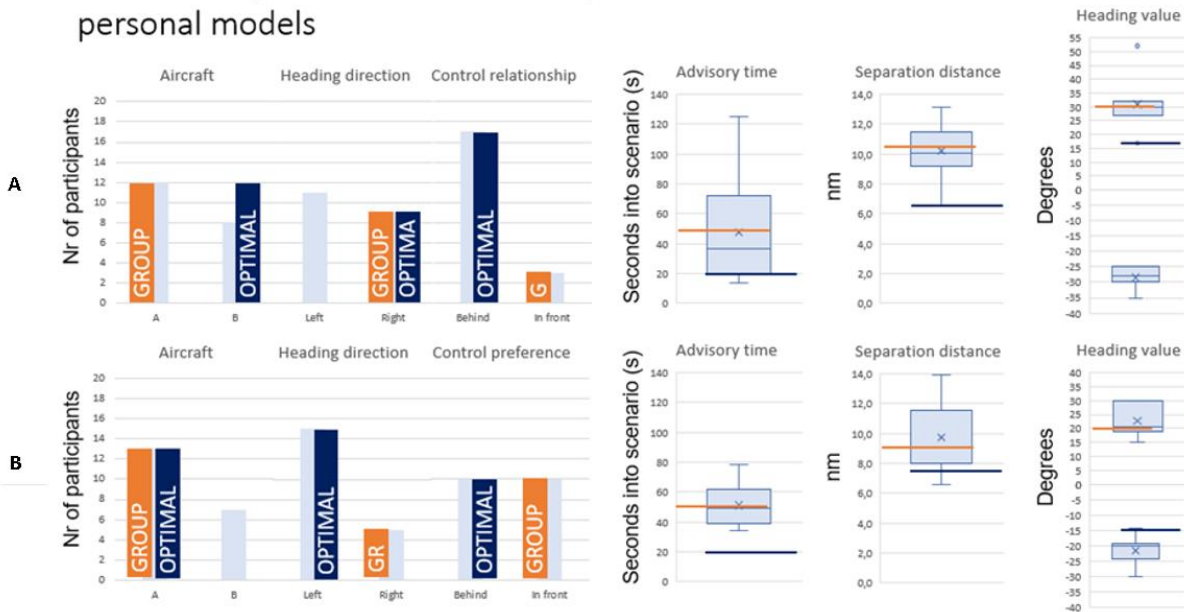


Figure 8. SIM2A models.

Advisory conformance, SIM 2B personal models

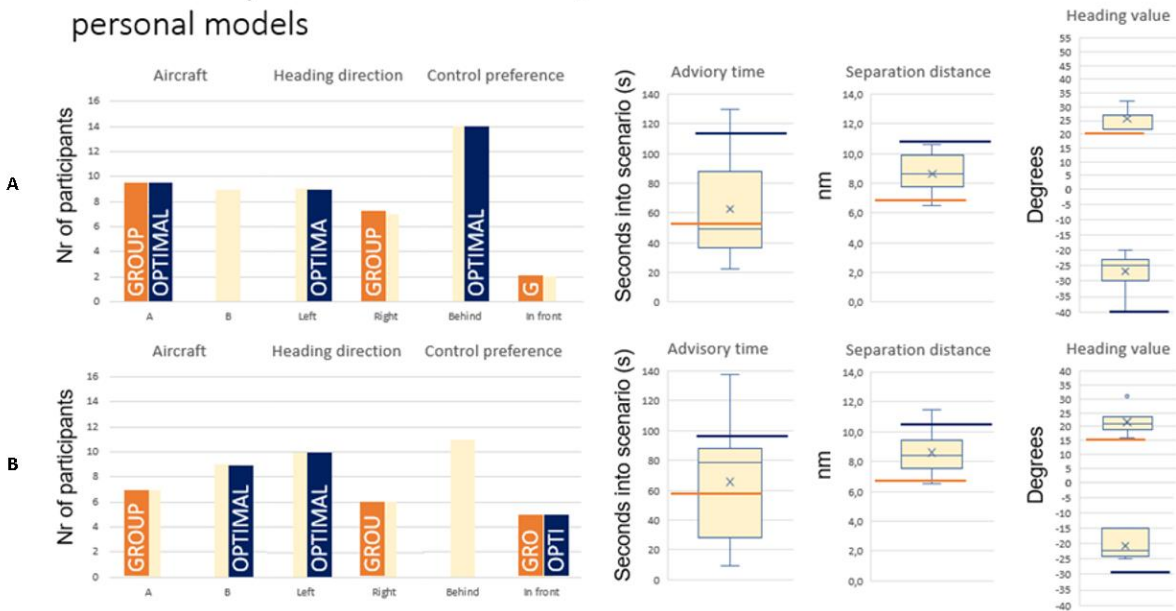


Figure 9. SIM2B models.

In the figures, the personal models have been combined so that aircraft choice, heading direction, and control preference are summarised for all participants for sample and scenario. Advisory time, separation distance, and heading value for all participants are shown as boxplots. The two figures depict the differences between personal model (when aggregating them for all participants), the group models, and optimal models.

When comparing group and optimal models in SIM2A and SIM2B, we can see that the optimal advisory is at the very edge or outside the boxplot of conformal models for advisory time, separation distance, and heading deviation. On the other hand, it matches fairly well the majority of personal models for aircraft choice, heading direction, and control preference (except for scenario B in SIM2B). We expected the group model to provide advisories that are close to what the average preference among controllers would be. When looking at SIM2A we can see that the group advisories do match personal models fairly well, with the advisory being roughly in the middle of the advisory time, separation distance, and heading deviation box plots for both scenarios. However, when looking at control preference for scenario A, and heading direction for scenario B, the group advisory is aligned with what the minority of personal models. An explanation for this can be that the group model was trained on all scenarios in the training pre-test. It may be that overall, aircraft were vectored more often in front than behind.

When looking at SIM2B, we can see that the group model is deviating more from the most common preferences among all personal models. Notable is that separation distance and heading deviation are outside the boxplots for the personal models, and that the control preference is in front, whereas the majority of personal models specify vectoring one aircraft behind the other. An explanation for this can be that the group model for SIM2B was trained with a dataset combining SIM2A and SIM2B. Further, there was more data in SIM2A, especially because altitude solutions were frequent in SIM2B, the group model is heavily weighted towards sample SIM2A preference.

Advisory conformance, SIM2A Scenario A

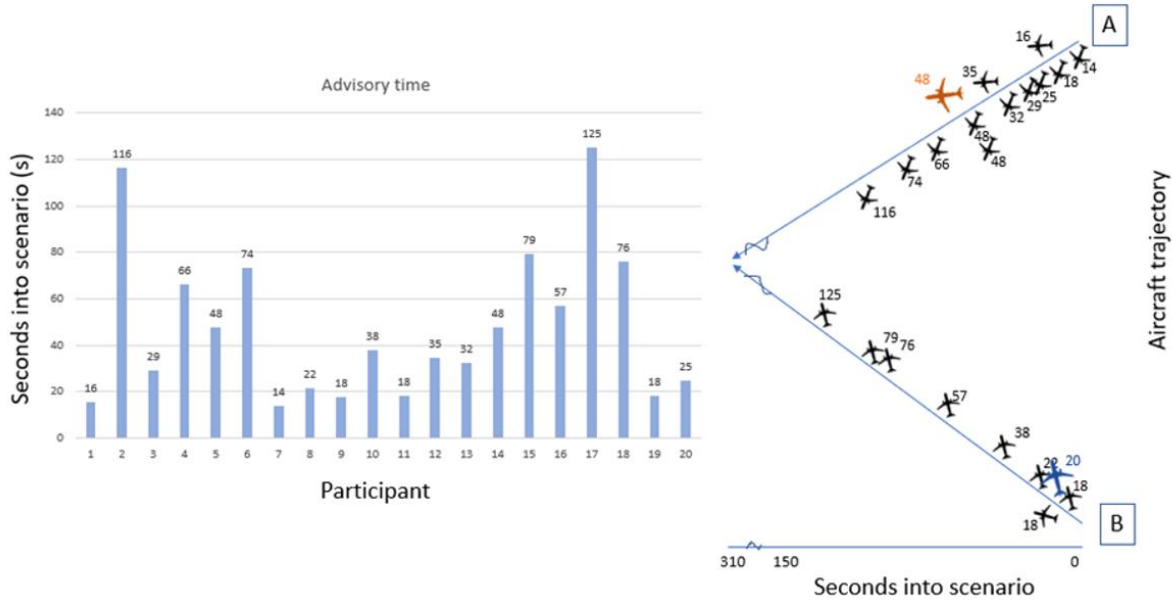


Figure 10. Advisory conformance, SIM2A

This illustration shows the optimal advisory (blue aircraft), group advisory (orange aircraft), and personal advisories (black aircraft) for scenario A in SIM2A. The bar chart shows the advisory time for all personal models, with numbers matching those next to aircraft. The illustration only depicts aircraft choice, control preference (behind or ahead), heading direction (left or right) and advisory time. Separation distance and heading value are not depicted. What can be seen is that the group advisory is similar only to two personal models in terms of aircraft choice, only three personal models in terms

of heading direction (left turn) and control preference (ahead of the other aircraft). Advisory time is, however, similar to the advisory time of many personal models. The optimal advisory has a higher similarity to personal models. The illustration also show that the group and optimal advisory are just two advisories in a spectrum of advisories.

2.5.4 Main experiment procedures

Participants

The experimental design of the field simulations required the same participants to be available for both pre-test and main experiment. The same participants (n=20 in SIM2A; n=16 in SIM2B) returned for the main experiment. Due to illness, the final sample size in the main experiment was 19 participants for SIM2A, and 15 participants for simulation SIM2B.

Simulation sessions

The main experiment sessions were conducted from 19-26 Jan 2022 (SIM2A) and 26-29 Apr 2022 (SIM2B). In a few cases illness forced rescheduling of main experiment dates.

The main experiment sessions used the same SectorX simulation, en route airspace, simplifying assumptions, and basic UI functionality as used during the training pre-test sessions.

For the main experiment, a subset pair of training scenarios were selected for use. These two were chosen (from the six scenario variants of the training pre-test) for reasons of eventual experimental control. For example, these two had minimal altitude solutions, few separation losses and other complicating factors, and an even balance of resolution aircraft choice. The eventual two test scenarios were dubbed scenarios A and B. Two variants of scenarios A and B are shown in figure 11 (colours inverted for visibility).

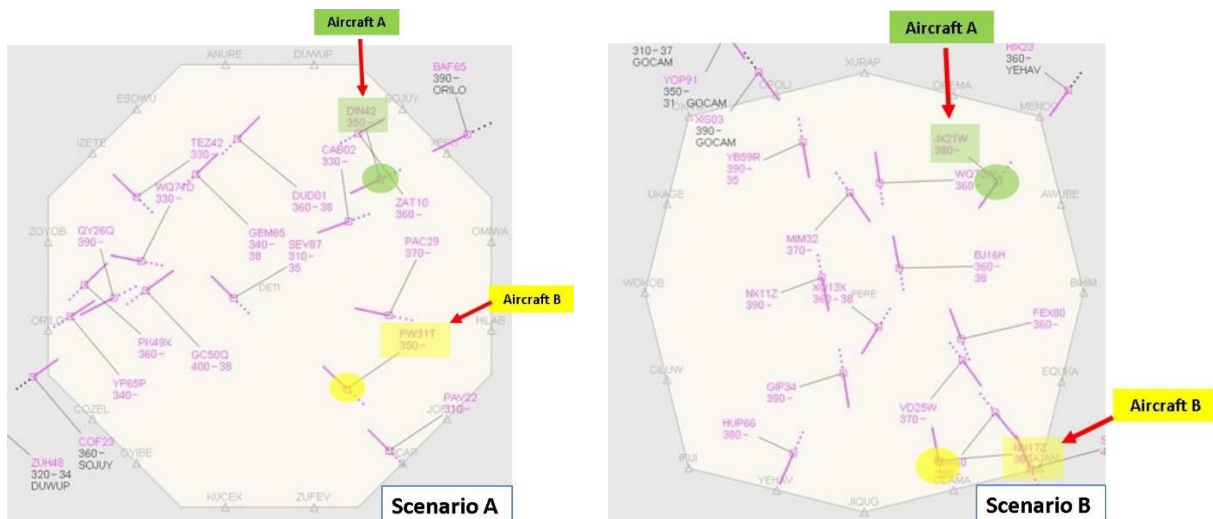


Figure 11. Examples of test scenarios A and B, as used in the main experiment.

After they introduced for participants the aims of the main experiment, researchers briefed participants on the UI as configured for the main experiment. In the main experiment, participants

were assigned via a Latin square procedure to different Conformance and Transparency presentation orders. Participants were unaware that Conformance was varied within each of the three sessions. For each participant, a given session corresponded to one of three Transparency conditions—vector, diagram, or text -- as depicted in figure 12.

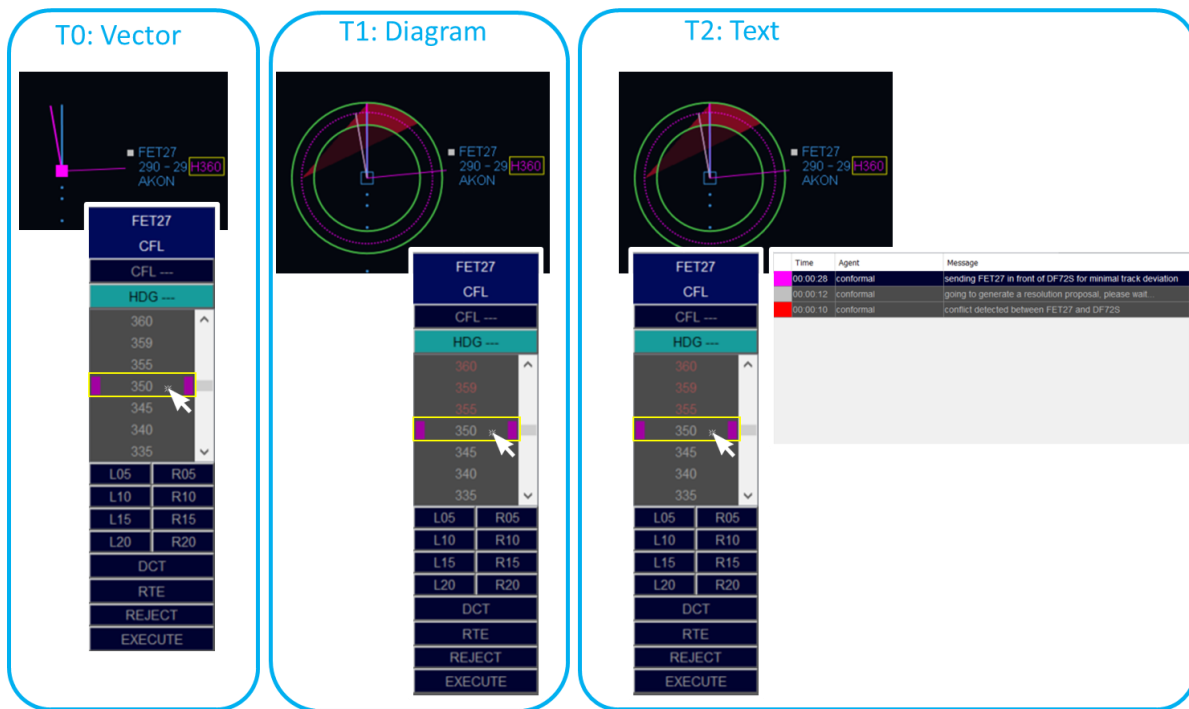


Figure 12. MAHALO transparency conditions.

The vector (or T₀) presentation corresponded to the baseline level of Transparency, and presented only a heading resolution vector. The diagram presentation (T₁) was based on the Solution Space Diagram (SSD) that integrates speed (by concentric green rings) and heading (via red NO GO zones) in proposing conflict resolutions. The text (or T₂) presentation provided both the SSD and text explanation of the chosen solution including target CPA. For example: “Turn [aircraft A] behind [aircraft B] to aim at 8.0 nm separation.”

Prior to each session (which corresponded to one of the three Transparency levels), each controller was given a Walkthrough (5 mins), Self-Training (5 mins), and Training Test (5 mins). During the Training Test, controllers were queried on their understanding of the interface, and required to demonstrate proficiency on interacting with traffic and advisory functionality.

Each session consisted of six test scenarios, each roughly 2.5 mins in length. Participants were instructed to passively supervise automation, and respond only if an advisory were presented. Flight labels were inactive and only after automation issued a proposal was the interface enabled. Once a resolution advisory was presented (a single advisory was scripted per scenario) controllers could open a clearance menu in conjunction with the advisory. Controllers interacted with a given advisory by either executing the proposed solution, or rejecting the proposal and either adjusting the proposed heading or selecting another clearance type and implementing that alternative. The interaction possibilities are shown in Figure 13-15.

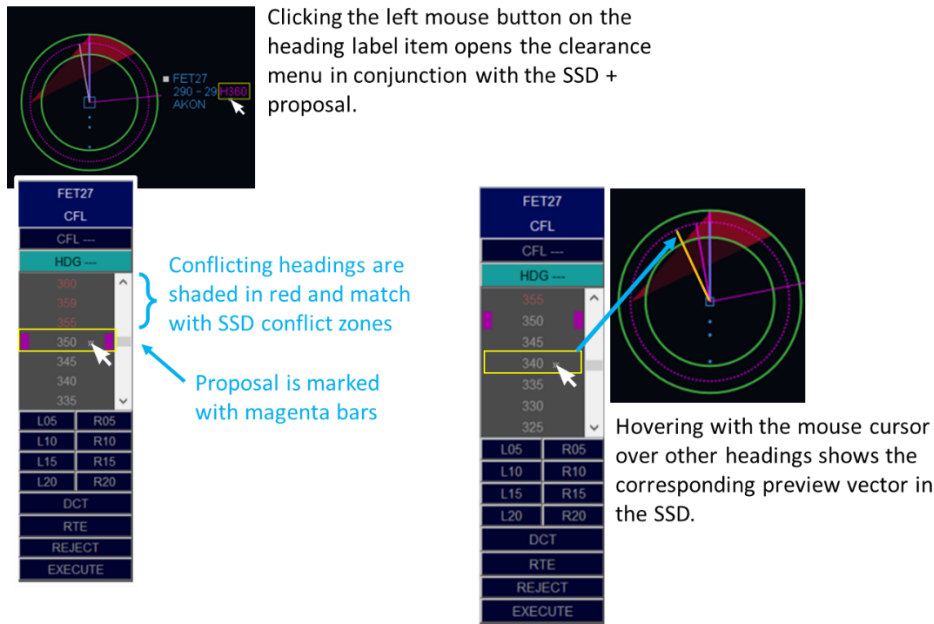


Figure 13. Advisory/proposal interaction in Diagram and Text transparency conditions.

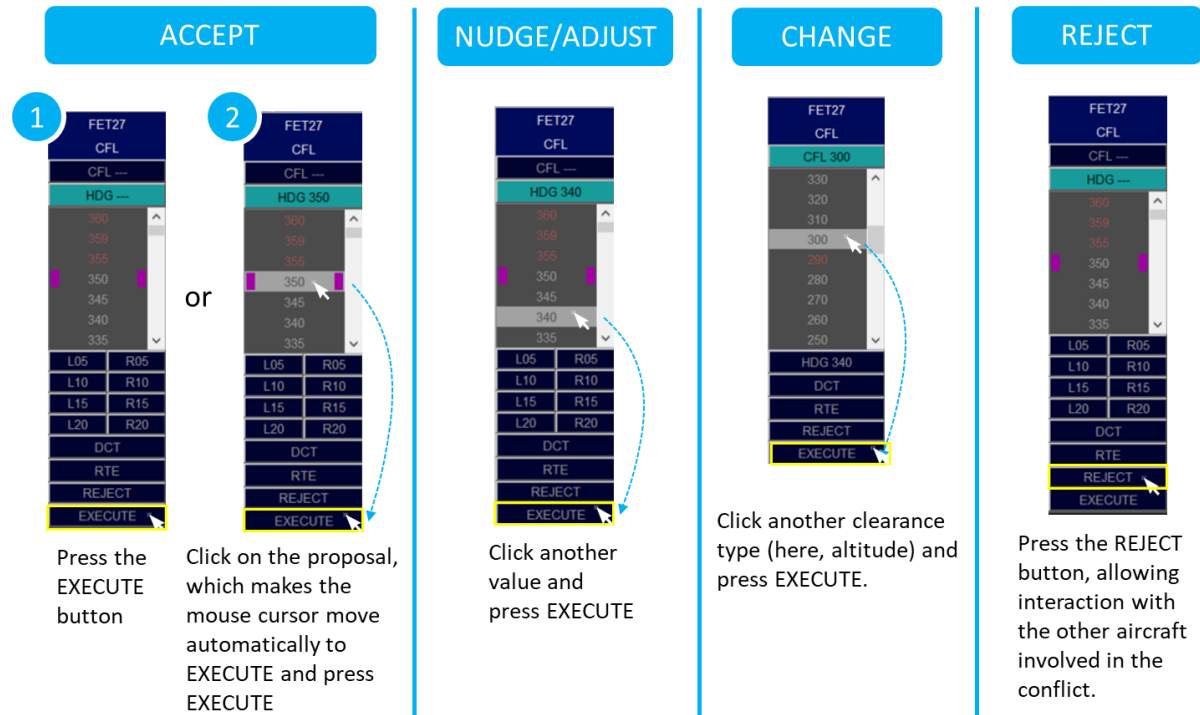


Figure 14. Proposal/advisory interaction.

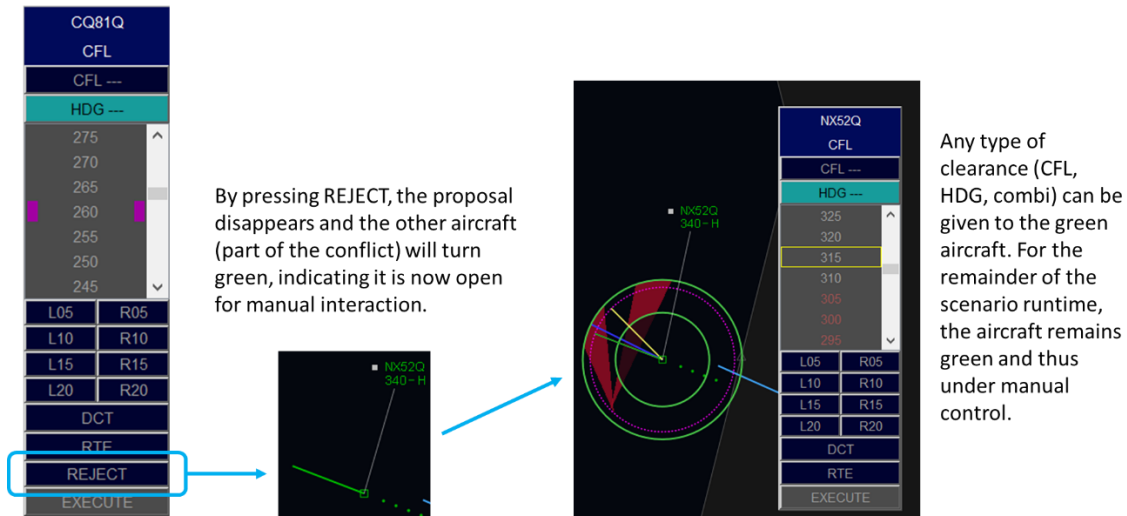


Figure 15. Proposal rejection (only in SIM2B).

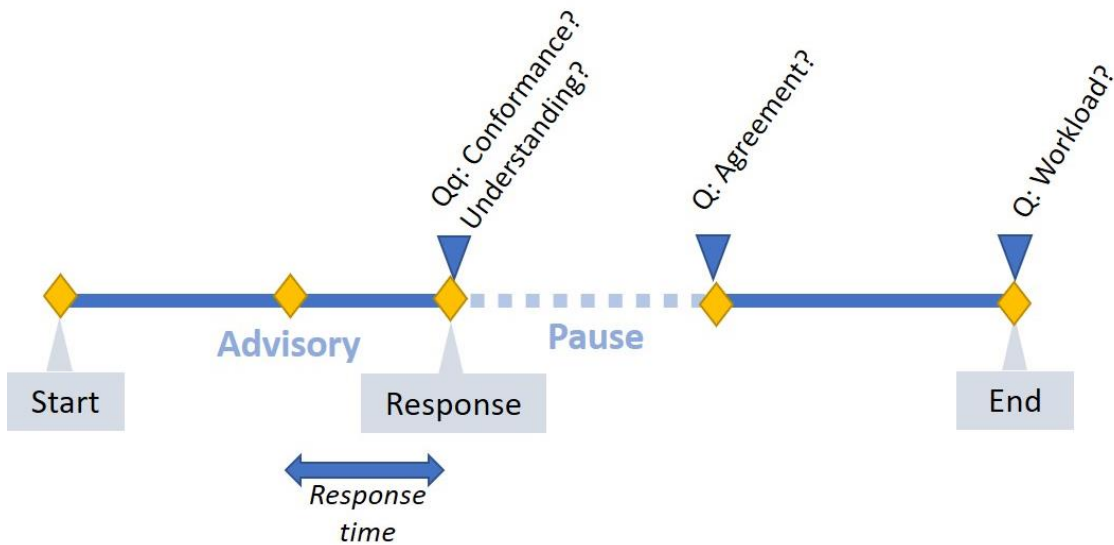


Figure 16. Scenario timing, main experiment.

As shown in figure 16 each main experiment scenario consisted of a number of discrete events. At some point during each scenario an advisory was issued with respect to a conflict aircraft pair. Once the controller responded by either agreeing with or without revising the clearance (by clicking “EXECUTE”) or by rejecting (by clicking “REJECT”), the simulation paused. At this point the controller was prompted to provide ratings of conformance (“The system solved the conflict the same way I would have.”) and understanding (“I can understand why the system suggested that solution.”) on a secondary laptop screen. The controller was also then required to indicate agreement with the proposed solution. Once the controller had entered his/her agreement rating, the simulation resumed.

During this post-pause segment of the scenario, the controller had no interaction with traffic, and the segment was only included to prevent workload rating confounds if scenario lengths had differed. At the end of the scenario run time, the controller was prompted to provide an on-screen 0-100 rating of workload. To encourage response variability, this workload rating prompt defaulted to the workload rating given during the previous scenario (or 50 in the case of the first scenario within a given session).

After completion of each scenario, the controller completed a Post-session Questionnaire. After all three sessions, participants were also administered an Exit Questionnaire, and provided a debriefing on the experimental design and other aspects of the simulation. Including between-session breaks, each main experiment session lasted roughly three hours per participant.

Annex B provides all of the briefing and participant data collection materials used for the main experiment sessions.

Data logging

For the main experiment, the same basic file format structure was used as for the training pre-test, and the same state and event data were written to file on the simulator PC. A flag in the data files captured the agent (human or automation) responsible for a given event. For technical reasons the main experiment used a combination of on-screen prompts and secondary laptop questionnaires for collecting subjective feedback, and these data were integrated in post-processing.

2.5.5 Data analysis procedures

General data analysis strategy

The highest-level aim of data analysis was to address the research hypotheses identified in section 2.3.3. In the first instance, analysis tried to identify the main and interaction effects of conformance and transparency on agreement and acceptance.

The general approach to data analysis was to begin with a high-level overview, looking for overall patterns and global trends in the data. After that, finer-grained analysis would address individual controller differences, and other contextual factors. For example, and as will be discussed later in the results section, Simulation (2A vs 2B) and Scenario (A vs B) were ultimately considered extraneous variables that forced separate analyses.

Data analysis set out to rely on both descriptive and, where appropriate, inferential statistics. Examples of the latter included

- Repeated measures ANOVA - for interval and ratio data
- Friedman test – for non-normally distributed data
- Cochran's Q-test – for binary data

Given the small sample sizes, it was decided that decision rules might be relaxed (above the typical $p=.05$ threshold) to explore trends in the data.

In the first instance, data analysis set out to assess acceptance in a binary (accept versus reject) way. However, initial data analysis quickly suggested that a finer-grained analysis might provide richer data. Specifically, acceptance was classified using a five-point (roughly ordinal) scale, as follows (note that the possibility to interact with other conflict aircraft was not possible in SIM 2A):

- Accept—Fully **accept** and implement the solution, as presented.
- Nudge—accept aircraft choice, heading, and heading direction, but **nudge** the degree (e.g. HDG 10 to HDG 15)
- Adjust—accept aircraft choice and heading, but **adjust** the solution pattern (e.g. implement a RIGHT turn instead of a LEFT turn)
- Change—accept aircraft choice but **change** the clearance type, from heading to altitude.
- Reject—reject and choose to interact with the other conflict aircraft.

Post-processing (and expert feedback from the two project workshops) suggested an additional analysis strategy: given that achieved separation margin is an important element of conflict resolution strategy, analysis set out to ‘bin’ controllers via binary split on average separation margin in the training pre-test.

3. Methods

Analysis of main experiment results started from the assumption that data could not be pooled across simulations, participants, and scenarios. Preliminary analysis revealed non-normal data distributions (e.g. abnormal skewness or kurtosis, or multimodality) that suggested normal inferential statistical methods (e.g. ANOVA) were inappropriate when considering data pooled across scenario and simulation. Moreover, there were also logical reasons to consider the simulations and scenarios separately. For example, optimal advisory was calculated differently (using local group averages) across simulations, and there were slight refinements to the simulation methodology that rendered simulations qualitatively different. Also, scenarios unavoidably differed in traffic geometry, context traffic, etc, and these differences were judged (on the basis of preliminary analysis) to render them fundamentally distinct for reasons of data analysis. . In the end, this led to a segmented data analysis approach, broken out by scenario and simulation. For the sake of completeness, pooled data results are shown in section 3, but these results should be treated with caution.

Another source of variability revealed in preliminary data analysis concerned between-controller differences in preferred resolution strategy. Annex C shows the large inter-controller variability in preferred resolution strategy, specifically in the CPA they chose (i.e., how tight a gap they tended to use in deconflicting aircraft). Some controllers aimed for a tighter, and some a tighter separation margin. Based on this discovery, together with feedback from controllers in two workshops (where it was noted that target CPA would be a crucial aspect of individual strategy in personalising a CD&R algorithm), it was decided to analyse main experiment results from SIM2A and SIM2B based on participants' average separation distance. This was done using a binary split: Participants were divided into two groups depending on their average separation distance achieved in the six scenario repetitions of the training pre-test simulation. Figures 17 and 18 show the boxplots for the two groups of participants for both scenarios n SIM2A and SIM2B. Note that the group of participants differ between scenarios A and B.

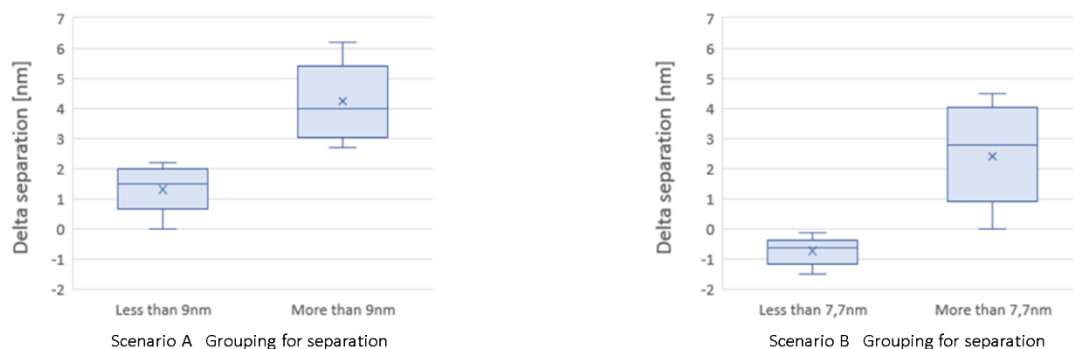


Figure 17. Grouping of participants in SIM 2A based on individual separation margin preference in training pre-test.

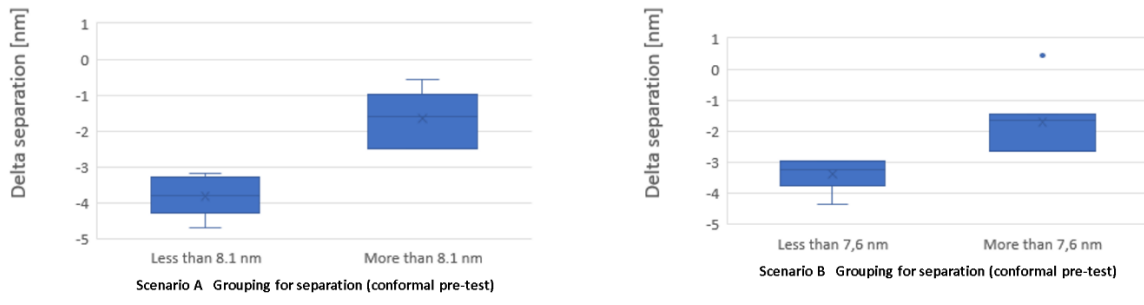


Figure 18. Grouping of participants in SIM 2B based on individual separation margin preference in training pre-test.

3.1. Conformance and Transparency effects

3.1.1 Acceptance of advisories

It was hypothesized at the outset that controllers would be more likely to accept system advisories if these both matched their own personal strategies (i.e. conformal) and were presented with enhanced transparency. In other words, the research team expected to find the main effect of both conformance and transparency on the acceptance of advisories. In the first instance acceptance was seen as a binary decision to either use or disuse a given advisory. As discussed earlier, preliminary analysis led the team to refine the definition of acceptance along a five point roughly ordinal scale (accept, nudge, adjust, change, or reject).

Table 4 shows the breakdown of acceptance by conformance and transparency, in absolute count. Data are collapsed across both simulations and scenarios.

Table 4. Acceptance by Conformance and Transparency (absolute count), pooled data.

	Vector			Diagram			Text		
	Pers	Group	Opt	Pers	Group	Opt	Pers	Group	Opt
Accept	40	35	51	43	35	40	36	37	29
Nudge	21	12	13	20	15	19	17	17	26
Adjust	5	14	4	5	11	3	7	10	7
Change	0	1	1	0	1	1	0	1	1
Reject	2	6	3	0	6	5	4	3	5

Overall, the slight majority (56.5%) of advisories were fully accepted. Following in order were: Nudge (26.1%); Adjust (10.8%); Reject (5.6%); and Change (<1%).

Again, agreement was originally intended as a binary behavioural measure—either an ATCO accepted an advisory, or s/he rejected it. After initial analysis, the research team decided on a five-point scale from: Accept; Nudge; Adjust; Change; and Reject. At the extremes, acceptance meant implementing the proposed resolution as-is, and rejection meant overriding the proposed resolution and acting on the other aircraft, that is the conflicting aircraft that the system had not chosen. From the data shown

in table 4, 346/612 (56.5%) of all advisories were fully accepted, and 506/612 (82.7%) were either accepted or nudged. If we consider those that were either accepted, nudged, or adjusted, the total jumps to 572/612 (93.5%). As shown in Table 5 acceptance was fairly close across the three conformance levels, however it was consistently slightly lower for the Group condition. Notice that row totals should equal about 100%, but column total do not (because the acceptance categories are additive not discrete).

Table 5. Acceptance (fine-grained) by Conformance model, pooled data.

	Personal	Group	Optimal
Accept fully	34.4%	30.9%	34.7%
Accept + Nudge	35.0%	29.8%	35.2%
Accept+Nudge+Adjust	33.9%	32.5%	33.6%

As shown in Table 6, acceptance was also very close across transparency levels. Looking at full acceptance only, the text condition was associated with noticeably lower acceptance. When adding in the nudge and adjust categories, however, the effect seems to wash out. Text was still associated with the lowest level of acceptance, but the effect diminishes.

Table 6. Acceptance (fine-grained) by Transparency level, pooled data.

	Vector	Diagram	Text
Accept fully	35.9%	34.5%	29.8%
Accept + Nudge	33.2%	34.1%	32.9%
Accept+Nudge+Adjust	31.2%	31.2%	31.0%

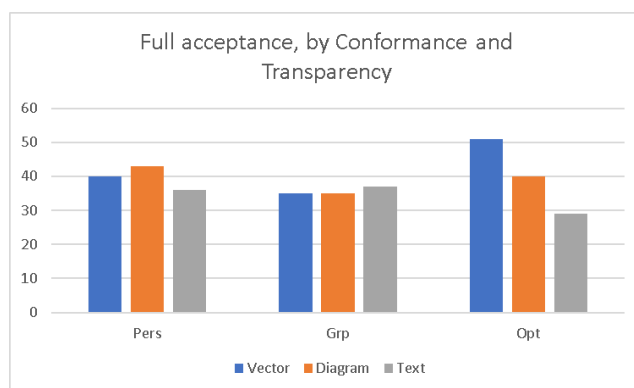


Figure 19. Full acceptance, by Conformance and Transparency, pooled data.

Figure 19 shows the breakdown by conformance and transparency of the 56.5% of advisories that were ‘fully accepted’ across both simulations and scenarios.

Figures 20 and 21 present the combined effects of transparency and conformance on acceptance (for SIM2A and SIM2B, respectively). Notice that each graph also breaks out these effects for Scenario A (the top set of bar charts) and Scenario B (at the bottom).

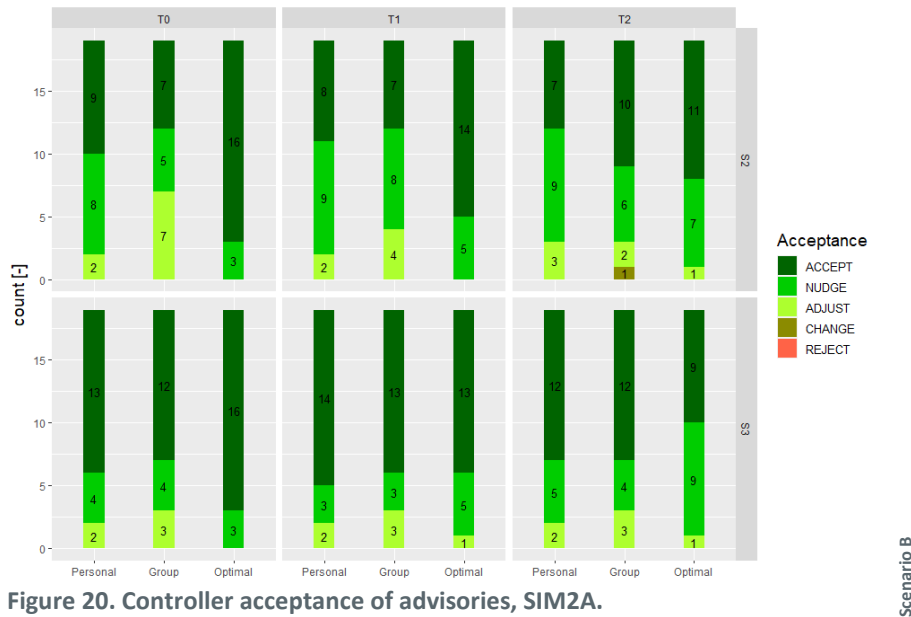


Figure 20. Controller acceptance of advisories, SIM2A.



Figure 21. Controller acceptance of advisories, SIM2B.

First, notice simulation effects. The tendency to reject advisories was different across the two simulations. SIM2A controllers never rejected (which meant rejecting and acting on the other aircraft) whereas this was fairly common in SIM2B. Also, the tendency to fully accept optimal advisories (over personal and group advisories) was slightly greater in SIM2A than SIM2B.

Also, there was an apparent scenario effect. Notice that within each simulation, scenario B (the bottom series of bar charts) tends to show greater full agreement than does scenario A (the top bar charts).

When considering conformance levels, Fig 20 and Fig 21 show that acceptance of personal and group advisories varies little across transparency levels in respective scenario. However, for optimal

advisories, a change in acceptance responses can be seen in the T2 condition compared to the T0 and T1 condition (which are more similar).

A follow-up analysis looked at acceptance by dividing participants in two groups, depending on how similar their personal model was to the optimal model in terms of separation margin. For both scenarios in SIM2A and scenario B in SIM2B, we see fewer changes to an advisory (e.g., nudging, adjusting) in higher transparency conditions (T1 and T2) with the group who has an average separation distance preference closer or less to the optimal advisory’s separation distance. In figure 22, the “less than 9 nm” grouping had a personal model aiming for a separation margin of less than 9 nm, which was closer to the optimal model that had a corrected separation target of 6.6 nm.

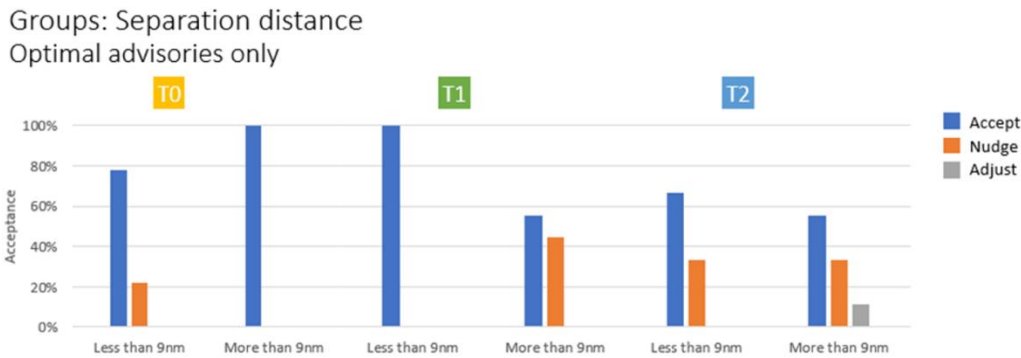


Figure 22. Acceptance of advisories in scenario A, SIM2A.

In figure 23, the “less than 7.7 nm” group had a personal model aiming for a separation margin of less than 7.7 nm, which was less than the optimal model that had a corrected separation target of 7.7 nm.

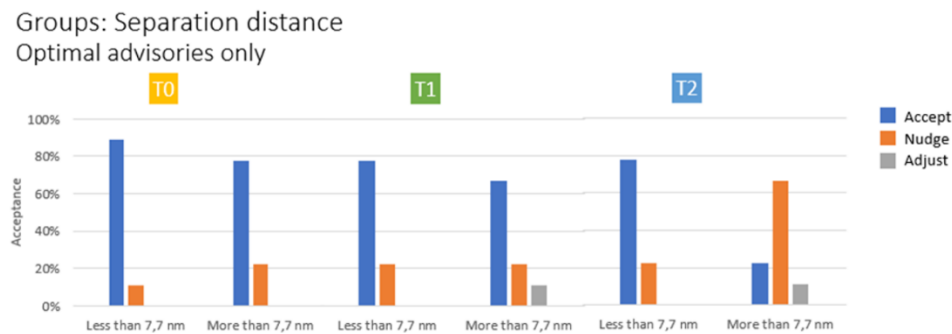


Figure 23. Acceptance of advisories in scenario B, SIM2A.

In figure 24, the “more than 8.1 nm” group had a personal model aiming for a separation margin of more than 8.1 nm, which was closer to the optimal model that had a corrected separation target of 10.7-10.8 nm.

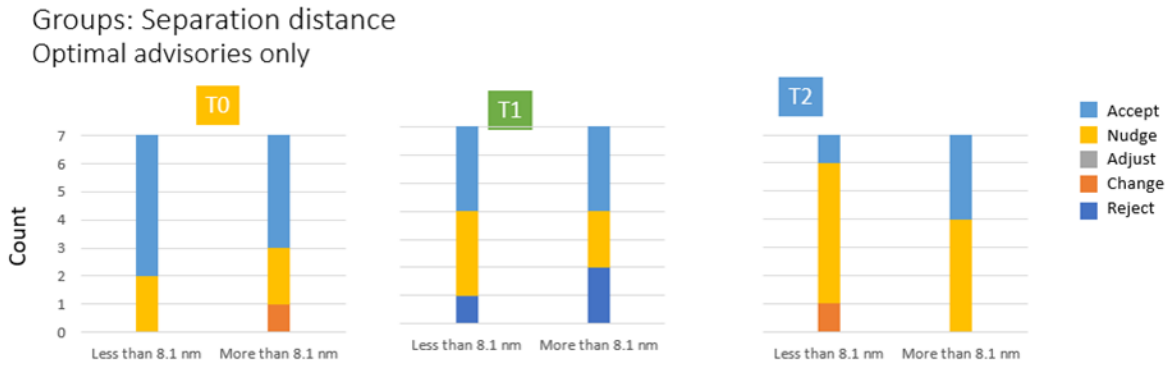


Figure 24. Acceptance of advisories in scenario A, SIM2B

In figure 25, the “more than 7.6 nm” group had a personal model aiming for a separation margin of more than 7.6 nm, which was closer to the optimal model that had a corrected separation target of 10.3-10.6 nm.

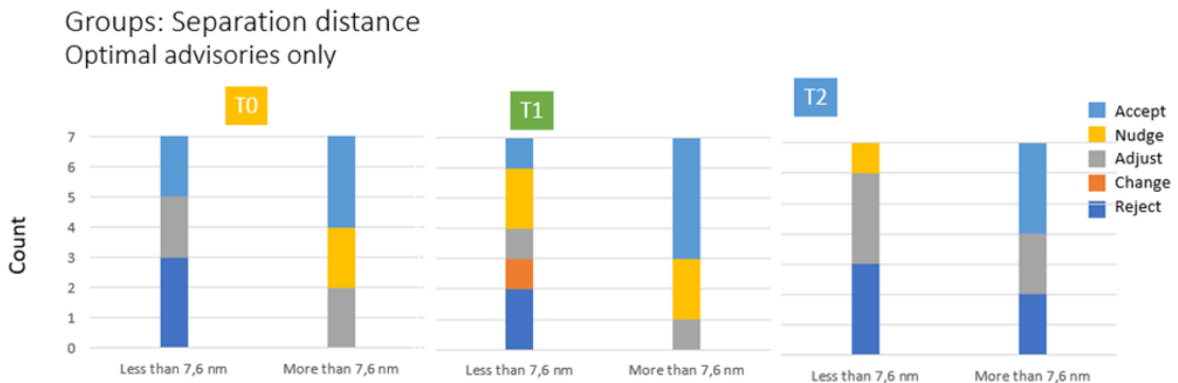


Figure 25. Acceptance of advisories in scenario B, SIM2B.

Given the fine-grained distinction in acceptance (i.e., redefining the binary measure into five levels), data were judged too sparse for inferential statistical analysis

3.1.2 Agreement with advisories

Whereas acceptance was an objective performance measure, agreement was obtained via self-reported ratings. Self-report ratings are notoriously variable across people (some tend to rate high,

some low, use different ranges, etc). To enable meaningful comparison, each controller’s agreement ratings were standardized as Z scores computed within participant. Agreement ratings are shown in figure 26 and Table 7, pooled across scenarios and simulations.

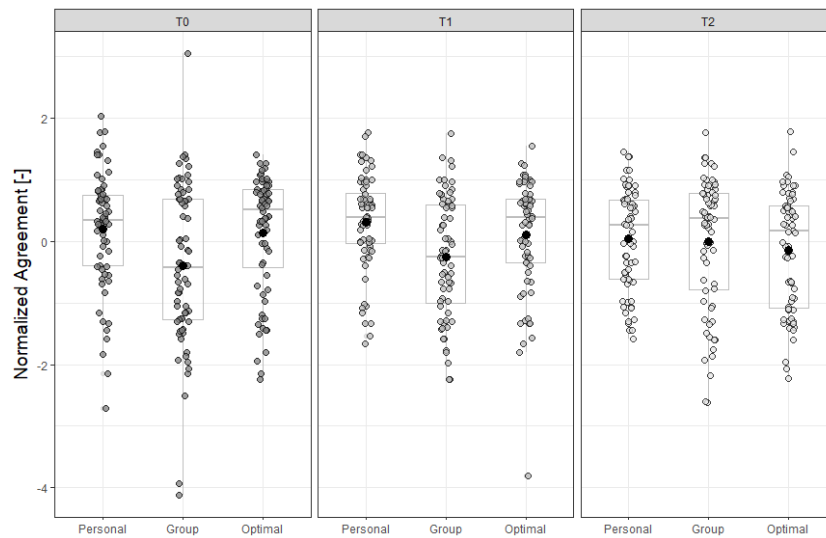


Figure 26. Agreement ratings by transparency and conformance, pooled data.

Table 7. Agreement ratings by transparency and conformance, pooled data.

	Mode ^a	Mean	Std. Deviation
conformal_T0	0.256	0.201	0.945
conformal_T1	0.256	0.314	0.777
conformal_T2	-1.071	0.041	0.819
group_T0	-1.302	-0.393	1.306
group_T1	-1.302	-0.260	0.977
group_T2	-1.608	-0.002	1.070
optimal_T0	-0.040	0.139	0.969
optimal_T1	-1.334	0.103	0.941
optimal_T2	-1.309	-0.143	0.958

Based on the observed high variability of responses, data were broken out separately by simulation and scenario, as shown in figures 27 and 28. Within each graph, scenarios A and B are presented in the top and bottom half, respectively.

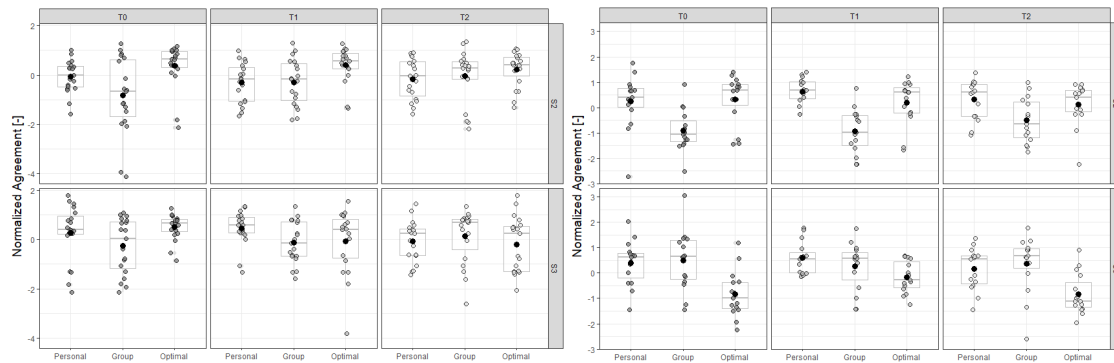


Figure 27. Controller agreement with advisories, SIM2A (left) and SIM 2B (right).

Table 8. Controller agreement with advisories, SIM2A.

Scenario A				Scenario B			
Conformance	Transparency	Mean	SD	Conformance	Transparency	Mean	SD
Personal	T0	-0.065	0.668	Personal	T0	0.273	1.135
	T1	-0.308	0.808		T1	0.455	0.698
	T2	-0.166	0.814		T2	-0.065	0.843
Group	T0	-0.831	1.553	Group	T0	-0.263	1.122
	T1	-0.289	0.945		T1	-0.118	0.857
	T2	-0.047	1.063		T2	0.142	1.087
Optimal	T0	0.390	0.902	Optimal	T0	0.507	0.535
	T1	0.407	0.715		T1	-0.072	1.299
	T2	0.238	0.709		T2	-0.190	1.125

Table 9. Controller agreement with advisories, SIM2B.

Scenario A				Scenario B			
Conformance	Transparency	Mean	SD	Conformance	Transparency	Mean	SD
Personal	T0	0.248	1.062	Personal	T0	0.399	0.877
	T1	0.641	0.527		T1	0.596	0.630
	T2	0.314	0.778		T2	0.162	0.821
Group	T0	-0.902	0.813	Group	T0	0.508	1.161
	T1	-0.928	0.874		T1	0.266	0.952
	T2	-0.503	0.878		T2	0.374	1.134
Optimal	T0	0.328	0.961	Optimal	T0	-0.835	0.912
	T1	0.202	0.879		T1	-0.160	0.622
	T2	0.122	0.833		T2	-0.832	0.799

Acceptance results suggest an interaction trend between conformance and transparency. For the personal model, the vector display produced the highest acceptance. For the group model, T2 (text) showed the highest acceptance; For the optimal model, T1 vector showed the highest acceptance. One suggestion from these data is that the ‘best’ level of transparency, in terms of controller acceptance, might vary with the type of conformance model (personal, group, or optimal) in use.

As shown in table 10, Conformance showed a significant main effect on normalised agreement rating for scenario A in SIM2A, and for both scenarios in SIM2B. For SIM2A scenario A, post-hoc t tests

revealed significant differences between the Personal and Optimal models ($t=-2.56$, $p<.05$) and between the Group and Optimal models ($t=3.57$, $p<.01$). This means that:

In SIM2A Scenario A, normalized agreement ratings were significantly higher for the Optimal model (marginal mean=.35) than for either the Personal (-.18)- or the Group model (-.39).

For simulation 2B, conformance showed a significant main effect for both scenarios (see Table 10). For SIM2A Scenario A, post-hoc t tests revealed significant differences between the Personal and Group models ($t=4.72$, $p<.001$) and between the Optimal and Group models ($t=-4.83$, $p<.001$). This means that:

In SIM2B Scenario A, normalized agreement ratings were significantly lower for the Group model (marginal mean=-.78) than for either the Personal (.40)- or Optimal (.22) models.

For SIM2A Scenario B, post-hoc t tests revealed significant differences between the Optimal model and both the Personal ($t=6.9$, $p<.001$) and the Group ($t=5.85$, $p<.001$). This means that:

In SIM2B Scenario B, normalized agreement ratings were significantly lower for the Optimal model (marginal mean=-.61) than for either the Personal (.38)- or Optimal (.38) models.

The statistically significant effects on conformance can thus be summarized as follows:

- In SIMA Scenario A, the optimal model produced higher agreement than the other models;
- In SIM2B Scenario A, the group model produced the lowest agreement;
- In SIM2B Scenario B, the optimal model produced the lowest agreement.

No main effect of transparency was found. The conformance x transparency interaction trend approached significance for SIM2A Scenario B.

Table 10. Repeated measures ANOVA of normalised agreement

SIM2A	Scenario A	Conformance	$F(2,18) = 6.78$, $p<.01$ ***
		Transparency	$F(2,18) = .47$, $p=.63$
		C x T	No trend ($p=.11$)
	Scenario B	Conformance	$F(2,18) = .80$, $p=.46$
		Transparency	$F(2,18) = .51$, $p=.61$
		C x T	Slight trend ($p=.059$)
SIM2B	Scenario A	Conformance	$F(2,14)= 16.81$, $p<.001$ ***
		Transparency	$F(2,14)= .17$, $p<.85$
		C x T	No trend ($p=.31$)
	Scenario B	Conformance	$F(2,14)= 21.14$, $p<.001$ ***
		Transparency	$F(2,14) = 1.33$, $p=.28$
		C x T	No trend ($p=.33$)

Agreement ratings by preferred separation margin

It was observed that controllers tended to form a bimodal distribution (i.e., cluster in two groups) in terms of their average separation margin during the training pre-test. We therefore conducted an exploratory analysis of this issue by dividing participants into separation margin groups. Because this was considered exploratory, no inferential statistical tests have been run so far.

For scenario A and B in SIM2A, and scenario B in SIM2B, agreement ratings were higher for the group of participants whose average separation margin was closer to the optimal advisory at higher transparency conditions (T1 and T2. Again, scenario A in SIM2B does not follow this pattern.



Figure 29. Agreement rating by separation margin, SIM2A Scenarios A (left) and B (right).

In figure 28, the “less than 9 nm” and “less than 7.7 nm” groups had a personal model aiming for a separation margin of less than 9 nm, which was closer to, or less than the optimal model that had a corrected separation target of 6.6 nm and 7.7 nm, respectively.

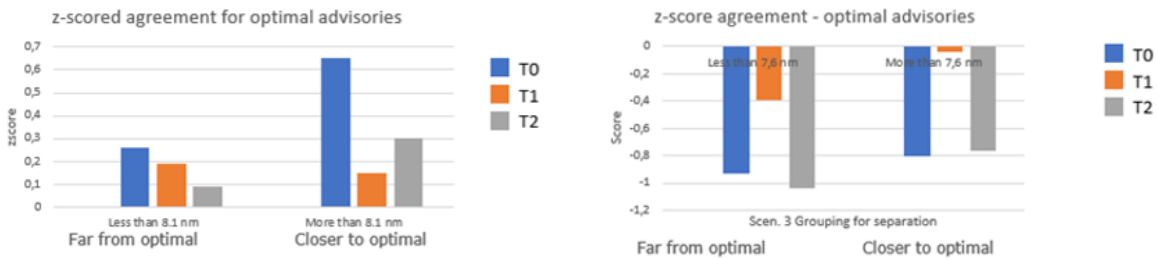


Figure 28. Agreement rating by separation margin, SIM2B Scenarios A (left) and B (right).

In figure 29, the “more than 8.1 nm” group and “more than 7.6 nm” group had a personal model aiming for a separation margin of less than 9 nm, which was closer to the optimal model that had a corrected separation target of 10.7-10.8 nm and 10.3-10.6 nm, respectively.

3.1.3 Self-reported workload

Given the typical inter-respondent variability in workload ratings, workload data were normalized into within-participant calculated Z scores. As a first analysis, simulation and scenario were collapsed to look at pooled data trends in the conformance and transparency impact on rated workload. As shown in figure 41, the pooled data showed extreme variance. Workload data were also therefore broken out by scenario and simulation, as shown in figures 42 (left and right).

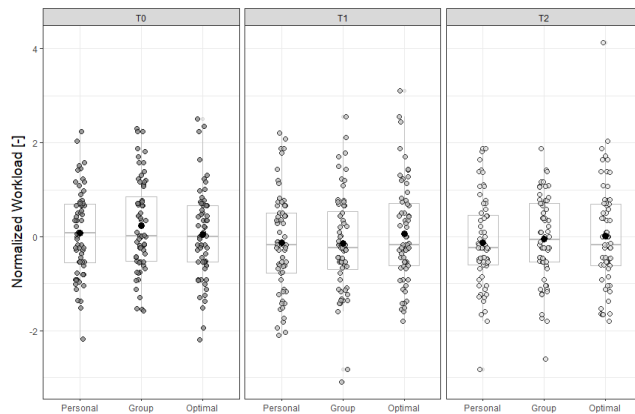


Figure 30. Normalised workload ratings by transparency and conformance, pooled data.

Table 11. Workload ratings by transparency and conformance, pooled data.

	Mode	Mean	Std. Deviation
conformal_T0	0.000	0.071	0.889
conformal_T1	0.000	-0.118	1.011
conformal_T2	0.000	-0.119	0.940
group_T0	0.000	0.227	0.945
group_T1	0.000	-0.137	1.036
group_T2	0.000	-0.045	0.881
optimal_T0	0.000	0.041	0.923
optimal_T1	0.000	0.064	1.031
optimal_T2	0.000	0.015	1.063

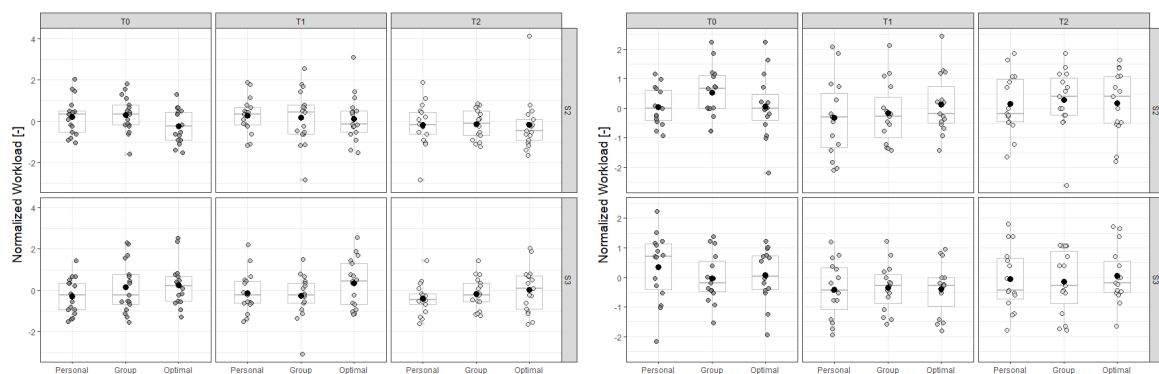


Figure 31. Normalised workload ratings by transparency and conformance, SIM2A (left) and SIM2B (right).

Table 12. Workload ratings, SIM2A.

Scenario A				Scenario B			
Conformance	Transparency	Mean	SD	Conformance	Transparency	Mean	SD
Personal	T0	0.195	0.842	Personal	T0	-0.251	0.824
	T1	0.261	0.804		T1	-0.107	0.918
	T2	-0.147	0.982		T2	-0.345	0.704
Group	T0	0.275	0.822	Group	T0	0.141	1.157
	T1	0.154	1.249		T1	-0.238	0.969
	T2	-0.124	0.676		T2	-0.133	0.723
Optimal	T0	-0.200	0.776	Optimal	T0	0.233	0.962
	T1	0.120	1.046		T1	0.307	1.103
	T2	-0.163	1.219		T2	0.024	1.020

Table 13. Workload ratings, SIM2B.

Scenario A				Scenario B			
Conformance	Transparency	Mean	SD	Conformance	Transparency	Mean	SD
Personal	T0	0.044	0.640	Personal	T0	0.350	1.160
	T1	-0.325	1.311		T1	-0.404	0.964
	T2	0.148	1.022		T2	-0.065	1.075
Group	T0	0.528	0.887	Group	T0	-0.024	0.848
	T1	-0.164	1.053		T1	-0.351	0.800
	T2	0.278	1.071		T2	-0.155	1.078
Optimal	T0	0.067	1.089	Optimal	T0	0.078	0.894
	T1	0.137	1.017		T1	-0.386	0.877
	T2	0.184	1.079		T2	0.060	0.956

Table 14. Repeated measures ANOVA of workload ratings.

SIM2A	Scenario A	Conformance	$F(2,18) = .492, p=.62$
		Transparency	$F(2,18) = 1.57, p=.22$
		C x T	No trend ($p=.75$)
	Scenario B	Conformance	$F(2,18) = 3.84, p<.05^{***}$
		Transparency	$F(2,18) = .473, p=.63$
		C x T	No trend ($p=.76$)
SIM2B	Scenario A	Conformance	$F(2,14)= .747, p=.48$
		Transparency	$F(2,14)= .971, p=.39$
		C x T	No trend ($p=.57$)
	Scenario B	Conformance	$F(2,14)= .556, p=.58$
		Transparency	$F(2,14) = 1.56, p=.23$
		C x T	No trend ($p=.86$)

SIM2A Scenario B showed a significant main effect of conformance, and post-hoc t tests revealed a significant difference between the Personal (margin mean =-.234) and Optimal (marginal mean=+.188) models, $t=-2.74, p<.05$. This means that

For SIM2A Scenario B, the Personal model produced significantly lower workload ratings than did the Optimal model.

Although other workload effects failed to reach statistical significance, it does not mean that they are not practically significant. It is instructive to examine the conformance and transparency trends in the workload data, by looking at interaction plots or marginal means. For example, in both scenarios of SIM2B, the Diagram condition is associated with a workload decrease for five of the six transparency conditions. This pattern is quite clear in figure 32 which plots the interaction for SIM2B Scenario B. Notice the workload dip for the Diagram condition, at each level of conformance model.

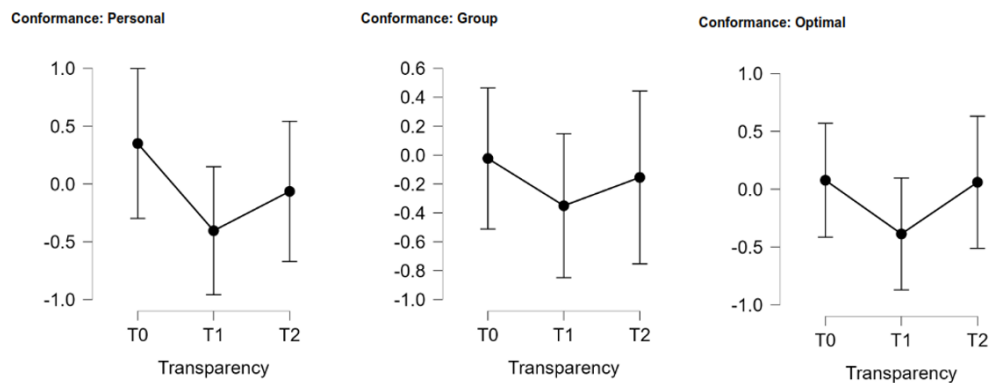


Figure 32. Workload by conformance and transparency, for SIM2B Scenario B.

This pattern is not at all apparent in SIM2A, in which four of the six transparency conditions show a diagram increase in workload, except for the Group model. In both scenarios of SIM2A, the Group model shows a workload decrease in the diagram condition.

Together, these results are a reminder of how the scenarios and simulations cannot be treated as a single data sample.

3.1.4 Delta CPA

As another exploratory analysis, we also analyzed the difference between groups for *Delta CPA*, which was defined as the difference between the separation distance and the target distance. For both scenarios in Sim2A (Fig 33), the group closer to the optimal advisory in terms of separation margin (“less than 9 nm” and “less than 7.7 nm”, respectively for SIM2A) made smaller changes when interacting with the advisory, compared with the group who was further away from the optimal.

In SIM2B, delta CPA for scenario B shows a similar pattern to that observed in SIM2A, where the group with a separation margin closer to the optimal (“more than 7.6 nm” in scenario B for SIM2B) made smaller changes to the suggested advisory. In comparison, the group with a separation distance further from the optimal (average less than 7.6 nm and further away from optimal 10.4 nm) made larger changes.

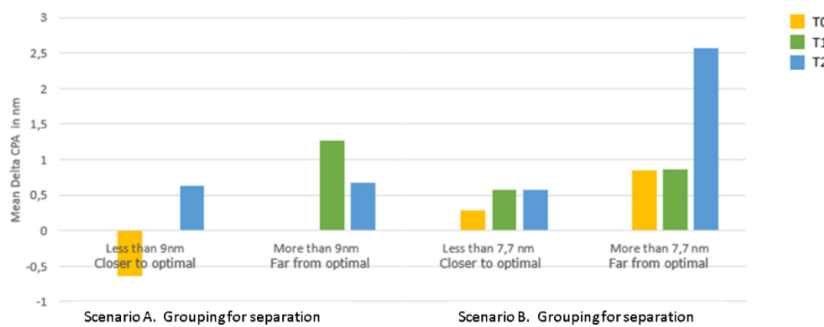


Figure 33. Delta CPA by participant group, SIM2A.

The “less than 9 nm” and “less than 7.7 nm” groups had a personal model aiming for a separation margin of less than 9 nm, which was closer to, or less than the optimal model that had a corrected separation target of 6.6 nm and 7.7 nm, respectively.

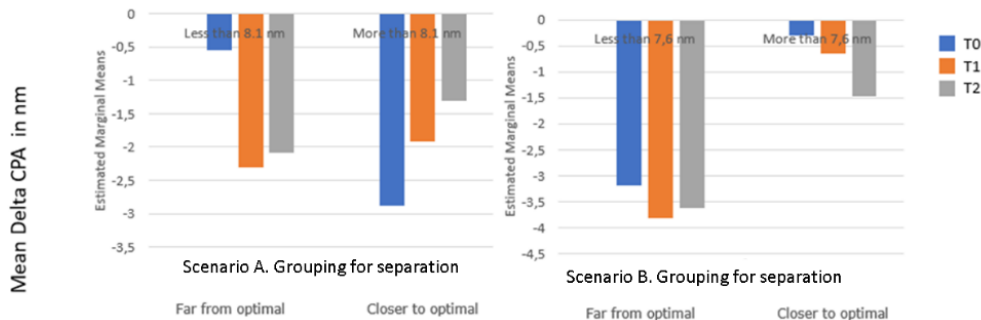


Figure 34. Delta CPA by participant group, SIM2B.

The “more than 8.1 nm” and the “more than 7.6 nm” group had a personal model aiming for a separation margin of less than 9 nm, which was closer to the optimal model that had a corrected separation target of 10.7-10.8 nm and 10.3-10.6 nm, respectively.

3.1.5 Survey results

Post-advisory questions

Within each 2.5 minute scenario during the main experiment, participants had to respond to the presented advisory. At this point the scenario paused and controllers were instructed via on-screen prompt to indicate their agreement with two statements, as follows.

- Statement 1: “The system solved the conflict the same way I would have.”
- Statement 2: “I can understand why the system suggested that solution.”

Statement 1: Similarity of solution

In SIM2A (Fig 35), participants generally rated the optimal advisory as most similar to their own solution strategy, except for the Text condition in Scenario B. However, in both scenarios ratings of

optimal advisories decreased at the highest transparency level (Text). In contrast, ratings of group advisories increased in the Text condition in both Scenarios. Note that personal advisories generally received ratings between ratings of group and optimal advisories.

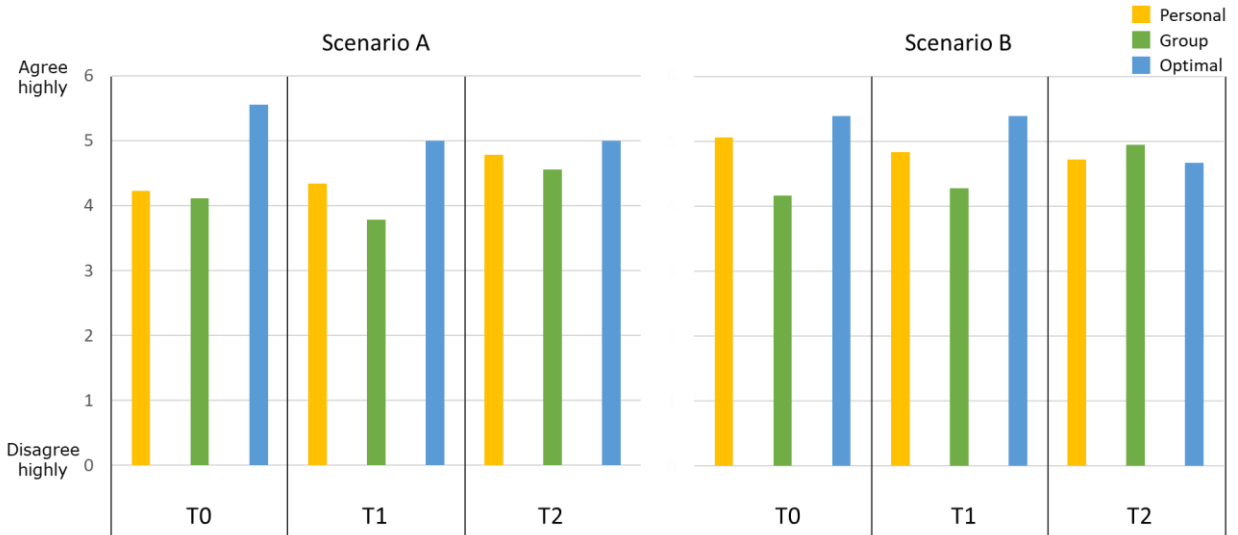


Figure 35. Reported similarity with own solution strategy across conformance and transparency conditions in scenario A and Scenario B, SIM2A (n=18).

In SIM2B (Fig 36), participants generally rated the personal advisory as most similar to their own solution strategy. Group advisories were rated as least similar with own solution strategy in Scenario A, while optimal advisories were rated least similar with own solution strategy in Scenario B. There is no apparent difference between transparency levels.

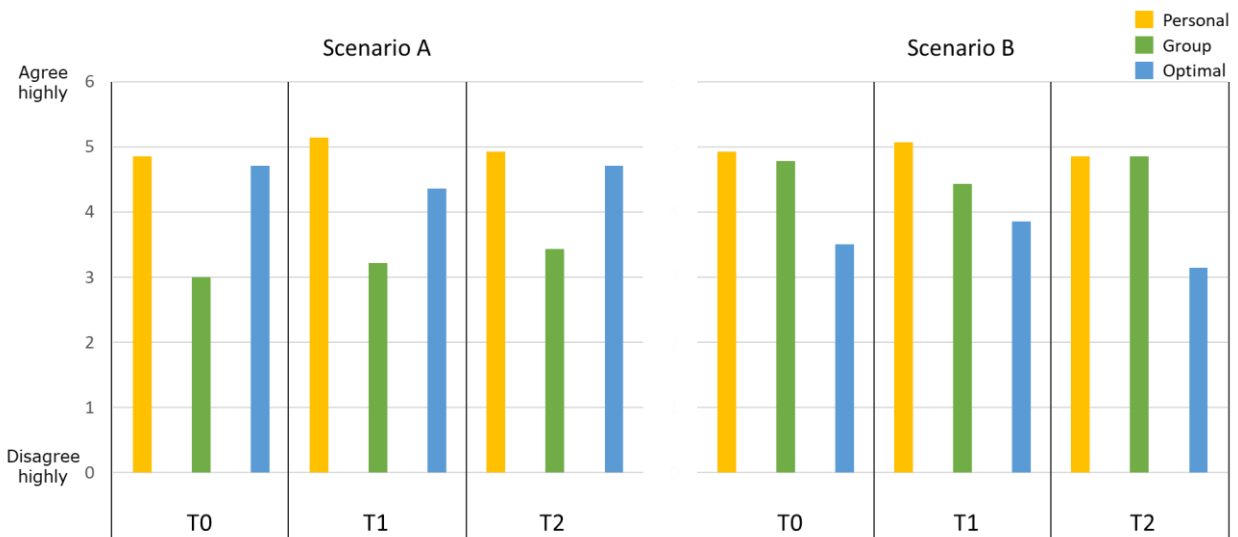


Figure 36. Reported similarity with own solution strategy across conformance and transparency conditions in scenario A and Scenario B, SIM2B (n=14).

In Fig 37 and Fig 38 we can see that groups for both scenarios in SIM2A, and scenario B in SIM2B rated advisories to be more similar to their own solution strategy when optimal advisories were closer to their preferred separation margin (“less than 9 nm” and “less than 7.7 nm”, respectively for SIM2A, and “more than 7.6 nm” in scenario B for SIM2B), especially for higher transparency levels (i.e. diagram and text). For scenario A in SIM2B, however, there is no difference between the groups. Again, this pattern follows what can be observed for acceptance, agreement, and delta CPA.

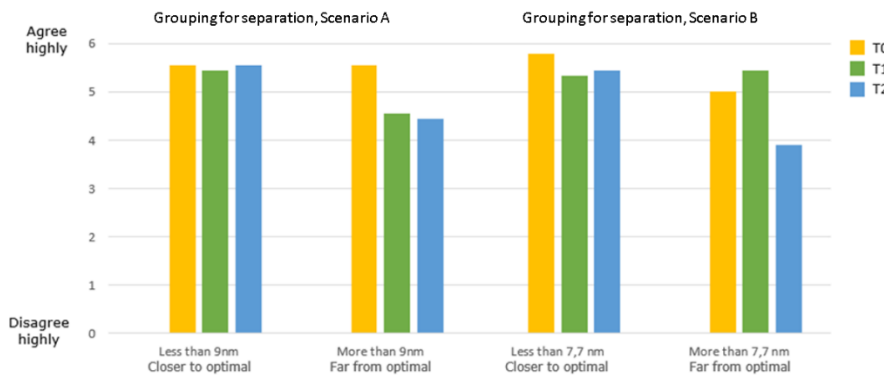


Figure 37. Reported similarity with own solution strategy, SIM2A.

The “less than 9 nm” and “less than 7.7” nm group had a personal model aiming for a separation margin of less than 9 nm, which was closer to, or less than the optimal model that had a corrected separation target of 6.6 nm and 7.7 nm, respectively.

6 point scale: “the system solved the conflict the same way I would have”
Optimal advisories only, where CPA target is 10.8 and 10.4 nm

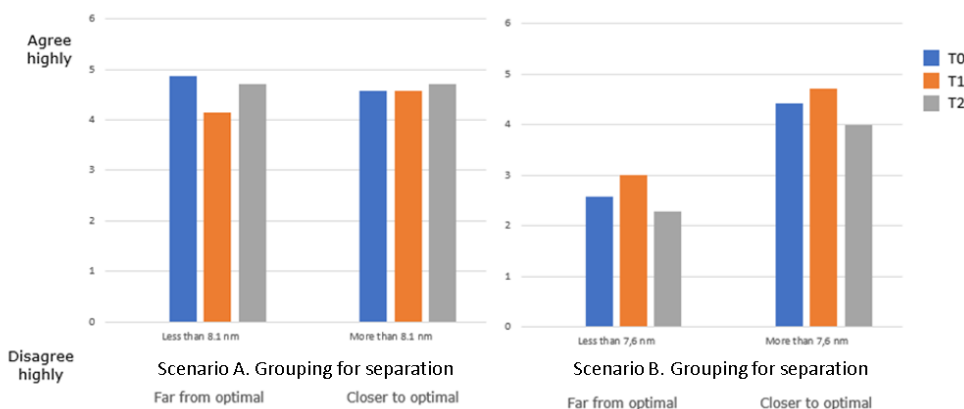


Figure 38. Reported similarity with own solution strategy, SIM2B.

The “more than 8.1 nm” and “more than 7.6 nm” group had a personal model aiming for a separation margin of less than 9 nm, which was closer to the optimal model that had a corrected separation target of 10.7-10.8 nm and 10.3-10.6 nm, respectively.

Statement 2: Understanding of solution

In SIM2A (Fig 39), ratings of understanding advisory were overall high, indicating that participants understood why the advisory was proposed. There is no apparent difference across conformance and transparency levels.

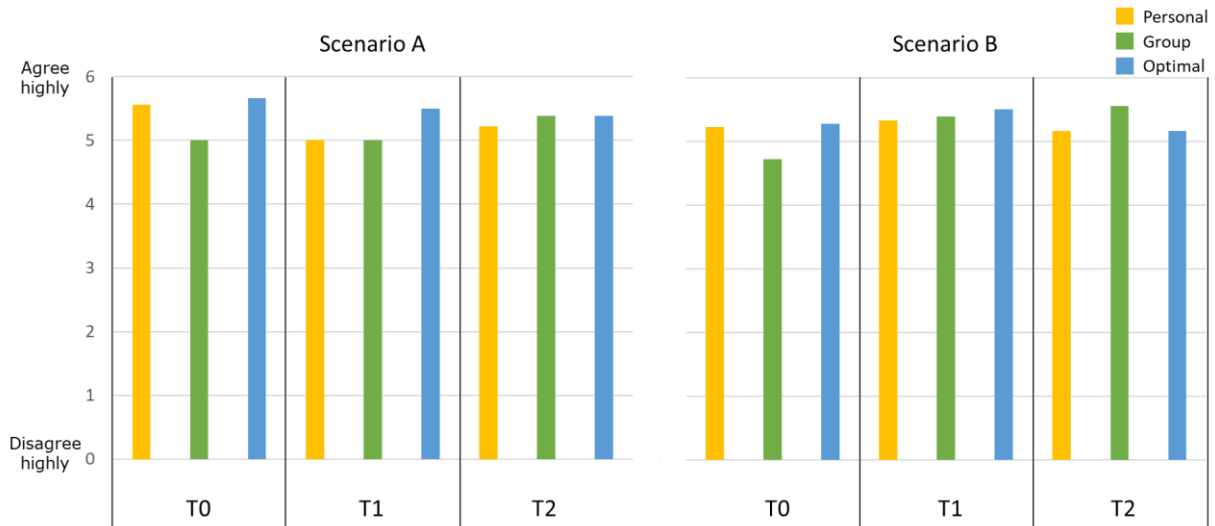


Figure 39. Reported understanding of advisory across conformance and transparency conditions in scenario A and Scenario B, SIM2A (n=18).

In SIM2B (Fig 40), ratings of understanding advisory were overall high, indicating that participants understood why the advisory was proposed. A difference can be observed in Scenario A where group advisories received lower ratings compared to personal and optimal advisories across transparency conditions. In Scenario B it is instead optimal advisories that received lower ratings. These patterns match those observed with ratings of how similar the advisory was perceived to be that of a given participant’s own solution strategy.

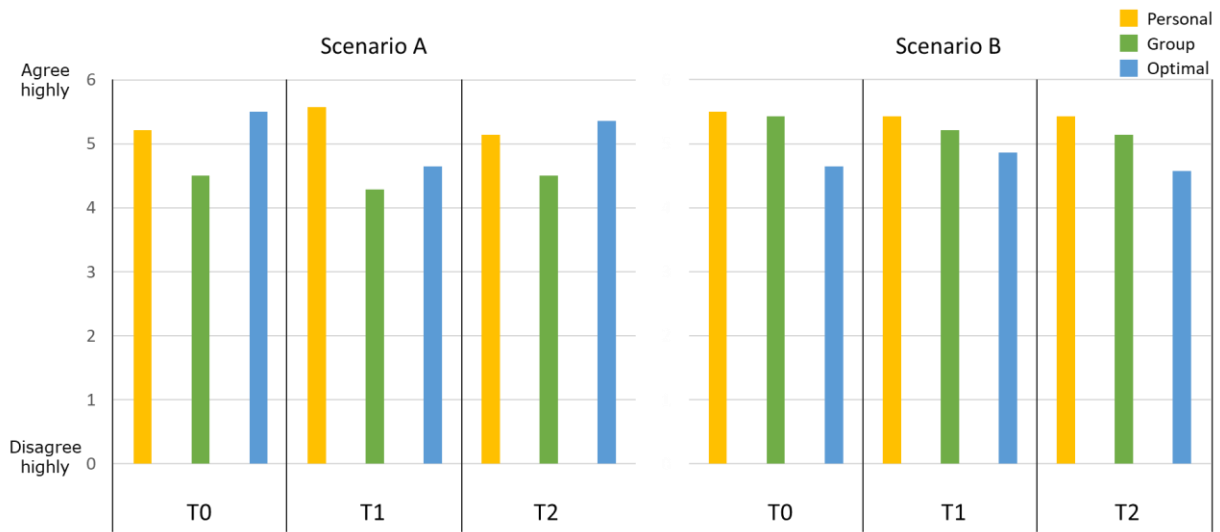


Figure 40. Reported understanding of advisory across conformance and transparency conditions in scenario A and Scenario B, SIM2B (n=14).

In Fig 41 and Fig 42 we can see that in scenario A and B in SIM2A and scenario B in SIM2B, participants in the group with an average separation margin closer to the optimal advisory (“less than 9 nm” and “less than 7.7 nm”, respectively for SIM2A, and “more than 7.6 nm” in scenario B for SIM2B), rated their understanding of advisories to be higher compared to the groups whose average separation margin was further away from the optimal advisory. Again, this effect was apparent with higher transparency levels (diagram and text). Note that scenario A in SIM2B does not show the same pattern.

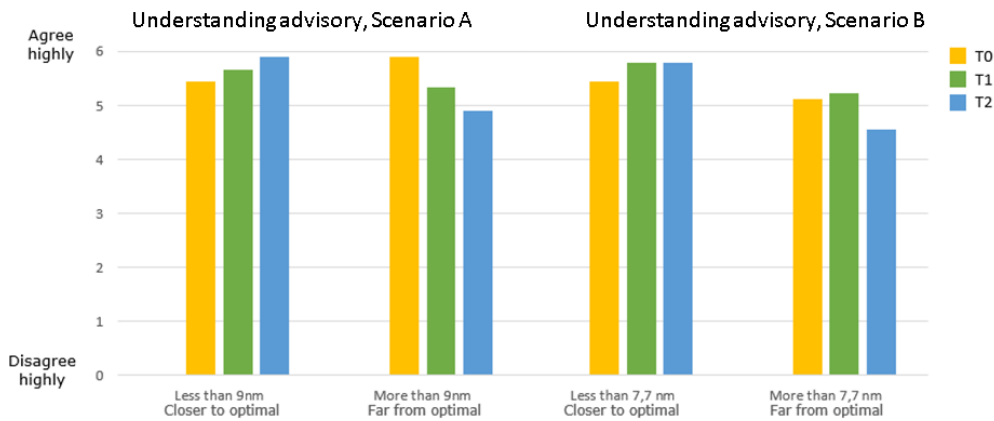


Figure 41. Reported understanding of solution, SIM2A.

The “less than 9 nm” and “less than 7.7 nm” group had a personal model aiming for a separation margin of less than 9 nm, which was closer to, or less than the optimal model that had a corrected separation target of 6.6 nm and 7.7 nm, respectively.

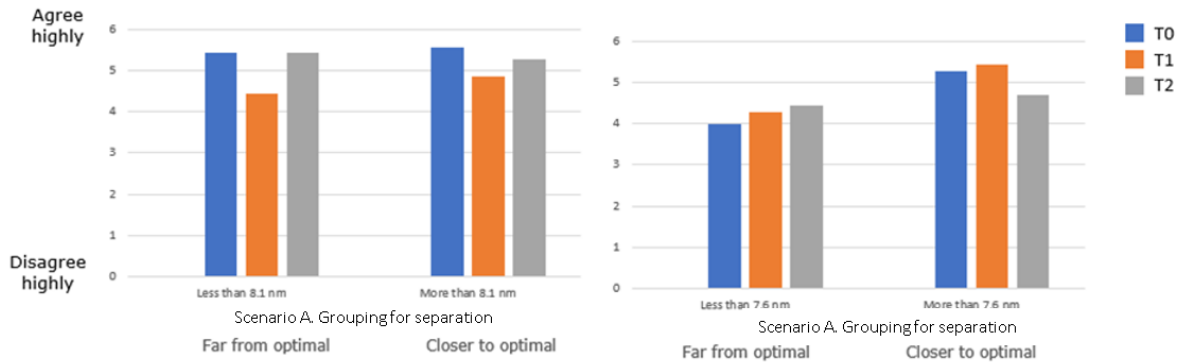


Figure 42. Reported understanding of solution, SIM2B.

The “more than 8.1 nm” and “more than 7.6 nm” group had a personal model aiming for a separation margin of less than 9 nm, which was closer to the optimal model that had a corrected separation target of 10.7-10.8 nm and 10.3-10.6 nm, respectively.

Post-session questionnaires

Post-session questionnaires were administered three times per participant, once after each Transparency session. Participants indicated agreement (on a 1-6 scale, from “Highly Disagree” to “Highly Agree”) with 13 statements.

Following are the results of post-session questionnaires, collapsed across simulations 2A and 2B. The following figures show absolute number of responses and can also be read (n=102) as approximate cumulative percentages.

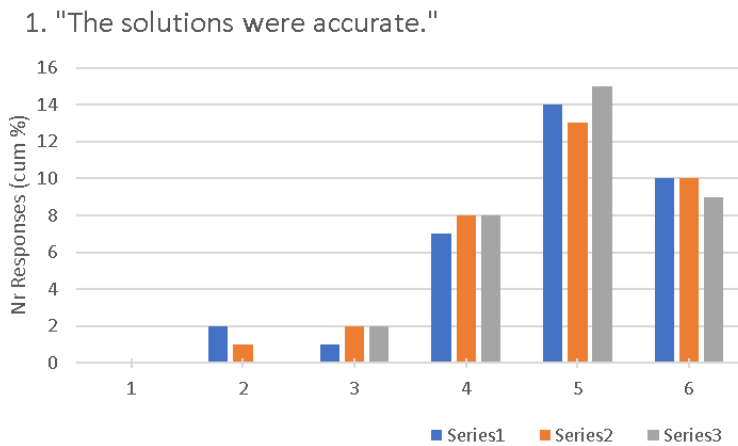


Figure 43. Post-session questionnaire item 1: Solution accuracy.

Controllers overwhelmingly agreed that solutions were accurate (figure 43, showing Question 1 results). In terms of a binary split (ratings 1-3 vs 4-6) 94% agreed and 6% disagreed. No transparency trend was apparent.

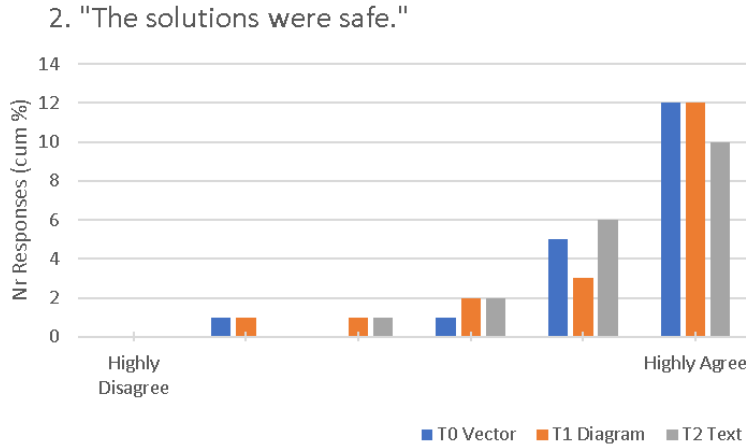


Figure 44. Post-session questionnaire item 2: Solution safety.

Controllers also tended to agree with the statement that solutions were safe (Figure 44, Question 2). 97% were in binary agreement with the statement, and a majority (60%) highly agreed. Again, no transparency trend was clear.

Combining responses to Questions 3, 4, and 5 (shown in figures 45-47 respectively), controllers tended to agree that solutions were efficient. They also tended to agree with system solutions generally, even though those solutions were different from the ones they would have generated themselves.

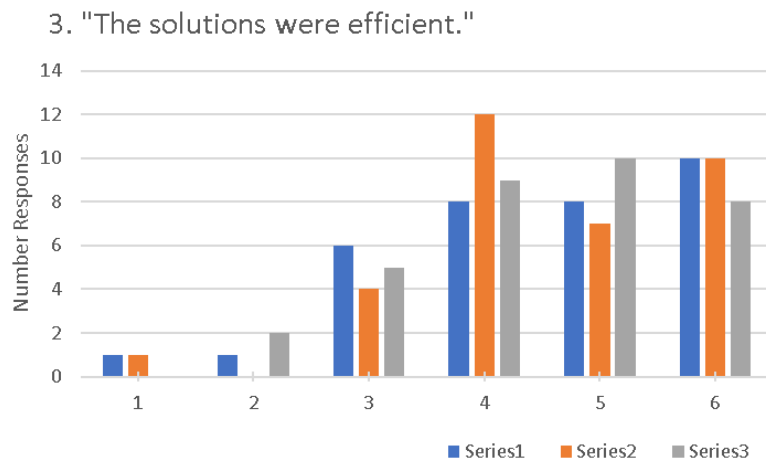


Figure 45. Post-session questionnaire item 3: Solution efficiency.

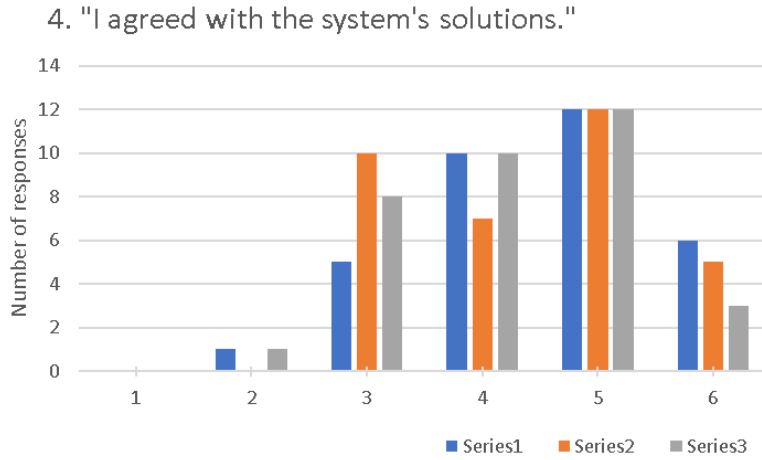


Figure 46. Post-session questionnaire item 4: General agreement.

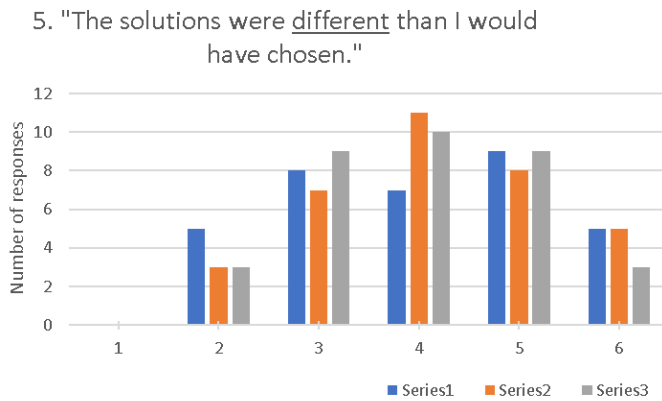


Figure 47. Post-session questionnaire item 5: Solution difference.

As shown in figure 47, 67% agreed that system solutions were different than those they would have chosen themselves. As shown in figure 48 (Question 6), 74% of controllers disagreed that the system solutions were better than the ones they would've chosen himself. However the disagreement tended to be fairly weak, with only 18% highly disagreeing with the statement.

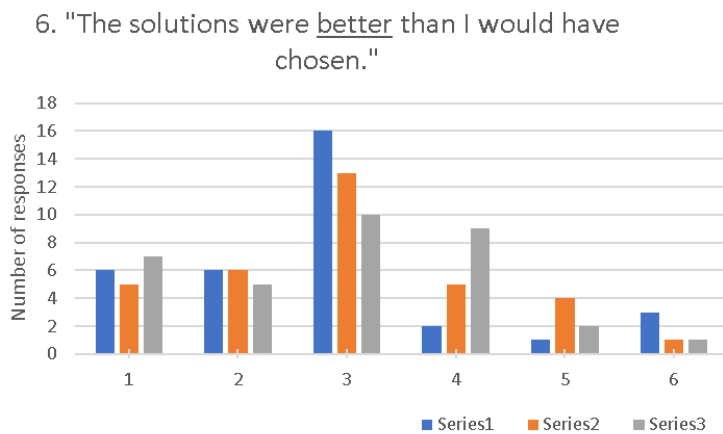


Figure 48. Post-session questionnaire item 6: Solution superiority.

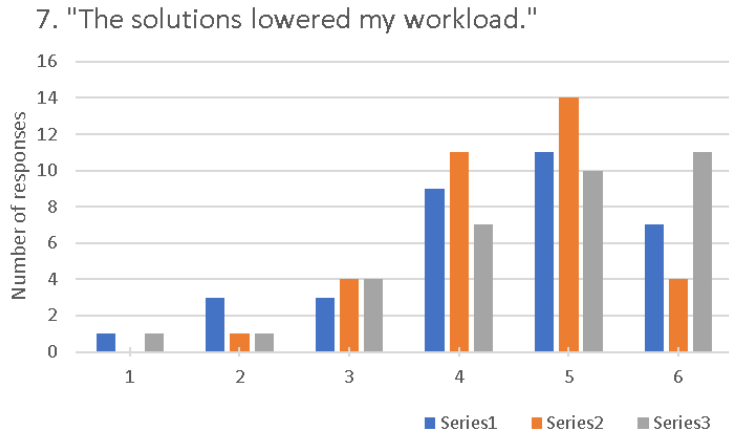


Figure 49. Post-session questionnaire item 7: Lower workload.

As shown in figure 49 (Question 7), controllers highly agreed that system solutions lowered their workload, with 84% rating their agreement four or higher.

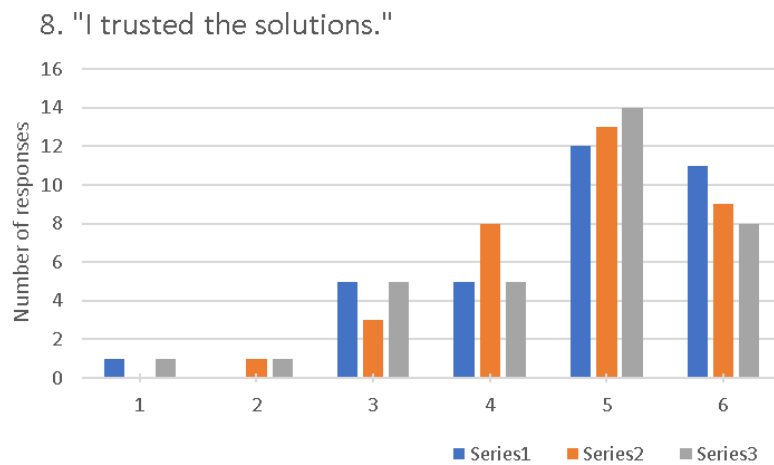


Figure 50. Post-session questionnaire item 8: Trust.

As shown in figure 50 (Question 8) controllers also reported highly agreeing with the statement that they trusted the system solutions. In a binary split almost 85% reported agreement.

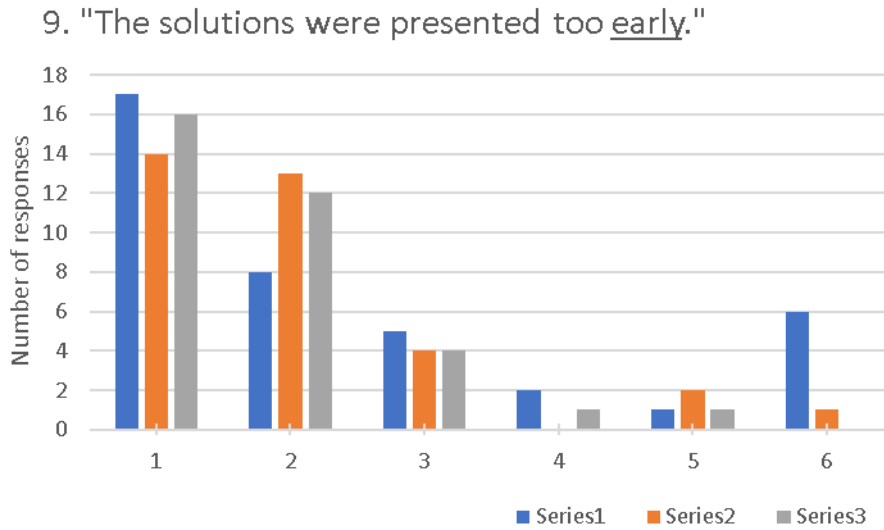


Figure 51. Post-session questionnaire item 9: Solutions too early.

Questions 9 and 10 concern the timing of solutions, and when there they were presented either too early or too late. Generally speaking, most controllers highly disagreed with the statements that solutions were either too early or too late. Notice that there would be no reason to expect transparency effects in Questions 9 or 10 since conformance (and thus for example the presentation of personal versus group versus optimal model scenarios, which can have systematic differences in solution timing) was varied within each session.

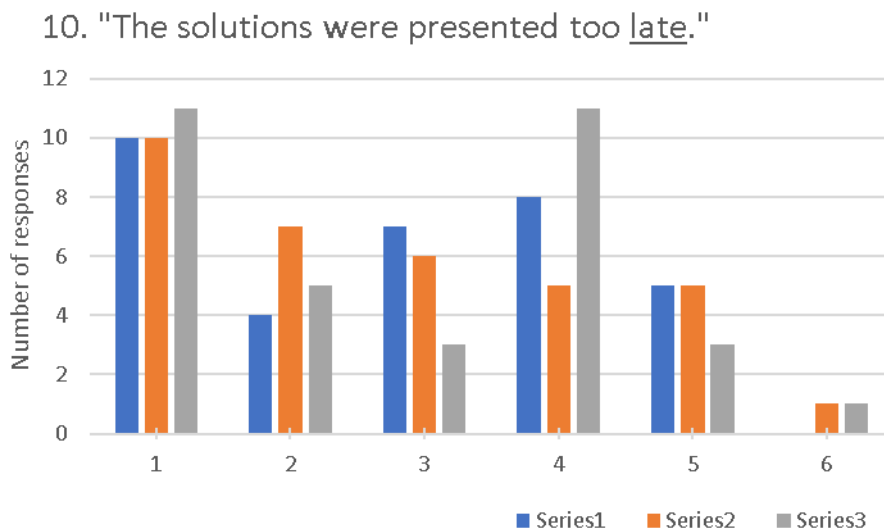


Figure 52. Post-session questionnaire item 10: Solutions too late.

11. "The solutions helped me resolve conflicts quicker."

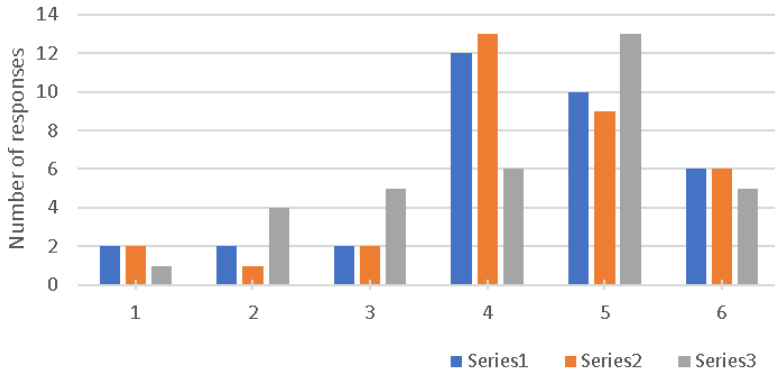


Figure 53. Post-session questionnaire item 11: Quicker resolutions.

Question 11 (figure 53) showed that controllers highly agreed with the statement that solutions help them resolve conflicts more quickly, with 80% above the agreement midpoint.

12. "The system was easy to use."

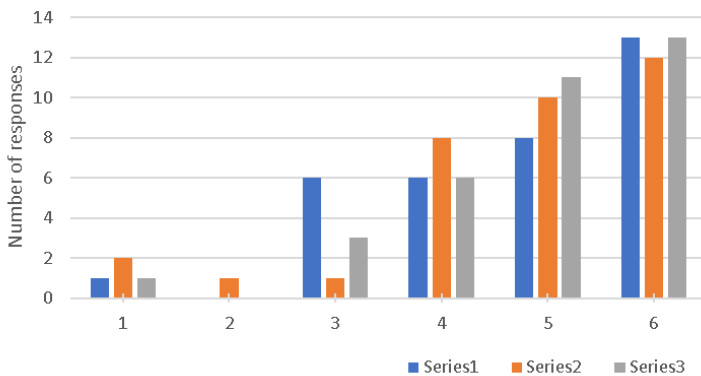


Figure 54. Post-session questionnaire item 12: Ease of use.

13. "The presentation format made it easy to understand the solution."

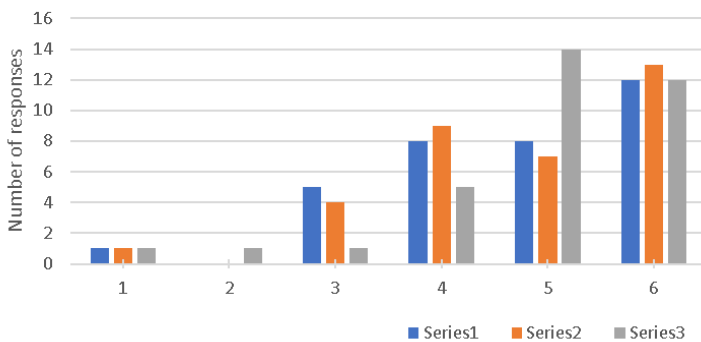


Figure 55. Post-session questionnaire item 13: Understandable format.

Question 12 showed overall high agreement (87% above midpoint) that the system was easy to use. Controllers were also in general agreement (Question 13) that presentation format made it easy to understand the solution. Transparency effects are not clearly apparent in these data.

Exit Questionnaires

Each participant completed exit interview questionnaires after completion of all three simulation sessions. Exit questionnaires consisted of seven agreement-scale items. Each item instructed the respondent to indicated agreement with the given statement on a scale of 1-6. Responses are shown below.

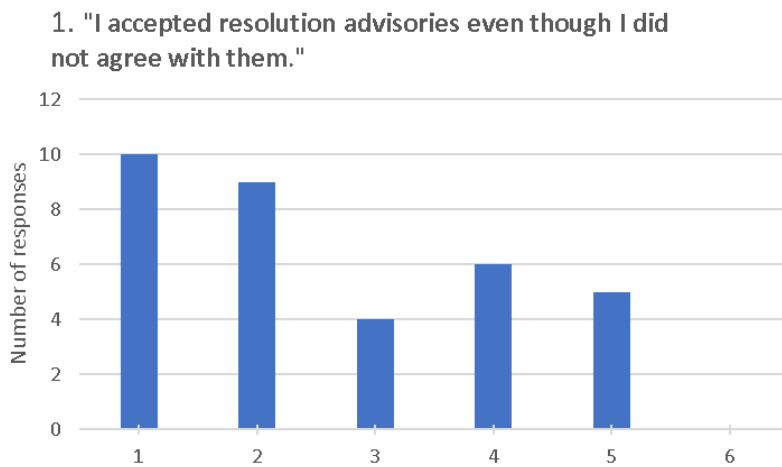


Figure 56. Exit questionnaire item 1: Accepted without agreeing.

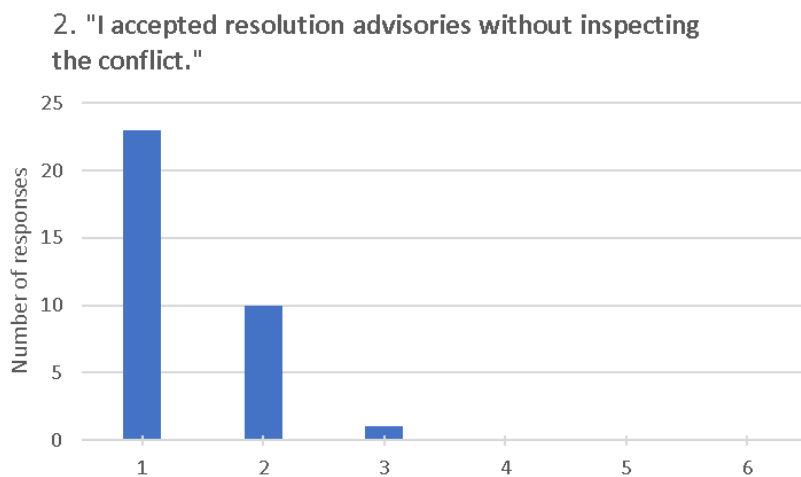


Figure 57. Exit questionnaire item 2: Accepted without inspecting.

3. "In the future, computers will do more and more of the controller's job."

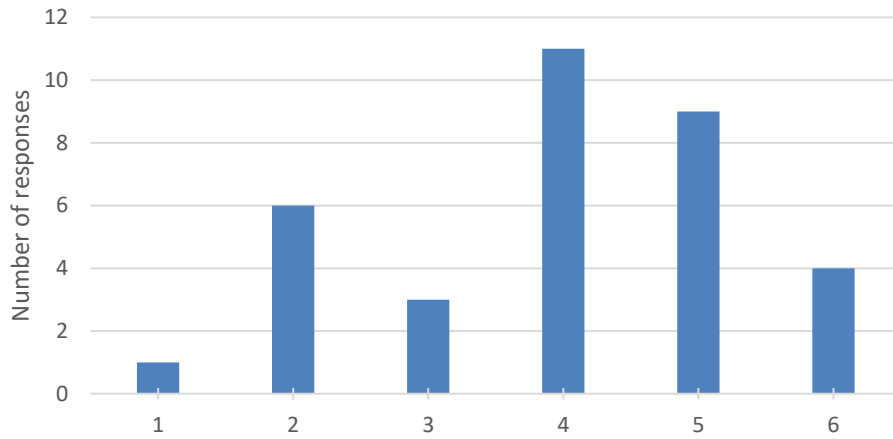


Figure 58. Exit questionnaire item 3: Computers will do more.

4. "In the future, computers might be able to perform ATC conflict resolution as well as I can."

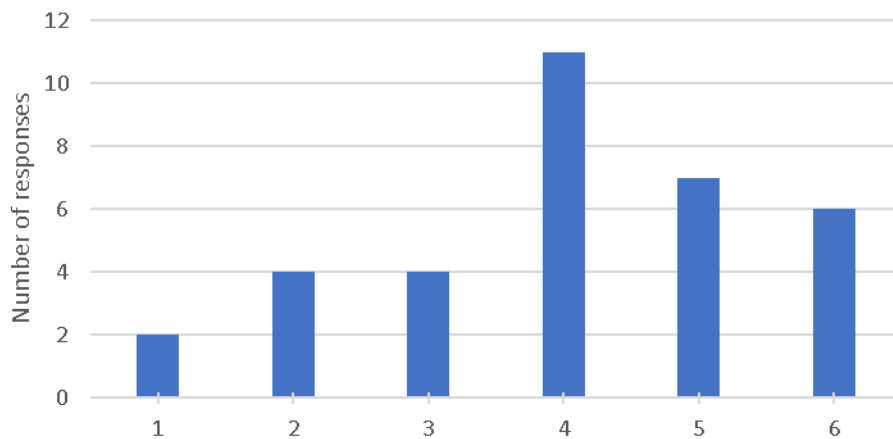


Figure 59. Exit questionnaire item 4: Computers will equal me.

5. "A system like this would make my job less rewarding."

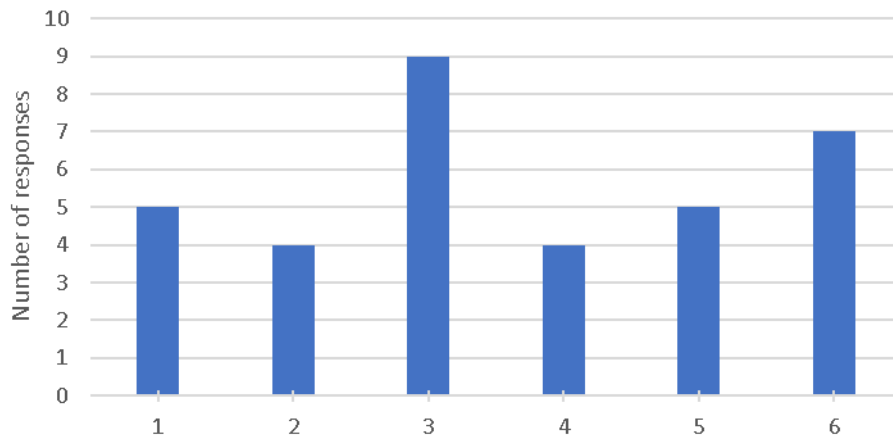


Figure 60. Exit questionnaire item 5: Less rewarding job.

6. "There is generally more than one acceptable solution to an air traffic conflict."

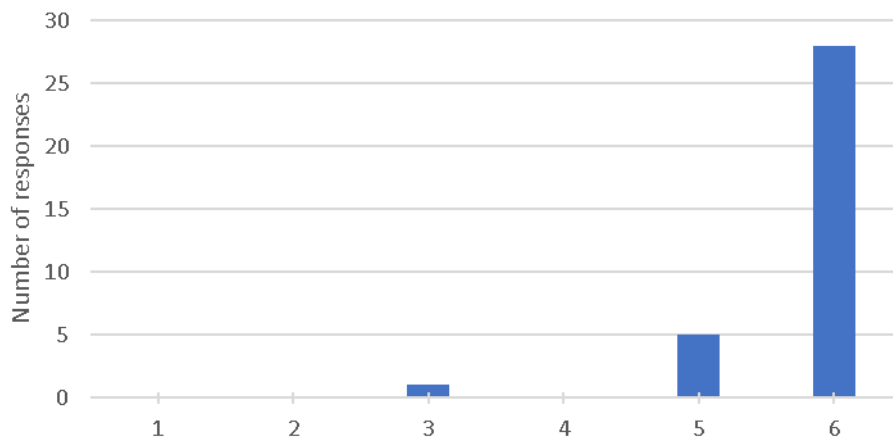


Figure 61. Exit questionnaire item 6: More than one solution.

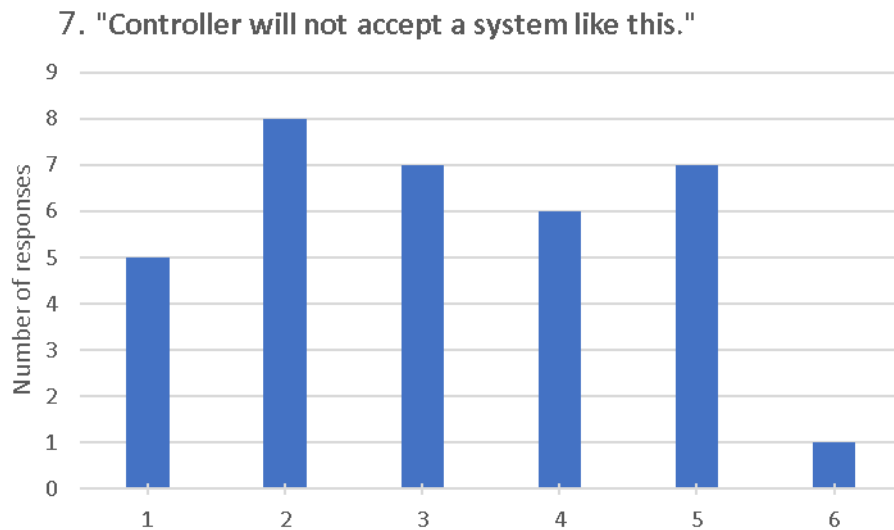


Figure 62. Exit questionnaire item 7: Controllers' acceptance.

The exit interview also included two conflict resolution exercises. Participants were shown static images of enroute conflict situations (items 8 and 9). Controllers were asked to indicate (a) their preferred strategy for resolution, (b) why they would choose this strategy, and (c) whether they would solve this conflict the same way every time. Overall, 64% of controllers indicated that they would repeat the same resolution, and this was fairly similar between simulation sites (68% vs 57% for sim2A and 2B). Controller responses were not probed further, so it is not known how controllers might vary in their solution strategies in future.

3.2.4 Debrief Sessions: General Impressions

Debriefs were conducted at the end of each participant's main experiment session. Debriefs consisted of open-ended questions, with prompts intended to stimulate discussion and elaboration. General impressions from the debrief sessions are organised according to the general themes that emerged.

Simulation realism

Participants generally reported that they found the simulation realistic at a high level. However, controllers also reported several specific aspects that they found unrealistic:

- lack of wind effect
- no airline specific information and data labels
- no ground speed information available
- lack of airspace outside the simulated sector
- octagonal sector shape
- no destination information available in labels
- flight legs that were unrealistically simple
- insufficient performance differences observed between aircraft.

Controllers also noted that the simulated conflict patterns were highly unlikely. Conflicting aircraft at the same flight level on converging courses would be very rare, because in real life aircraft would be separated by flight level. One controller also noted that a closest point of approach (CPA) of zero would be very rare.

Some controllers also reported gaming the simulation and becoming less attentive in their responses. As one controller put it

“I focused attention on the conflict, leaving out the rest. But I don't think such an object would make you lose focus. It was my approach”.

Display (i.e. transparency condition) preference

Display preference varied across controllers, with a slight preference for the diagram over the text display. Some expressed concern that the text display presented too much (or unnecessary) information. A few controllers did, however, feel that CPA was a useful field to display in the text presentation. In summary, there were individual differences in preference but for those who had a clear preference, across controllers the order of preference was Diagram, Text, and Vector.

Conformance

Although both transparency and conformance were manipulated within participant, participants were not aware of the conformance manipulation within a given transparency session. Therefore we could not get direct feedback on preferences regarding the personal versus group versus optimal models. Nonetheless, in the course of discussion some relevant points were made with respect to the match between human and machine CD&R strategies.

It seems that “incorrect” aircraft choice drove some participants to view some solutions as suboptimal. This suggests that the choice of aircraft is an important one in defining conformance. Another factor was a perceived tendency by some controllers for the system to act too aggressively and issue too great a heading change clearance, which also drove them to perceive the system as suboptimal.

Broader controller acceptance of advisory systems

Controllers were asked to think generally about advisory systems like they had just seen (but to disregard this specific simulation and its interface), and to speculate on how they think controllers in general would accept such automation in the future.

Controllers’ views on general acceptance often came down to the issue of trust and whether controllers could develop a working relationship with such advisory automation. A few controllers noted that at anything other than a zero error rate, trust would degrade immediately across controllers, and that even a single failure would mean the trust would become unrecoverable. It was also noted that introduction of such advisory automation would have to be very gradual and step-by-step to overcome skepticism. Further, controllers would be unwilling to just passively monitor an advisory system, without any active role to play.

4. Discussion

4.1 Pooled data vs fine-grained analysis

The observed nonnormality seen in the pooled data can be linked to a few interrelated issues, as follows:

- controller variability-- Differences both within and between controllers presented challenges to the field study methods. For example, if within-controller consistency (i.e., whether a given controller solves the same conflict the same way every time) is low this complicates the definition of a personal model. If between-controller consistency (i.e., whether controllers as a group tend to choose the same solution) is low this complicates creating a robust group model.
- The data demands of ML training, in particular the relatively small number of training samples, meant that the personal models were built on only 36 scenario interactions per participant in the training pre-test. In the end, this led to a synthetic scenario creation process for the Personal model (based on analysis, rather than an actual trained model). When pooled, however, training pre-test data did provide a sufficient number of data samples to train the Group model to the point of stabilized performance. However, this means that the Personal and Group models were in fact qualitatively different and built on different processes.
- Experimental simplifications-- given the above data demands, and the desire for experimental control in field simulations, certain concessions had to be made to operational realism. One example of this was limiting ML to heading solutions only. For those controllers who solved conflicts using altitude, this also complicated creation of personal models (a personal model of heading only solutions would be a poor fit for a controller who prefers altitude solutions).
- scenario effects-- a great deal of pre-processing and analysis went into selecting a subset of the six training scenario into two (A and B). One of the main drivers of this selection was controller consistency, given the critical role it plays in creating robust personal models. Having said that, post simulation analysis revealed differences in transparency and conformance trends across the two scenarios, and there are a few reasons to suspect systematic differences between the two scenarios. In short, it has to be recognize that any traffic scenario carries with it a specific context that makes it qualitatively different from the next.

4.2 Challenges in comparing personal and group models

As noted, creating personal models faced at least a few challenges. Primary among these is that robustness of the personal model is highly dependent on the internal consistency of that controller in choosing solutions. For some controllers, it was observed that the personal model was a poor match, for example when a given controller had a tendency to prefer altitude solutions (see ANNEX C).

In addition to challenges in creating robust personal models, the experimental design also faced challenges in comparing across models. Since between-controller variability (in terms of solution strategy) was high, personal models differed in their similarities and differences to the group and optimal models. Some personal models were a good fit to the other model types, but others were notably poor.

4.3 The CD&R context

A broader question is whether enroute CD&R represents a compelling use case for ML. This issue is reflected in the challenge the team experienced in defining transparency, particularly at the highest level. Notice that the vector versus diagram versus text manipulations are qualitatively different. Moreover it is not clear that CD&R is a context in which adding text-based rationale for advisories necessarily added beneficial transparency.

4.4 The benefits of transparency

Further, transparency seemed to have a different effect across scenarios A and B. An implicit assumption going into this research was that Transparency fosters understanding, acceptance and agreement. As a thought experiment however, consider the case where poorly functional automation is outputting advisories. In this case, transparency might have the opposite effect and lower controllers' agreement and acceptance of the system.

The notion here is that if transparency involves making clear to the operator the inner workings of the algorithm, it does not necessarily increase agreement and acceptance, but should optimize them. Transparency and explainability should increase acceptance and agreement for an optimal algorithm, which should also decrease acceptance and agreement for a sub optimal algorithm.

4.5 On personalization and optimal systems

Although personalization of ML systems is held as a positive goal, there is one potential challenge that we need to consider. Namely, attempts to personalize advisory systems introduce the risk that they drive the operator to solve the problem in a particular way. For example, the simulated advisories aimed to solve en route conflicts using a single intervention with only one of two involved aircraft. This approach is inconsistent with controllers who solved the conflict with two interactions (for example, slightly turning both aircraft). It should be noted that the way advisories are framed can give a suggestion for how the system proposed to solve a given conflict, and offers an implicit reference against which controllers' judgment and decision is formed. Without an advisory system the controller would search for information and cues with regard to traffic pattern, speeds, altitude etc. in deciding how to solve a conflict. Past research has noted that advisory systems can have the unintended consequence of increasing task load. The notion is that whereas a current controller has to devise a solution, under an automated advisory system that controller has the additional task of processing the advisory, and comparing that to their own strategy.

How do we define optimal? One potential irony is that attempts to achieve the optimal solution in all situations may prove sub optimal. What is considered optimal from an airspace and traffic geometric standpoint might run counter to human performance demands. If we are to design for human performance demands, the optimal solution might be one that best manages the controller's workload and situation awareness.

4.6 Challenges in training ML

It is a widely known problem in ML that enormous amounts of data are required to train such models to the point of stabilization. Further, RL models add the additional challenge that there is still a great deal of artistry required in designing the reward structure. How a designer constructs this reward structure has a large impact on the performance of the eventual model. As an extreme example, if an RL reward structure is built that heavily penalizes loss of separation, the model might simply learn to turn aircraft 180° once they enter a sector, to avoid loss of separation. The point here is that structuring RL models and their goal trade-offs, is one of the most critical elements of designing and tuning an RL model.

4.7 Experimental control vs operational realism

MAHALO field simulations used a very abstracted airspace and simplified task. The main reasons for this included experimental control, and narrowing of the context so as to facilitate ML training on a simplified task. Again, controllers are accustomed to dealing with wind effects, the hemispherical rule, and other aspects of their operational reality which field simulation did not capture. One interesting example was feedback received from more than one controller that a system that provides early conflict visualization prior to solution would be preferable. Notice that at the SIM2B site, controllers operationally use the CARD system that presents potential conflicts. In MAHALO, the advisory was presented without information about conflicts being detected.

4.8 Conformance and the potential importance of personalization

Agreement of ‘optimal’ advisories is very dependent on conformance with personal strategies. We saw that ATCOs whose preferred CPAs were farther from the proposed optimal solutions tended to disagree more with those advisories. This likely points to a broader trend, beyond only CPAs. Controllers who, for example, prefer quicker solutions might also tend to disagree with solutions later than their own.

4.9 Noise in decision making

When discussing bias and noise in decision making where the parameters underlying the decision task are vague or unknown, a challenge lies in figuring out what parameters to look at and how to measure the resulting decision. This is the case in ATC CD&R. In CD&R decisions, the outcome appears clear and unambiguous. Two aircraft previously on collision course (or predicted to lose separation to be precise) are, following the implemented decision, safely separated. But when looking at the details of what has happened, how do we classify the decision made? While the high level goal was to increase separation, a question remains why separation was sought in a particular way. Since separation horizontally is measured in nautical miles, it makes sense to conclude that the decision targeted a particular distance between the aircraft to be achieved. But it is not possible to solve a conflict by simply deciding on a separation distance. To achieve that, one or both of the aircraft have to alter its course. Which one? How much? When?

These are parameters that require an intervention in order for achieving adequate separation. Thus, it makes sense to measure these when trying to capture CD&R decisions. Moreover, additional parameters can be considered as part of trying to maximize efficiency if the solution. For instance

minimizing track deviation, fuel consumption, extra distance flown etc. The problem grows in complexity.

Another question is what the human operator solving these situations think is important. Does that person strive to achieve a certain separation distance? Are factors such as aircraft choice and timing important for this decision?

If we want to measure bias and noise in CD&R decisions we need to figure out how to classify or define the solution in a way that represents the operators goals. Looking at data from one perspective may lead us to believe that CD&R decision making is noisy. Looking at it from another perspective may lead us to believe that it is biased. Or likely both.

To measure bias, we need a reference against which it can be measured – that is, an objectively correct answer. In CD&R there is no objectively correct answer when dealing with human decision making where subjective factors such as workload, stress, and fatigue play a part. In contrast, noise can be measured regardless of a known target (or correct answer). But to determine noise, we need to know something about the scale at which data should be looked at. At a too close range, any data may appear noisy. At a too distant range, any data may appear coherent. Moreover, we need a lot of data, exactly how much is difficult to say, before we can draw conclusions about bias and noise.

That humans are noisy in their judgements and decisions is known. Experts are expected to be less noisy, but research show that this is not a rule. This is not necessarily a bad thing, even though there are many examples when it is. A benefit of algorithms and deterministic problem solvers is their noiselessness. A problem is sometimes bias when compared to what humans prefer (which ironically suggests that humans are not noisy). A weakness of algorithms and deterministic systems is their inability to adapt to new situations or changed circumstances and find more optimized solutions. This is where AI systems able to learn and adapt are expected to excel. A drawback of these systems is that they become more noisy, hence behaving more like humans.

The noise in data can also be traced to ambiguity in the problem to be solved. If there are conflicting cues preventing the formulation of a coherent interpretation of how to best solve the solution, we can expect noise. One such parameter may be aircraft choice in CD&R - one has to be prioritized over the other (I.e interacted with first) even if both can be interacted with in serial. And following the first interaction, the situation has slightly changed. Therefore it is not surprising that controlled are noisy in terms of aircraft choice. Depending on how different controllers view the situation, or the same person does over time, it may fall naturally to interact with different aircraft.

5. Conclusions

Analysis of conformance and transparency effects was challenged in the field study by the fact that scenario and simulation both emerged as extraneous variables that required separate analyses. Having said that, we can make certain broad statements about the impacts of conformance and transparency as they influenced the measures dependent variables.

First is the case of controller acceptance which was defined along five dimensions, and which was not subjected to statistical analysis. Even so, trends in the plots suggest that acceptance of personal and group advisories varied little across transparency levels. However, for optimal advisories a change in acceptance responses could be seen in the text condition compared to both the vector and diagram conditions.

In terms of controller agreement with advisories, results varied by both scenario and simulation. For both scenarios of SIM2B significant main effects of agreement ratings were found, and post hoc analysis revealed that: in scenario A, agreement ratings were significantly lower for the group model than for either the personal or optimal. In scenario B, agreement ratings were significantly lower for the optimal model than for either the personal or group models. In SIM2A Scenario A, a statistical main effect of conformance was found, and post hoc tests revealed that agreement was significantly higher for the optimal model than for either the personal or group model. This effect was not found for Scenario B of AIM2A.

For rated workload, a significant main effect of conformance was found, but only for SIM2A Scenario B. In this condition, the personal model was associated with significantly lower workload than was the optimal model.

5.1 Addressing the research hypotheses

This section addresses how well the field study answered each of the research hypotheses.

5.1.1 Relationship between conformance and acceptance / agreement

Hypothesis 1 stated that acceptance and agreement be higher for solutions that conform to the controller's preferred solution (i.e., the personal model). Conformance effects were found on agreement ratings, and inferential analysis of SIM2B showed that agreement was significantly lower for the group model in one scenario, and significantly lower for the optimal model another scenario. Conversely, SIM2A results showed a statistically significant higher agreement rating for the optimal model, in one of the two scenarios. These results underscore again the impact that simulation and scenario had as extraneous variables in the analysis.

In terms of the relationship between acceptance and conformance, neither the pooled data nor the fine grained breakout data present a clear picture. As shown in table 5, acceptance was very close across conformance levels. Acceptance (again, a five level scale) was however consistently lower for the group condition. Recall that the acceptance data were not analyzed using inferential statistics.

On balance, the data partially support Hypothesis 1 for agreement ratings data, but not for acceptance data.

5.1.2 Relationship between transparency and acceptance / agreement

Hypothesis 2 stated that controller acceptance of advisories will be higher if those advisories are presented in a high transparency display format. In terms of acceptance, descriptive data trends showed that whereas acceptance was very close across transparency levels, the text condition was associated with noticeably lower acceptance considering only full acceptance as the measure. This effect persisted but diminished as additional categories (nudge, adjust) were considered. Data trends also suggested the impact of simulation, scenario, and separation distance.

In terms of advisory agreement, no main effects reached statistical significance. However, simulation and scenario effects were again apparent. Moreover, data trends suggested an interaction between conformance and transparency (which approached significance for SIM2A Scenario B). For the group model, the text condition showed the highest acceptance whereas for the optimal model, the vector condition showed the highest acceptance. One possible interpretation of these data is that a conformance by transparency interaction would suggest that in terms of controller acceptance, the most appropriate level of transparency display type might vary with the type of underlying conformance model.

On balance, data do not support Hypothesis 2 regarding the relationship between transparency and acceptance / agreement; however, the data suggest that transparency might interact with conformance in driving acceptance. Finally, the impact of both simulation and scenario differences was apparent.

5.1.3 Relationship between workload and transparency / conformance

Hypothesis 3 stated that transparency and conformance manipulations would be associated with a change in reported workload. In terms of inferential analysis, a main effect of conformance showed that, at least for one Simulation / Scenario combination, the personal model produced significantly lower workload ratings than did the optimal model. Although other workload effects failed to reach statistical significance, data plots suggest strong differences between the simulation sites. At one site, the vector condition showed a noticeable decrease in reported workload, whereas at the other site the vector condition was associated with a reported workload increase.

On balance, hypothesis 3 was weakly supported, with the stipulation that strong simulation site differences might be influencing the results.

5.1.4 Interactive effects of Transparency and Conformance on acceptance and agreement

Hypotheses 4a and 4b together stated that conformance and transparency will interactively influence acceptance and agreement, such that

- under low transparency, personal advisories would be more accepted and agreed upon; and
- under high transparency, this effect would be less pronounced.

In terms of five-point acceptance, data trends suggest an interaction for optimal advisories. For both the personal and group conditions however, acceptance varied only slightly across transparency levels. Acceptance results, however, suggest an interaction trend between conformance and transparency. For the personal model, the vector display produced the highest acceptance. On the other hand, for the

group model the text condition showed the highest acceptance. For the optimal model, the baseline vector condition showed the highest acceptance. As stated earlier, one suggestion from these data is that the level of transparency, if it is to meet with controller acceptance, must be keyed to the level of conformance. On balance, the data do not support the directional interaction hypotheses as stated. Nonetheless, it is essential to keep in mind that transparency may not be a continuous construct (that can simply be adjusted up or down) but rather a condition that must be keyed to the system and conformance of an advisory system.

REFERENCES

- [1] van Rooijen, S.J. et al (2019). Conformal automation for air traffic control using convolutional neural networks. Presented at ATM 2019, Thirteenth USA/Europe Air Traffic Management Research and Development Seminar.
- [2] Eby, M. (1994). A self-organizational approach for resolving air traffic conflicts. Lincoln Lab Journal, 7 (2).
- [3] Westin, C. (2017). Strategic Conformance: Exploring Acceptance of Individual-Sensitive Automation for Air Traffic Control. PhD thesis. Control and Simulation Section, Aerospace Engineering Faculty, Delft University of Technology, The Netherlands. ISBN 978-94-6299-659-5.
- [4] MUFASA (2013). E.02.08 – MUFASA Final Project Report WP-E. SESAR SJU.

ANNEXES

ANNEX A: TRAINING PRE-TEST MATERIALS

- A1. Experiment briefing
- A2. Introduction
- A3. Consent form
- A4. Demographics questionnaire
- A5. Training instructions
- A6. Debrief

A1. Experiment Briefing

Read to participant

Scenarios

In today's simulation, you are asked to play a lot of short scenarios acting as an air traffic controller. You will work with the ATC simulator SectorX. Each scenario is 2 ½ minutes. You interact with aircraft using the mouse and keyboard. All scenarios display a hypothetical (not real) sector in an octagon shape, about 100 x 100 NM in size. You are controlling RVSM airspace, covering FL290-410. The radar screen is updated in 2.5 second intervals. The simulator runs at 2x time of speed. This means that aircraft move 2 times faster than normal, which affects aspects such as the time lapsed for each scenario and rate of climb and rate of descent.

Your task

Your task is to ensure separation between aircraft and make sure aircraft leave the sector through their assigned exit waypoint at the correct flight level as indicated by their flight plan. **Each task is equally important.** The Exit Points are located on the perimeter of the sector and are shown as triangles with five-letter names around. The designated Exit Point of each aircraft is shown in the label of each aircraft. You have a printout of what aircraft labels look like and what information they contain, depending on the state of the aircraft.

From time to time there may be conflicts between aircraft. A conflict occurs when aircraft are predicted to close within 5 NM and 1000 ft of one another. When the system predicts a loss of separation within the next 120 seconds, a short-term conflict alert (STCA) is provided. Because the simulation is played at 2 times faster than normal, this means 60 seconds prior to loss of separation.

Note that:

- Flights do not have a departure or destination aerodrome.
- The rules for semicircular cruising levels are not applied.
- You are not able to coordinate with adjacent sectors.
- Communication with aircraft is carried out by means of data link CPDLC, so there is no verbal communication with pilots.
- You can instruct aircraft to change heading and/or altitude by interacting with aircraft through their labels. Executed changes will be implemented by the aircraft right away.

So, most importantly, make sure that:

- aircraft are safely separated: 5 NM horizontally and 1000ft vertically.
- all traffic leaves the sector at their assigned exit waypoint and their assigned exit flight level.

Training

- You will first be trained on how to use SectorX. The training consists of three parts. The first two parts are intended to build your knowledge and skills in using SectorX. The last part is a test of your ability to understand information and use SectorX.

A2. Introduction

1. Welcome and Presentation of researchers

2. Short presentation of MAHALO

Artificial Intelligence (AI) is becoming more common every day, across many applications. MAHALO (Modern ATM via Human/Automation Learning Optimisation) is a SESAR research project exploring possible application of AI in ATC. Specifically, MAHALO is exploring how AI *explainability* (i.e., how well can the controller understand AI inner processes) and *conformance* (i.e., how much does AI have to match the controller's own style) impact AI use in ATC.

As part of this effort, MAHALO is conducting a series of human-in-the-loop simulations, in which controllers are being asked to control a series of short air traffic scenarios.

3. Presentation of simulation

We have asked you here to take part in an ATC simulation. This simulation is realistic in some respects, however we have simplified several aspects of the ATC task. Today is the first phase of a two-step simulation. The second phase will take place in late April under week 17.

Your participation today is expected to last about three hours total, including three short breaks. After reading through the simulation briefing, signing the consent form, and completing a short questionnaire, you will spend around 25 minutes doing a training session. The goal of this training is for you to get to know the simulator and learn how to interact with it. During this training, I will sit next to you providing support and answering any questions you might have.

When you feel comfortable using the simulator, we will run three measurement sessions, each 30 minutes long. In these three sessions, you will see a number of short (2 ½ minutes) en-route air traffic scenarios. Once the simulation is completed, we will have a short debriefing, to ask some general questions about the simulation and the scenarios.

WE want to stress that we are NOT judging individual controllers, nor will we collect any identifying information. Your participation is both voluntary and anonymous.

Do you have any questions at this point?

4. Agenda

- Consent form
- Demographics Questionnaire
- Simulation briefing
- Training session
- Measurement session
- Debriefing

A3. Consent form

Information sheet ex Art. 13 of the European General Data Protection Regulation n. 2016/679

You are being invited to take part in a research study forming part of the MAHALO project. MAHALO Consortium would like to process your personal data in order to carry out its research activities.

Before you decide to give your consent to the processing of your personal data, it is important that you receive and understand all the relevant information about the processing of your personal data, in a transparent, intelligible, clear form. Please take time to read the following information carefully. If there is anything that is not clear, or you would like more information, please get in touch with the Project Coordinator (contact details are provided below). After having read and understood the following information, please feel free to give your consent to the processing of your personal data.

In accordance with Article 13 of the European General Data Protection Regulation n. 2016/679 (GDPR), MAHALO is committed to provide you with any information about the lawful processing of your personal data, in full respect of the principle of transparency.

With reference to the EXPERIMENTS you are about to participate in, we inform you that:

1. The Project Coordinator is Stefano Bonelli (Deep Blue s.r.l., Via Manin 53, 00185, VAT: 06458931000). You can contact the Project Coordinator at the following email address: stefano.bonelli@dblue.it.

The Project Coordinator is defined as the natural or legal person, public authority, agency or another body which, alone or jointly with others, determines the purposes and means of the processing of personal data.

2. For any matter related to personal data protection, please contact the Data Protection Officer (DPO) Stefano Bonelli at the following email address: stefano.bonelli@dblue.it.

3. The purposes of the processing of your personal data are:

- recruitment in the experiment (common personal data)
- carrying out the experiment (common and special categories of personal data)
- analysis of data collected (common and special categories of personal data)
- dissemination (common and special categories of personal data)

4. The MAHALO consortium will process personal data provided by you. The provision of your personal data is necessary for your participation in the experiment. Your refusal to provide data will not allow you to participate in the activities.

You will be asked to provide personal data such as first name and last name, date and place of birth, years of working experience, ID number (passport / driver license), ID valid until data, email address. In the performance of the experiment, other categories of personal data of yours will be processed: performance data, eye-tracking data, observation of behaviours and actions will be collected and recorded by a Subject Matter Expert (affiliated to one of the MAHALO partners).

You will be asked to answer questionnaires and interviews, with only audio recording and notes taken by the interviewer.

The experiment may be video recorded, and photos may be taken for dissemination and communication purposes.

The experiment may collect interaction logs and performance data from the simulation environment.

Your personal data will be collected and handled in paper or digital format.

5. Your personal data will be only processed on the basis of explicit consent, given specifically for each of the above-mentioned purposes. You have the right to withdraw consent at any time, without affecting the lawfulness of former processing.

6. Your personal data could be transmitted to all the members of the MAHALO consortium. If necessary, these subjects will be appointed as personal data processors.

Member from the MAHALO consortium will not disclose your data to any other parties, under any circumstances.

7. Your personal data will be processed by authorised and duly instructed subjects, able to ensure the safe and lawful processing of your personal data. These persons authorised to process personal data will also be bound by full confidentiality.

8. The processing of your personal data is based on the principles of correctness, lawfulness, transparency and minimization.

9. The personal data provided will only be stored for the time needed to fulfil the purposes they are collected and processed for.

After the experiment, name and email address of the participants are recorded on the simulation roster and schedule, so they can typically be traced back. Roster and schedule are only available to the simulation manager that has to respect a strict confidentiality commitment.

Your common personal data will be stored until the end of MAHALO project (December 2022).

Audio/screen recordings, and information about yourself will be treated as confidential by the MAHALO Consortium. These recordings will be stored until the end of MAHALO project (December 2022).

10. As a data subject, you have the right to request from the MAHALO consortium access to and rectification or erasure of personal data or restriction of processing your personal data as well as the right to data portability; where the processing is based on consent, you have the right to withdraw your consent at any time, without affecting the lawfulness of processing based on consent before its withdrawal; you also have the right to lodge a complaint with a supervisory authority.

Informed Consent Form - EXPERIMENTS

I, _____, Born in _____ on ____/____/____

- declare that I have carefully read the above information sheet, that I have fully understood and accepted its content.

_____ on ____/____/____

Signature of the Data subject _____

- give my explicit consent to the processing of my personal data, also belonging to the so-called special categories of personal data, in particular, eye-tracking data, according to the above information sheet.

_____ on ____/____/____

Signature of the Data subject _____

- give my explicit consent to the processing of my personal data, also belonging to the so-called special categories of personal data according to the above information sheet.

_____ on ____/____/____

Signature of the Data subject _____

- give my explicit consent so that short extracts of a video or photographs, in which what I say or what I do cannot be precisely determined and that cannot in any way damage my reputation, may be used by the MAHALO consortium for dissemination and illustrative purposes of the research results, according to the above information sheet.

_____ on ____/____/____

Signature of the Data subject _____

- request to be included in the newsletter service and to this end I attach my email address and give my consent to the processing of my personal data for this purpose, according to the above information sheet. Email _____

_____ on ____/____/____

Signature of the Data subject _____

A4. Demographic questionnaire

Participant ID: _____ (e.g. P1)

Date: _____

Personal information.

Age: _____

Which ratings do you have/have you had?

En-route TMA Tower Procedural Other

What is your operational experience as an air traffic controller?

Years: _____ Months: _____

Please indicate your experience per rating in years.

<i>Rating</i>	<i>Years</i>
En-route:	_____
TMA:	_____
Tower:	_____
Procedural:	_____
Other:	_____

How much in % of full time have you been working this past year (i.e., 2021)? _____

A5. Training instructions

MAHALO TRAINING Session SectorX			
1 STEP 1: TRAINING WALKTHROUGH			
Time	Scenario	INSTRUCTIONS to Instructor	INSTRUCTIONS to Participant on how to interact with the system
00:10:00	TW_1	Show participant label image printed on paper (can be placed next to workspace). Show what different colours mean.	
		Start training by typing in TW . Provide instructions as the scenario plays out.	Move the CONFLICT ALERT MESSAGE Window to the right as it obscures the VERA tool window .
			The time at the top of the interface shows the time in the scenario.
		Present the Radar menu .	Please look at the rows on the top. Our focus is the 3rd row starting with MARKER button. It's called Radar Menu . We'll be using some of the functions as follows.
			Zoom in and out: On the Radar Menu, to the left. Press -, to zoom in . Press +, to zoom out . You can also hold the Ctrl button on the keyboard and zoom in and out by using the mouse scroll wheel. To view the surrounding area (without adjusting the zoom), press and hold the Ctrl keyboard button and the right mouse button while moving the mouse.

			<p>VECTOR and TRAILS: On the Radar Menu, to the right. Adjust aircraft VECTOR length to your preferences by selecting a number... Each number represents the extrapolation in minutes [1 = 1 minute, 2 = 2 minutes, etc.]. You cannot turn off VECTOR. The flight trails can be turned off by clicking the button DOTS.</p>
			<p>Flight Information (FIM) Click on the FIM button in the top right corner. The FIM shows the flight information window with more flight details (similar to a flight plan window). Most of the information are also found in the flight label. Hover the mouse over different aircraft call signs to see information.</p>
			<p>Label rotation: To manually rotate an aircraft label, left-click and hold down the label for aircraft ZIT13. Drag the mouse to position the label to your preference, and then release the mouse.</p>
	Present an aircraft label and explain how to interact with it. See also printout of label.		<p>Hover the mouse over flight XAV06 inside the sector. The 1st row presents aircraft callsign. The 2nd row presents Actual Flight Level, Cleared Flight Level (CFL), and Heading. The 3rd row presents Exit Flight Level and Exit Waypoint. On XAV06 label, you see that it is currently flying at FL290 as also indicated on CFL. But its exit flight level is FL330, meaning that you must climb this flight to FL330. HILAB is the exit waypoint.</p>
			<p>Assume an aircraft: In order to interact with an aircraft, an aircraft label must be assumed first. Let's assume flight YIR89, left-click the callsign and select "ASSUME". The label colour changes from green-white to all-green. Please assume flight ZUC57, LAC82, and WAD46.</p>

		<p>Transfer an aircraft: To transfer ZIT13, left-click the callsign and select “TRANSFER”. This transfers the flight to the next sector. The callsign changes color from green to white.</p>
		<p>View FPL route: To see an aircraft’s flight plan route, hover the cursor over flight YIR89 label, then left-click on the “Heading” field and hold. The route appears. When you release the mouse, the route disappears.</p>
		<p>To interact with an aircraft: To interact with an aircraft, hover the cursor over flight ZUC57 label and select a label item associated with your instruction (e.g. CFL, Heading). A menu will appear. You then can insert a new value.</p> <p>Note: If you mistakenly click on wrong aircraft or no longer want to interact with that aircraft, you can remove the menu by middle-click anywhere outside the menu box, or press the Esc key on the keyboard.</p>
		<p>Assign a new Cleared Flight Level (CFL): On COW70 label, click “CFL” label item. Clearance Menu opens and centers at the current flight level. Select a new flight level. This makes the cursor automatically move to the EXECUTE button. Press the EXECUTE button to make the aircraft change its flight level.</p> <p>If the Clearance Menu is blocking your view or it is on top of other aircraft, move it to empty space by left-click and hold the mouse. Please try it.</p>

			<p>Change Cleared Flight Level to Exit Flight Level: On XAV06 label, click “CFL” label item. Clearance Menu opens and centers at the required exit flight level, which is FL330. Press the exit flight level 330 in the CFL list. The cursor moves automatically to the EXECUTE button. Press the EXECUTE button to make the aircraft change its flight level.</p>
			<p>Assign a Heading: On ZUC57 label, click “Heading” label item. Clearance Menu opens and centers at the current heading. Select a new heading. The cursor moves automatically to the EXECUTE button. Press EXECUTE to make the aircraft change its heading.</p>
		<p>When the STCA alert goes off between aircraft YIR89 and REG05 OR between aircraft LAC82 and WAD46:</p>	<p>Short-Term Conflict Alert (STCA) and conflict alert window: When STCA issues an alert 2 min before the involved aircraft reach their Closest Point of Approach (CPA), the “Conflict Alert Message” window shows the callsigns of the conflict pair, remaining distance in NM and the CPA distance in NM. When CPA distance is displayed in red colour, the separation is smaller than the standard separation (< 5 NM).</p>
		<p>Present the VERA tool and the MTCD VERA window.</p>	<p>Use the “VERA” tool to probe conflicts On flight YIR89 label, left-click the callsign and select VERA. The cursor changes into a circle and a dashed line (showing radial and distance in NM) appears connecting the cursor and aircraft blip. Now you can measure the distance and direction from this aircraft. Move the cursor to the blip or label of flight REG05 and left-click to commit the selection. You will see white squares appear to the left of both flight labels and the conflict pair is added to the MTCD VERA window. You can add more pairs as needed. Hovering the cursor over the white squares will portray the extrapolated location (diabolo) on the map view where the CPA will be reached. The VERA window details the time to CPA (in minutes) and CPA distance (in NM) between the conflict pair. Left-clicking on a white square makes the diabolo persistent. Hovering the mouse cursor over the MTCD VERA window will also reveal the diabolo.</p>

			<p>Deactivate VERA tool To deactivate the VERA tool for YIR89 and REG05, right-click on the VERA table item in the MTC D VERA Winow belonging to the conflict pair.</p>
			<p>Solve conflict Solve the conflict between YIR89 and REG05 using heading.</p>
			<p>Use VERA tool Use the VERA tool on WAD46 and LAC82. <i>What is their CPA? (correct answer is 2.0)</i></p>
			<p>Solve conflict Solve the conflict between WAD46 and LAC82 using heading.</p>
		End of TW	<p>Do you have any questions? If not, end the training and move to step 2. Encourage them to try out the VERA tool on different flights.</p>
2		STEP 2: PARTICIPANT SELF TRAINING	
Time		INSTRUCTIONS to Instructor	INSTRUCTIONS to Participant
00:10:00	TST_1 TST_2 TST_3 TST_4	<p>Start self-training by typing in TST.</p> <p>The participant plays four scenarios á 2.5 minutes each without any specific instructions. Answer questions as they come up.</p>	<p>You will now play 4 scenarios on your own. Each scenario is 2.5 minutes long. You can ask questions about the interface and system and I will answer them.</p>

--	--	--	--

3		STEP 3: TRAINING TEST		
Time		INSTRUCTIONS to Instructor	INSTRUCTIONS to Participant	CHECKLIST
00:05:00	TT_1	Start training test playlist. The participant plays 2 scenarios á 2.5 minutes long. If the participant does not answer the question correctly, then provide instructions and reload the training test playlist.	You will now play 2 scenarios. The purpose is to test your knowledge of the system. We want to ascertain that you know how to use it well enough before we continue with the actual experiment. I will ask you questions and provide some simple instructions.	Mark with "X" when done or answered correctly.
	Scenario 1	Type in playlist TT and ask the followings.		
	1		Zoom out to range 145. Then adjust Zoom to your preference.	
	2		Rotate all flight labels to your preference.	
	3		Assume flight YH23W and XOQ65 .	
	4		Transfer flight NUL82 to the next sector.	
	5		Rotate the flight label of ZEJ99 .	

	6		Zoom in to 90 range with aircraft CQ81Q in the center.	
	7	Answer: FL340	What FL is CQ81Q flying?	
	8	Answer: FL340	At what FL should CQ81Q leave the sector?	
	9	Answer: YEHAV	At what waypoint should CQ81Q leave the sector?	
	10	Answer: Heading 222	What heading is CQ81Q on?	
	11	Answer: 0.7 NM (unless heading has been changed)	What is the CPA between CQ81Q and NX52Q ?	
	12		Change the heading of flight NX52Q .	
	13	Answer: UKAGE	Direct flight NX52Q to its exit waypoint. What is it called?	
			Continue playing the scenario until the end. When you are ready, press the Next Scenario button.	
	Scenario 2	When the next scenario starts, ask the following.		
	1		Open the aircraft label for flight RS44C and select the VERA tool.	
	2	Correct action: Press middle scroll wheel on the mouse.	De-select the VERA tool.	
	3		Use the VERA tool on RS44C and VJ27D .	
	4	Answer: 2 NM (see MTCD VERA window)	What is the CPA between RS44C and VJ27D ?	
	5	Correct action: right-click on callsign in MTCD VERA window .	Remove the aircraft pair from the VERA window.	



6		Change FL of VJ27D .	
7		Use the VERA tool on JH24Z and NEH77 .	
8	Answer: 2 NM (see MTCD VERA window)	What is the CPA between JH24Z and NEH77 ?	
9	Correct action: right-click on callsign in MTCD VERA window .	Remove the aircraft pair from the VERA window.	
		Continue playing the scenario until the end.	
	End of TT		

A6. Debrief

Participant ID: _____

Introduction

1. You have now played 36 traffic scenarios, what are your thoughts?
2. Think of the different scenarios you played - what do you think of their complexity? Did it vary? Were scenarios complex enough? Did scenarios vary in their complexity?

Resolution strategy

3. What strategies did you use for solving conflicts?
 - a. Why did you use this/these strategies?
 - b. Is this strategy/are these strategies what you normally use?
4. On a scale between 1-5, when solving conflicts, how obvious is the solution to you?
Circle the answer.

1	2	3	4	5
<i>"Not at all - it took time for me to come up with a solution"</i>				<i>"Very - I knew directly what to do - the solution was obvious to me"</i>

- a. If you troubleshoot a conflict to derive a solution – what aspects impact your decision?

Attitudes toward automation

5. On a scale between 1-5, to what extent do you use automation in your workplace?

Circle the answer.

1	2	3	4	5
<i>"I use as little automation as possible"</i>				<i>"I use as much automation as possible"</i>

6. On a scale between 1-5, to what extent did you use the VERA tool?

Circle the answer.

1	2	3	4	5
<i>"Not at all"</i>				<i>"All the time"</i>

7. In the simulation, you could use the CD&R support tool VERA. What support tools for CD&R do you have at your workstation?

a. Can you mention some good things about this/these tools?

b. Can you mention some bad things about this/these tools?

8. The level of automation is foreseen to increase in ATC over the next year.

a. For the task of CD&R, what would be the best use of automation do you think?

b. What should automation not do in CD&R according to you?

Other notes:

ANNEX B: MAIN EXPERIMENT MATERIALS

- B1. Experiment briefing
- B2. Introduction
- B3. Training instructions
- B4. Post solution questionnaire
- B5. Post session questionnaires (x3, for vector, diagram, and text conditions)
- B6. Debrief
- B7. Exit questionnaire

B1. Experiment Briefing

Simulator and scenarios

In today's simulation, you are asked to play several short scenarios. You will act as an air traffic controller supervising an artificial intelligent agent, like a digital colleague, that will manage most of the ATC tasks, including assuming aircraft, transferring aircraft, routing aircraft to their exit points, clearing aircraft to their exit flight levels, and conflict detection and resolution. To be clear: **all aircraft are controlled by automation (and appear blue) and cannot be interacted with**. As like last time, you will have the VERA tool to probe for conflicts, as you wish.

Like last time, you will work with the ATC simulator SectorX. Each scenario is about 3-4 minutes long. You interact with aircraft using the mouse and keyboard. All scenarios display a hypothetical sector in an octagon shape, about 100 x 100 NM in size. You are controlling RVSM airspace, covering FL290-410. The radar screen is updated in 2.5 second intervals. The simulator runs at 2x speed. This means that aircraft move 2 times faster than normal, which affects aspects such as the time lapsed for each scenario and rate of climb and rate of descent.

- Read to participant

Your task

Your task is to supervise the system and ensure separation between aircraft. However, the only time you will be able to interact with traffic is when the system notifies you of a pending conflict between aircraft. Again, you have a printout of what aircraft labels look like and what information they contain, depending on the state of the aircraft. [provide printout of aircraft label]

From time to time there may be conflicts between aircraft. You may use the VERA tool to probe for conflicts. A conflict occurs when aircraft are predicted to close within 5 NM and 1000 ft of one another. When the system predicts a loss of separation within the next 120 seconds, a short-term conflict alert (STCA) is provided. Because the simulation is played at 2 times faster than normal, this means 60 seconds prior to loss of separation.

The system will warn you of foreseen conflicts and suggest solutions for solving them. When an advisory appears, the simulation is paused to give you time to understand and analyse the suggested solution before deciding to accept or reject it. If you reject it, you must implement another solution.

Note that flights do not have a departure or destination aerodrome. The rules for semi-circular cruising levels are not applied. You are not able to coordinate with adjacent sectors. Communication with aircraft is carried out by means of data link CPDLC so there is no verbal communication with pilots. Unlike last time, you will **not be able to interact with aircraft** to change heading and altitude, except for when the system notifies you of a conflict.

Read to participant

Automation advisory

The system will help you identify and proactively avoid conflicts by suggesting solutions. Before the experiment starts you will play several training scenarios where you get to learn about the system, observe how it operates and how to interact with it. The system is an artificial intelligent agent. The resolution advisory should always precede a STCA. However, the STCA is prioritized over the automation advisory. As such there may be instances where no advisory is provided before STCA is triggered. This occurs, for example, in instances where the time to a loss of conflict is too short to provide an advisory.

When a conflict is detected by the system, the following occurs:

- The simulation pauses.
- The resolution advisory is shown in magenta for the aircraft concerned.
- You interact with the label to inspect the advisory and accept or reject it.

When the automated advisory appears, you should **carefully inspect the advisory** before choosing to accept or reject it. The automation does not necessarily advise the most optimal resolution. You may want to solve the conflict otherwise.

After you have [1] made a decision to accept or reject the advisory, [2] implemented an alternative advisory if you reject the system's advice, [3] answered two brief paper questions, and [4] responded to an onscreen question prompt, the scenario will continue playing. At the end of each scenario, you will see another onscreen question prompt. This sounds confusing, but don't worry, it will all be explained in the training session later.

Read to participant

Training

You will first be trained on how to use SectorX and interact with the artificial intelligent agent. Training will be provided before each session. Before each session, training will be provided in three parts. The first two parts are intended to re-familiarize your knowledge and skills in using SectorX. The last part is a test of your ability to understand information and use SectorX, understand and interact with the artificial intelligent agent.

B2. Introduction

5. Welcome and Presentation of researchers

6. Short presentation of MAHALO

Artificial Intelligence (AI) is becoming more common every day, across many applications. MAHALO (Modern ATM via Human/Automation Learning Optimisation) is a SESAR research project exploring possible application of AI in ATC. Specifically, MAHALO is exploring how AI *explainability* (i.e., how well can the controller understand AI inner processes) and *conformance* (i.e., how much does AI have to match the controller's own style) impact AI use in ATC.

As part of this effort, MAHALO is conducting a series of human-in-the-loop simulations, in which controllers are being asked to control a series of short air traffic scenarios.

7. Presentation of simulation

We have asked you here to take part in an ATC simulation. This simulation is realistic in some respects, however we have simplified several aspects of the ATC task. Today is the first phase of a two-step simulation. The second phase will take place in late April under week 17.

Your participation today is expected to last about three hours total, including three short breaks. After reading through the simulation briefing, signing the consent form, and completing a short questionnaire, you will spend around 25 minutes doing a training session. The goal of this training is for you to get to know the simulator and learn how to interact with it. During this training, I will sit next to you providing support and answering any questions you might have.

When you feel comfortable using the simulator, we will run three measurement sessions, each 30 minutes long. In these three sessions, you will see a number of short (2 ½ minutes) en-route air traffic scenarios. Once the simulation is completed, we will have a short debriefing, to ask some general questions about the simulation and the scenarios.

WE want to stress that we are NOT judging individual controllers, nor will we collect any identifying information. Your participation is both voluntary and anonymous.

Do you have any questions at this point?

8. Agenda

- Consent form
- Demographics Questionnaire
- Simulation briefing
- Training session
- Measurement session

Debriefing


B3. Training instructions

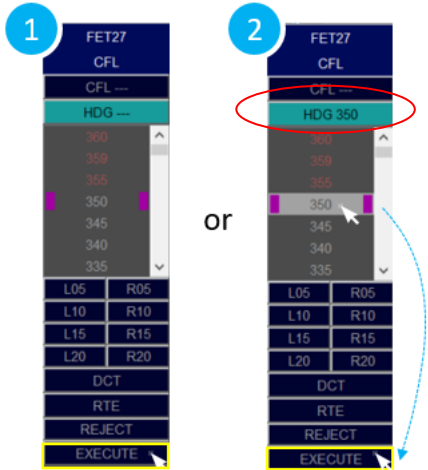
B3.1. T₀ (VECTOR) condition

MAHALO TRAINING Session SectorX			
1 STEP 1: TRAINING WALKTHROUGH			
Time	Playlist	INSTRUCTIONS to Instructor	INSTRUCTIONS to Participant on how to interact with the system
00:05:00	TWO	Show participant label image printed on paper (can be placed next to workspace). Show what different colours mean.	
		Start training by typing in playlist TWO . Provide instructions as the scenario plays out.	
		<i>REHEARSAL. If not read before:</i>	Aircraft are displayed in blue, indicating they are under control by “automation.” Flights cannot be controlled manually and the flight label items are inactive.
		<i>REHEARSAL. If not read before:</i>	The time at the top of the screen shows the time in the scenario.
		<i>REHEARSAL. If not read before:</i>	Please look at the rows on the top. Our focus is the 3rd row starting with MARKER button. This row is called Radar Menu . We'll be using some of the functions as follows.
		<i>REHEARSAL. If not read before:</i>	Zoom in and out: On the Radar Menu , to the left. Press -, to zoom in . Press +, to zoom out .

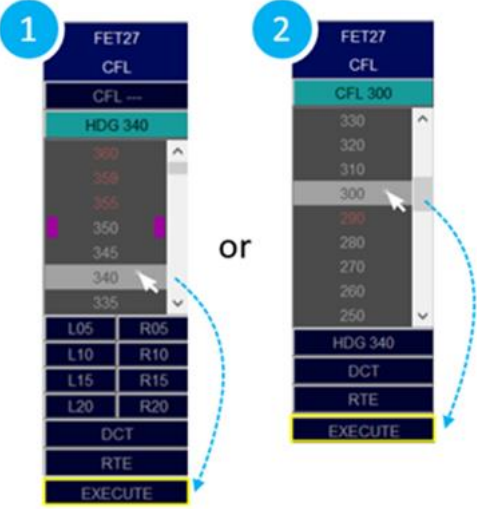
			To view the surrounding area (without adjusting the zoom), press and hold the Ctrl keyboard button and the right mouse button while moving the mouse.
		<i>REHEARSAL. If not read before:</i>	<p>VECTORS and TRAILS: On the Radar Menu, to the right.</p> <ul style="list-style-type: none"> i. To adjust aircraft VECTOR length, select a number 1, 2, 4 or 8. Each number represents the extrapolation in minutes [1 = 1 minute, 2 = 2 minutes etc.]. Note that you cannot turn off VECTOR. ii. To turn on/off the flight trails, click the DOTS button. Note that dots along the trail cannot be adjusted.
		<i>REHEARSAL. If not read before:</i>	<p>Flight Information (FIM): To view flight information, click on the FIM button in the top right corner. The FIM shows the flight information window with more detailed information about a flight. Most of which is also found in the flight label.</p>
		<i>REHEARSAL. If not read before:</i>	<p>Label rotation: To rotate an aircraft label, left-click and hold down the label.</p>
		<i>REHEARSAL. If not read before:</i>	<p>Please hover the mouse over flight XAV06. 1st row presents aircraft callsign. 2nd row presents Actual Flight Level, Cleared Flight Level (CFL), and Heading. 3rd row presents Exit Flight Level (XFL) and Exit Waypoint (XWP). XAV06 is flying at FL290. Automation detects that its XFL is FL330 and, therefore, clears XAV06 to FL330 (indicated on CFL item). XAV06 is directing to HILAB which is its XWP.</p>
		<i>REHEARSAL. If not read before:</i>	<p>Assume an aircraft: The system automatically assumes aircraft. The label colour changes from blue-white to all-blue. Please observe this action on flight ZUC57 or LAC82.</p>

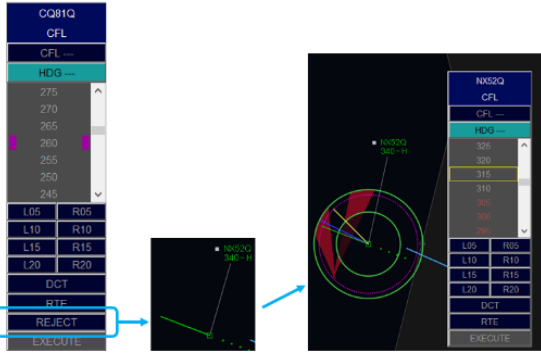
		<p><i>REHEARSAL. If not read before:</i></p>	<p>Transfer an aircraft: The system automatically transfers aircraft. The label colour then changes from all-blue to white-blue.</p>
		<p><i>REHEARSAL. If not read before:</i> Present the Message windows.</p>	<p>There are two message windows on the screen, i.e. CONFLICT ALERT MESSAGE, and MTCD VERA. (Arrange the windows to your preference. Ensure that they are neither on top of each other nor obscure the information.)</p> <ol style="list-style-type: none"> 1. "CONFLICT ALERT MESSAGE" window. When STCA issues an alert 2 min before the involved aircraft reach their Closest Point of Approach (CPA), the "Conflict Alert Message" window shows: <ol style="list-style-type: none"> i. the callsigns of the conflict pair, ii. remaining distance in Nautical Miles (NM), iii. CPA distance in NM. When CPA distance is displayed in red, separation is smaller than the required minima (< 5 NM). 2. "MTCD VERA" window. Use the "VERA" tool to probe conflicts On flight QUT47 label, left-click the callsign and select VERA. The cursor changes into a circle and a dashed line (showing radial and distance in NM) appears connecting the cursor and aircraft blip. Now you can measure the distance and direction from this aircraft. Move the cursor to the blip or label of flight REG05 and left-click to commit the selection. White squares will appear to the left of both flight labels and the aircraft pair is added to the MTCD VERA window. You can add more pairs as needed. Hovering the cursor over the white squares will portray the extrapolated location (diabolo) on the map view where the CPA will be reached. Left-clicking on a white square makes the diabolo persistent. Hovering the mouse cursor over the VERA window will also reveal the diabolo. The MTCD VERA window details: <ol style="list-style-type: none"> i. the callsigns of the conflict pair,

			<ul style="list-style-type: none"> ii. time to CPA in minutes, iii. CPA distance in NM. <p>Deactivate VERA To deactivate the VERA tool for QUT47 and REG05, right-click on the VERA table item belonging to the flight pair.</p>	
		<i>When a RESOLUTION Advisory is provided.</i>	The automation automatically detects conflicts and proposes advisories for solving them. When this happens, the simulation stops, the aircraft symbol for which a solution is proposed flashes in magenta colour, and a magenta vector line indicates the proposed heading.	
		BRIEFING	During this session, resolution advisories are presented as a magenta vector line .	
		<p>Interaction with proposals</p> <p><i>Demonstrate this on conflict QUT47 & REG05 (time: 03:20)</i></p>	<p>ACCEPT</p> 	<p>Accepting proposals</p> <p>For this conflict pair, the system suggest to solve the conflict by interacting with flight QUT47.</p> <ol style="list-style-type: none"> 1. Hover the mouse over the label of QUT47, the label is extended and magnified. 2. Left-click on the magenta HDG label item, a Clearance (CLR) menu opens, and advisory value is shown in magenta bars. In this scenario, it is H295. 3. Left-click on magenta HDG. The cursor automatically moves to EXECUTE button and the selected

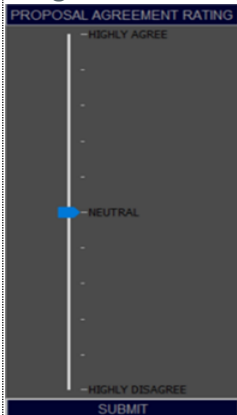
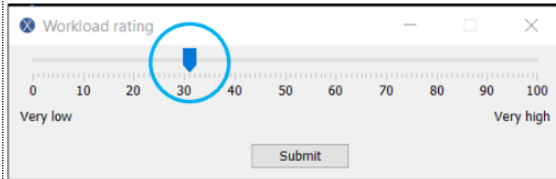
				<p>“HDG 295” is filled out in the HDG clearance box on the top. This is the heading that you have entered in the system – it is the same as the one proposed by the system as indicated in the magenta bars.</p> <p>4. Press EXECUTE.</p> <p>We will also train how to modify and reject the proposal, so DO NOT PRESS THE EXECUTE BUTTON now.</p>
--	--	--	--	--

		<p>AUGMENT</p>		<p>Modifying proposals</p> <p>If you want to modify the advisory on the selected aircraft QUT47, do the followings:</p> <ol style="list-style-type: none"> 1. Scroll up/down to the HDG you prefer. The yellow vector line visualizes the heading hovered over. 2. Select a new value. The cursor automatically moves to EXECUTE button and the selected “HDG --” is filled out in the HDG menu box on the top. Now, this is the heading that you have entered in the system. Press EXECUTE.
--	--	-----------------------	--	---

				<p>3. BUT If you also want to implement altitude solution together with HDG, select the HDG first (the selected “HDG --” is filled out in the HDG clearance box on the top). Then click on the CFL menu, select a level. Now you will have values filled in both CFL and HDG menu box. Press EXECUTE. Please try it.</p>
--	--	--	--	---

		<p>Demonstrate this on conflict LAC82 & WAD46 (time: 08:30)</p>	<p>REJECT</p> 	<p>Rejecting proposals</p> <p>If you instead <i>want to solve the conflict by interacting with the other aircraft (in this scenario, it's WAD46)</i>, do the followings:</p> <ol style="list-style-type: none"> 1. Press REJECT button, the proposal and CLR menu disappears and WAD46 will turn green, indicating it is now under your control. This means any type of clearance (CFL, HDG, combi) can be given to WAD46. 2. Click on HDG label item to open CLR menu.
--	--	---	--	--

				<p>3. For HDG solution, select a value. Note that when you scroll up/down on headings, the yellow vector line visualizes the heading hovered over. Now click on heading 075. The cursor automatically moves to EXECUTE button and the selected “HDG 075” is filled out in the HDG menu box on the top. Now, this is the heading that you have entered in the system. Press EXECUTE.</p> <p>4. For level solution, click CFL menu and select level 350. The cursor automatically moves to EXECUTE button and the selected “CFL 350” is filled out in the CFL menu box on the top. Now, this is the level that you have entered in the system. Press EXECUTE.</p> <p>5. For combined strategy, select values for both HDG and CFL item. The selected values are automatically filled out in its relative box. Press EXECUTE. Please try it. Note that WAD46 remains green and thus under manual control for the rest of scenario.</p>
			Questionnaires	Once you have executed the solution, you’ll be asked to answer the following questionnaires.
		Questionnaire must precede agreement rating	1. Post-solution questionnaire (online form)	Two items: conformance and understanding (scale 1- 6)

		<p><i>Prepare ATCo that sim will resume immediately after SUBMIT button is pressed</i></p>	<p>2. Agreement rating, onscreen prompt</p> 	<p>Agreement rating</p> <p>After accepting or rejecting proposals, a dialog pops up. This requires you to report your level of agreement (0-100 scale).</p> <p>Highly agree 100 Neutral 50 ← default selected value when dialog pops up Highly disagree 0</p> <p>After pressing the SUBMIT button, the simulation will resume.</p>
				<p>Simulation continues to end</p>
			<p>3. Workload rating, onscreen prompt</p> 	<p>When the scenario have ended, the workload prompt is shown.</p> <p>Indicate your workload (horizontal slider, scale 0-100). Think of your workload in terms of how difficult you found it to be to monitor the automation and specifically understanding the proposed solution.</p> <p>The initial position of the slider equals the reported workload rating from the previous scenario. In this way, it is easier for you to judge if the current scenario was more difficult or easier than the previous scenario.</p>

		End of TW	Do you have any questions? If not, end training and move to step 2.
--	--	------------------	---

2			
STEP 2: PARTICIPANT SELF TRAINING			
Time	Playlist	INSTRUCTIONS to Instructor	INSTRUCTIONS to Participant
00:05:00	TST0	Start self-training playlist TST0 . The participant plays two scenarios á 2.5 minutes each without any specific instructions. Answer questions as they come up.	You will now play 2 scenarios on your own. Each scenario is 2 minutes long. You can ask questions about the interface and system and I will answer them.

3				
STEP 3: TRAINING TEST				
Time	Playlist	INSTRUCTIONS to Instructor	INSTRUCTIONS to Participant	CHECKLIST
00:05:00	TTO	Start playlist TTO . The participant plays 2 scenarios á 2 minutes long. If the participant does not answer the question correctly, then provide instructions and reload the training test playlist.	You will now play 2 scenarios. The purpose is to test your knowledge of the system. We want to ascertain that you know how to use it well enough before we continue with the actual experiment. I will ask you questions and provide some simple instructions.	Mark with “X” when done or answered correctly.
	Scenario 1	Ask the followings.		
	1		Arrange the Message windows to your preference.	
	2		Rotate flight labels to your satisfaction.	

	3	Answer: 6.9NM	Use the VERA tool between FK72P and ZEJ99 , what is the CPA?	
	4		Remove the aircraft pair from the MTCD VERA window.	
	5	Answer: Heading 291	What is the HDG of NX52Q ?	
	6	Answer: FL340	What is FL of GK68T ?	
	7	Answer: Click open a CLR menu on HDG item.	When advisory pops up, show me how you inspect it.	
	8		Accept the proposal for the conflict pair.	
	9		Rate your agreement with the proposal.	
	10		Rate your workload.	
	Scenario 2	Ask the followings.		
	1		Select the VECTOR length to 2 minutes for all aircraft (then adjust as preferred).	
	2	Answer: HDG 260	According to the system, what is the proposal for the conflict pair RS44C and VJ27D ?	
	3		Modify the proposal.	
	4		Rate your agreement with the proposal.	
	5		For the conflict pair JH24Z and NEH77 , reject the proposal and implement your own solution.	
	6		Rate your agreement with the proposal.	

	7		Rate your workload.	
		END of TT		

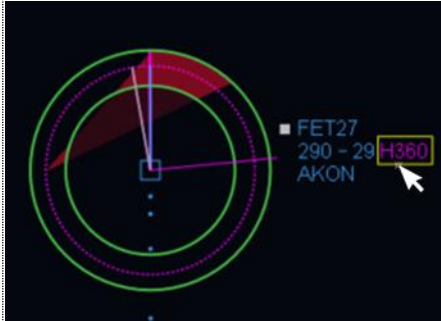
B3.2. T₁ (DIAGRAM) condition

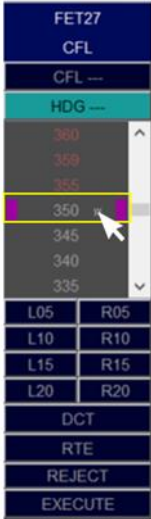
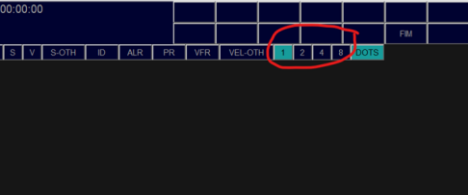

MAHALO TRAINING Session SectorX			
1 STEP 1: TRAINING WALKTHROUGH			
Time	Playlist	INSTRUCTIONS to Instructor	INSTRUCTIONS to Participant on how to interact with the system
00:05:00	TW1	Show participant label image printed on paper (can be placed next to workspace). Show what different colours mean.	
		Start training by typing in playlist TW1 . Provide instructions as the scenario plays out.	
		<i>REHEARSAL. If not read before:</i>	Aircraft are displayed in blue, indicating they are under control by “automation.” Flights cannot be controlled manually and the flight label items are inactive.
		<i>REHEARSAL. If not read before:</i>	The time at the top of the screen shows the time in the scenario.
		<i>REHEARSAL. If not read before:</i>	Please look at the rows on the top. Our focus is the 3rd row starting with MARKER button. This row is called Radar Menu . We'll be using some of the functions as follows.


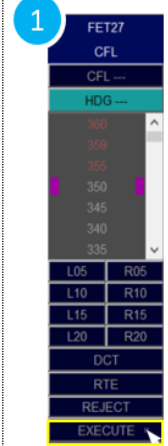
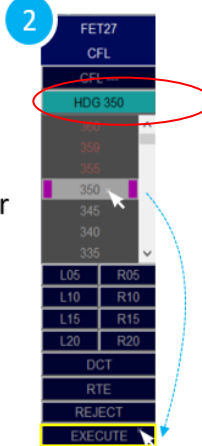
		<i>REHEARSAL. If not read before:</i>	<p>Zoom in and out: On the Radar Menu, to the left. Press -, to zoom in. Press +, to zoom out. To view the surrounding area (without adjusting the zoom), press and hold the Ctrl keyboard button and the right mouse button while moving the mouse.</p>
		<i>REHEARSAL. If not read before:</i>	<p>VECTORS and TRAILS: On the Radar Menu, to the right.</p> <ol style="list-style-type: none"> i. To adjust aircraft VECTOR length, select a number 1, 2, 4 or 8. Each number represents the extrapolation in minutes [1 = 1 minute, 2 = 2 minutes etc.]. Note that you cannot turn off VECTOR. ii. To turn on/off the flight trails, click the DOTS button. Note that dots along the trail cannot be adjusted.
		<i>REHEARSAL. If not read before:</i>	<p>Flight Information (FIM): To view flight information, click on the FIM button in the top right corner. The FIM shows the flight information window with more detailed information about a flight. Most of which is also found in the flight label.</p>
		<i>REHEARSAL. If not read before:</i>	<p>Label rotation: To rotate an aircraft label, left-click and hold down the label.</p>
		<i>REHEARSAL. If not read before:</i>	<p>Please hover the mouse over flight XAV06. 1st row presents aircraft callsign. 2nd row presents Actual Flight Level, Cleared Flight Level (CFL), and Heading. 3rd row presents Exit Flight Level (XFL) and Exit Waypoint (XWP). XAV06 is flying at FL290. Automation detects that its XFL is FL330 and, therefore, clears XAV06 to FL330 (indicated on CFL item). XAV06 is directing to HILAB which is its XWP.</p>

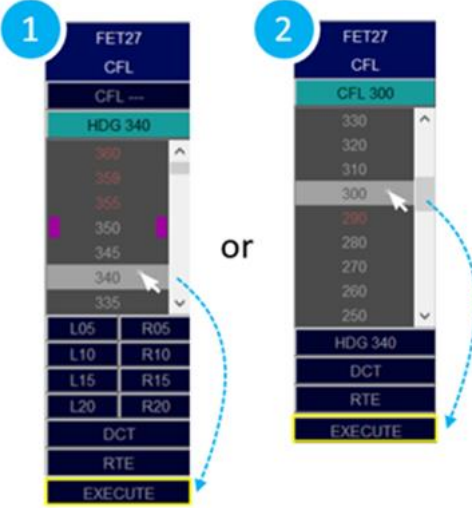
		<i>REHEARSAL. If not read before:</i>	<p>Assume an aircraft: The system automatically assumes aircraft. The label colour changes from blue-white to all-blue. Please observe this action on flight ZUC57 or LAC82.</p>
		<i>REHEARSAL. If not read before:</i>	<p>Transfer an aircraft: The system automatically transfers aircraft. The label colour then changes from all-blue to white-blue.</p>
		<p><i>REHEARSAL. If not read before:</i> Present the Message windows.</p>	<p>There are two message windows on the screen, i.e. CONFLICT ALERT MESSAGE, and MTCD VERA. (Arrange the windows to your preference. Ensure that they are neither on top of each other nor obscure the information.)</p> <ol style="list-style-type: none"> 1. "CONFLICT ALERT MESSAGE" window. When STCA issues an alert 2 min before the involved aircraft reach their Closest Point of Approach (CPA), the "Conflict Alert Message" window shows: <ol style="list-style-type: none"> i. the callsigns of the conflict pair, ii. remaining distance in Nautical Miles (NM), iii. CPA distance in NM. When CPA distance is displayed in red, separation is smaller than the required minima (< 5 NM). 2. "MTCD VERA" window. Use the "VERA" tool to probe conflicts On flight QUT47 label, left-click the callsign and select VERA. The cursor changes into a circle and a dashed line (showing radial and distance in NM) appears connecting the cursor and aircraft blip. Now you can measure the distance and direction from this aircraft. Move the cursor to the blip or label of flight REG05 and left-click to commit the selection. White squares will appear to the left of both flight labels and the aircraft pair is added to the MTCD VERA window. You can add more pairs as needed. Hovering the cursor over the white squares will portray the extrapolated location (diabolo) on the map view where the CPA will be reached. Left-clicking on a white square makes the diabolo

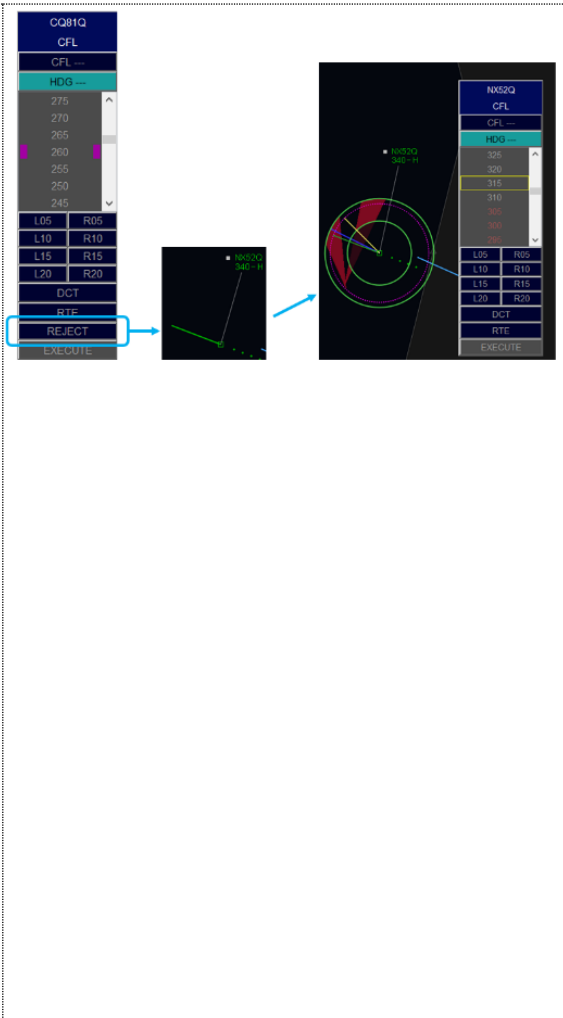
			<p>persistent. Hovering the mouse cursor over the VERA window will also reveal the diablo.</p> <p>The MTCD VERA window details:</p> <ol style="list-style-type: none"> i. the callsigns of the conflict pair, ii. time to CPA in minutes, iii. CPA distance in NM. <p>Deactivate VERA</p> <p>To deactivate the VERA tool for QUT47 and REG05, right-click on the VERA table item belonging to the flight pair.</p>
		<p><i>When a RESOLUTION Advisory is provided.</i></p>	<p>The automation automatically detects conflicts and proposes advisories for solving them. When this happens, the simulation stops, the aircraft symbol for which a solution is proposed flashes in magenta colour, and a magenta vector line indicates the proposed heading.</p>
		<p>BRIEFING:</p>	<p>When the system suggests a solution, the resolution advisory is indicated by a magenta vector line. In addition, when a resolution advisory is suggested, the system presents the solution in a circle Diagram for the selected aircraft. The Diagram becomes visible when the resolution is inspected in the flight label.</p> <p>[Skip if T2 has been completed] The Diagram only presents information related to the horizontal relationship between aircraft. It does not present information about the vertical relationship between aircraft. The Diagram consists of a circle diagram around a selected aircraft. The inner diameter represents the lower boundary of the selected aircraft’s speed envelope. The outer diameter represents the upper boundary of the speed envelope. Other aircraft are represented by triangles inside the Diagram, which are conflict zones showing all heading and speed combinations that will result in a loss of separation with another aircraft. Hovering the mouse cursor over a conflict triangle highlights its corresponding aircraft on the radar screen in red. Each aircraft will have its own unique triangle inside the diagram. Any “hole” in the circular Diagram</p>

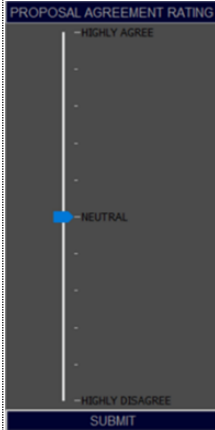
		<p>represents a “go” area, or a potential solution, to resolve a conflict. Resolving a conflict involves aiming the tip of the speed vector outside of any triangle. Note: a speed vector can intersect a triangle, but as long as the tip of the vector is located outside the triangle, the conflict will be resolved.</p>
		<p>Space Solution Diagram (SSD)</p>  <p>Circle Diagram</p> <p>Clicking the left mouse button on the heading label item opens the Clearance (CLR) menu in conjunction with the SSD + proposal.</p> <p>Select a longer vector range to increase the size of the SSD.</p> <p>Conflicting headings are shaded in red and match with SSD conflict zones. The red triangle indicates the relative position of other aircraft. If the system avoids these, there will be no conflict.</p> <p>By hovering over red triangle, the aircraft that it is in conflict with is shown in red colour.</p> <p>TIP: Use the VECTOR (right in radar menu) to “blow up” the SSD, and see in more detail where the vector is aimed inside the SSD.</p>

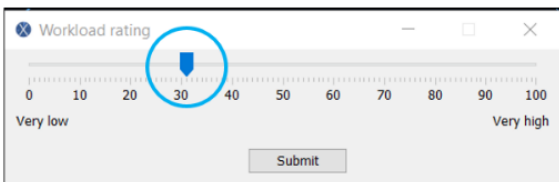
				 <p>Proposal is marked with magenta bars.</p>
				<p>Hovering with the mouse cursor over other headings shows the corresponding YELLOW preview vector in the SSD.</p>

		<p>Interaction with proposals</p> <p><i>Demonstrate this on conflict QUT47 & REG05 (time: 03:20)</i></p>	<p>ACCEPT</p>  <p>1</p>  <p>or</p> <p>2</p> 	<p>Accepting proposals</p> <p>For this conflict pair, the system suggest to solve the conflict by interacting with flight QUT47.</p> <ol style="list-style-type: none"> 1. Hover the mouse over the label of QUT47, the label is extended and magnified. 2. Left-click on the magenta HDG label item, a Clearance (CLR) menu opens, and advisory value is shown in magenta bars. In this scenario, it is H295. 3. Left-click on magenta HDG. The cursor automatically moves to EXECUTE button and the selected "HDG 295" is filled out in the HDG clearance box on the top. This is the heading that you have entered in the system – it is the same as the one proposed by the system as indicated in the magenta bars. 4. Press EXECUTE. <p>We will also train how to modify and reject the proposal, so DO NOT PRESS THE EXECUTE BUTTON now.</p>
--	--	---	--	--

			<p>AUGMENT</p> 	<p>Modifying proposals</p> <p>If you want to modify the advisory on the selected aircraft QUT47, do the followings:</p> <ol style="list-style-type: none"> 1. Scroll up/down to the HDG you prefer. The yellow vector line visualizes the heading hovered over. 2. Select a new value. The cursor automatically moves to EXECUTE button and the selected “HDG --” is filled out in the HDG menu box on the top. Now, this is the heading that you have entered in the system. Press EXECUTE. 3. BUT If you also want to implement altitude solution together with HDG, select the HDG first (the selected “HDG --” is filled out in the HDG clearance box on the top). Then click on the CFL menu, select a level. Now you will have values filled in both CFL and HDG menu box. Press EXECUTE. Please try it.
		<p><i>Demonstrate this on conflict LAC82 & WAD46 (time: 08:30)</i></p>	<p>REJECT</p>	<p>Rejecting proposals</p> <p>If you instead <i>want to solve the conflict by interacting with the other aircraft (in this scenario, it's WAD46)</i>, do the followings:</p> <ol style="list-style-type: none"> 1. Press REJECT button, the proposal and CLR menu disappears and WAD46 will turn

				<p>green, indicating it is now under your control. This means any type of clearance (CFL, HDG, combi) can be given to WAD46.</p> <ol style="list-style-type: none"> Click on HDG label item to open CLR menu. For HDG solution, select a value. Note that when you scroll up/down on headings, the yellow vector line visualizes the heading hovered over. Now click on heading 075. The cursor automatically moves to EXECUTE button and the selected “HDG 075” is filled out in the HDG menu box on the top. Now, this is the heading that you have entered in the system. Press EXECUTE. For level solution, click CFL menu and select level 350. The cursor automatically moves to EXECUTE button and the selected “CFL 350” is filled out in the CFL menu box on the top. Now, this is the level that you have entered in the system. Press EXECUTE. For combined strategy, select values for both HDG and CFL item. The selected values are automatically filled out in its relative box. Press EXECUTE. Please try it. Note that WAD46 remains green and thus under manual control for the rest of scenario.
--	--	--	---	---

			Questionnaires	Once you have executed the solution, you'll be asked to answer the following questionnaires.
		Questionnaire must precede agreement rating	1. Post-solution questionnaire	Two items: conformance and understanding (scale 1-6)
		<i>Prepare ATCo that sim will resume immediately after SUBMIT button is pressed</i>	2. Agreement rating, onscreen prompt 	Agreement rating After accepting or rejecting proposals, a dialog pops up. This requires you to report your level of agreement (0-100 scale). Highly agree 100 Neutral 50 ← default selected value when dialog pops up Highly disagree 0 After pressing the SUBMIT button, the simulation will resume.
				Simulation continues to end
			3. Workload rating, onscreen prompt	When the scenario have ended, the workload prompt is shown. Indicate your workload (horizontal slider, scale 0-100). Think of your workload in terms of how difficult you found it to be to monitor the

				<p>automation and specifically understanding the proposed solution.</p> <p>The initial position of the slider equals the reported workload rating from the previous scenario. In this way, it is easier for you to judge if the current scenario was more difficult or easier than the previous scenario.</p>
--	--	--	--	---

		End of TW	Do you have any questions? If not, end training and move to step 2.
--	--	------------------	---

2				STEP 2: PARTICIPANT SELF TRAINING			
Time	Playlist	INSTRUCTIONS to Instructor	INSTRUCTIONS to Participant				
00:05:00	TST1	Start self-training playlist TST1 . The participant plays two scenarios á 2 minutes each without any specific instructions. Answer questions as they come up.	You will now play 2 scenarios on your own. Each scenario is 2.5 minutes long. You can ask questions about the interface and system and I will answer them.				

3				STEP 3: TRAINING TEST			
Time	Playlist	INSTRUCTIONS to Instructor	INSTRUCTIONS to Participant				CHECKLIST

00:05:00	TT1	Start playlist TT1 . The participant plays 2 scenarios á 2 minutes long. If the participant does not answer the question correctly, then provide instructions and reload the training test playlist.	You will now play 2 scenarios. The purpose is to test your knowledge of the system. We want to ascertain that you know how to use it well enough before we continue with the actual experiment. I will ask you questions and provide some simple instructions.	Mark with “X” when done or answered correctly.
	Scenario 1	Ask the followings.		
	1		Arrange the Message windows to your preference.	
	2		Rotate flight labels to your satisfaction.	
	3	Answer: 6.9NM	Use VERA tool between FK72P and ZEJ99 , what is the CPA?	
	4		Remove the aircraft pair from the MTCD VERA window.	
	5	Answer: HDG 291	What is the HDG of NX52Q ?	
	6	Answer: FL340	What is FL of GK68T ?	
	7	Answer: A conflict with NX52Q	When an advisory pops up, open and inspect the Diagram. What does the red triangle indicate?	
	8	Answer: HDG 260	What is the proposed advisory?	
	9		Accept the proposal for the conflict pair.	
	10		Rate your agreement with the proposal.	
	11		Rate your workload.	
	Scenario 2	Ask the following.		



1		Select the VECTOR length to 2 minutes for all aircraft (then adjust as preferred).	
2		When an advisory pops up, open and inspect the Diagram. Set VECTOR length to 4 to increase the Diagram view.	
3		Modify the proposal for the conflict pair RS44C and VJ27D .	
4		Rate your agreement with the proposal.	
5		Reject the proposal for the conflict pair JH24Z and NEH77 , and implement your own solution.	
6		Rate your agreement with the proposal.	
7		Rate your workload.	
		END of TT	

B3.3. T₂ (TEXT) condition



MAHALO TRAINING Session SectorX			
1 STEP 1: TRAINING WALKTHROUGH			
Time	Playlist	INSTRUCTIONS to Instructor	INSTRUCTIONS to Participant on how to interact with the system
00:05:00	TW2	Show participant label image printed on paper (can be placed next to workspace). Show what different colours mean.	
		Start training by typing in playlist TW2 . Provide instructions as the scenario plays out.	
		<i>REHEARSAL. If not read before:</i>	Aircraft are displayed in blue, indicating they are under control by “automation.” Flights cannot be controlled manually and the flight label items are inactive.
		<i>READ</i>	<p>The time at the top of the screen shows the time in the scenario.</p> <p>Move the “Events” window to an empty space, if necessary. Adjust the size by extending it, to see entire messages.</p> <p>During this session, the system will inform you of what it is doing. A Text window will provide text information about actions taken by the automation such as assuming and transferring flights, conflicts detected, and proposed resolutions.</p>


		<i>REHEARSAL. If not read before:</i>	Please look at the rows on the top. Our focus is the 3rd row starting with MARKER button. This row is called Radar Menu . We'll be using some of the functions as follows.
		<i>REHEARSAL. If not read before:</i>	Zoom in and out: On the Radar Menu , to the left. Press -, to zoom in . Press +, to zoom out . To view the surrounding area (without adjusting the zoom), press and hold the Ctrl keyboard button and the right mouse button while moving the mouse.
		<i>REHEARSAL. If not read before:</i>	VECTORS and TRAILS: On the Radar Menu , to the right. i. To adjust aircraft VECTOR length, select a number 1, 2, 4 or 8. Each number represents the extrapolation in minutes [1 = 1 minute, 2 = 2 minutes etc.]. Note that you cannot turn off VECTOR . ii. To turn on/off the flight trails, click the DOTS button. Note that dots along the trail cannot be adjusted.
		<i>REHEARSAL. If not read before:</i>	Flight Information (FIM): To view flight information, click on the FIM button in the top right corner. The FIM shows the flight information window with more detailed information about a flight. Most of which is also found in the flight label.
		<i>REHEARSAL. If not read before:</i>	Label rotation: To rotate an aircraft label, left-click and hold down the label.
		<i>REHEARSAL. If not read before:</i>	Please hover the mouse over flight XAV06 . 1st row presents aircraft callsign . 2nd row presents Actual Flight Level , Cleared Flight Level (CFL) , and Heading . 3rd row presents Exit Flight Level (XFL) and Exit Waypoint (XWP) . XAV06 is flying at FL290. Automation detects that its XFL is FL330 and, therefore, clears XAV06 to FL330 (indicated on CFL item). XAV06 is directing to HILAB which is its XWP .


		<p><i>READ:</i></p>	<p>Assume an aircraft: The system automatically assumes aircraft. The label colour changes from blue-white to all-blue. Please observe this action on flight ZUC57 or LAC82. In the EVENT window, the system states that “assuming control over ZUC57.”</p>
		<p><i>REHEARSAL. If not read before:</i></p>	<p>Transfer an aircraft: The system automatically transfers aircraft. The label colour then changes from all-blue to white-blue.</p>
		<p><i>SKIP CONFLICT ALERT AND MTCD VERA if read before</i></p>	<p>There are three message windows on the screen, i.e. EVENT, CONFLICT ALERT MESSAGE, and MTCD VERA. (Arrange the windows to your preference. Ensure that they are neither on top of each other nor obscure the information.)</p> <ol style="list-style-type: none"> 1. “Events” window presents occurrences and actions performed by Automation. The window can be expanded/reduced by hovering the mouse over a window corner or edge followed by left-click, then drag to adjust the size. Three types of information are presented: Time, Agent and Message. <ol style="list-style-type: none"> i. Time shows when an action is performed. ii. Agent specifies who performs the action, i.e. automation or human. iii. Message details what action is performed. <p>An action performed is colour-coded. Each colour represents the followings:</p> <p>Red conflict detected Magenta resolution proposal / advisory Light Pink (Old Rose) action or clearance to aircraft (e.g. assume, transfer)</p> 2. “CONFLICT ALERT MESSAGE” window. When STCA issues an alert 2 min before the involved aircraft reach their Closest Point of Approach (CPA), the "Conflict Alert Message" window shows: <ol style="list-style-type: none"> i. the callsigns of the conflict pair, ii. remaining distance in Nautical Miles (NM),

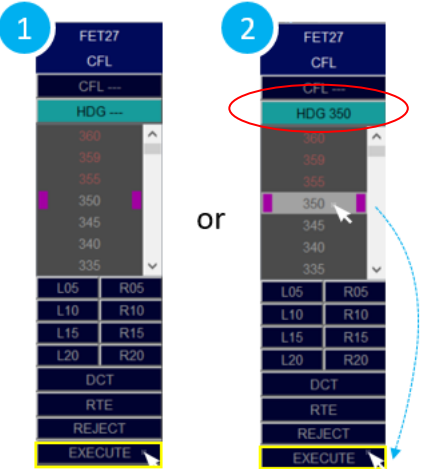
		<p>SKIP CONFLICT ALERT AND MTCD VERA if read before</p>	<ul style="list-style-type: none"> iii. CPA distance in NM. When CPA distance is displayed in red, separation is smaller than the required minima (< 5 NM). <p>3. "MTCD VERA" window.</p> <p>Use the "VERA" tool to probe conflicts</p> <p>On flight QUT47 label, left-click the callsign and select VERA. The cursor changes into a circle and a dashed line (showing radial and distance in NM) appears connecting the cursor and aircraft blip. Now you can measure the distance and direction from this aircraft. Move the cursor to the blip or label of flight REG05 and left-click to commit the selection.</p> <p>White squares will appear to the left of both flight labels and the aircraft pair is added to the MTCD VERA window. You can add more pairs as needed. Hovering the cursor over the white squares will portray the extrapolated location (diabolo) on the map view where the CPA will be reached. Left-clicking on a white square makes the diabolo persistent. Hovering the mouse cursor over the VERA window will also reveal the diabolo.</p> <p>The VERA window details:</p> <ul style="list-style-type: none"> i. the callsigns of the conflict pair, ii. time to CPA in minutes, iii. CPA distance in NM. <p>Deactivate VERA</p> <p>To deactivate the VERA tool for QUT47 and REG05, right-click on the VERA table item belonging to the flight pair.</p>
		<p>When a RESOLUTION Advisory is provided.</p>	<p>The automation automatically detects conflicts and proposes advisories for solving them. When this happens, the simulation stops, the aircraft symbol for which a solution is proposed flashes in magenta colour, and a magenta vector line indicates the proposed heading.</p>

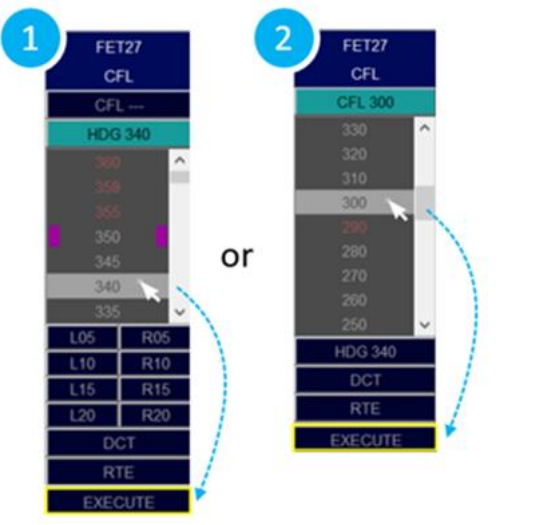
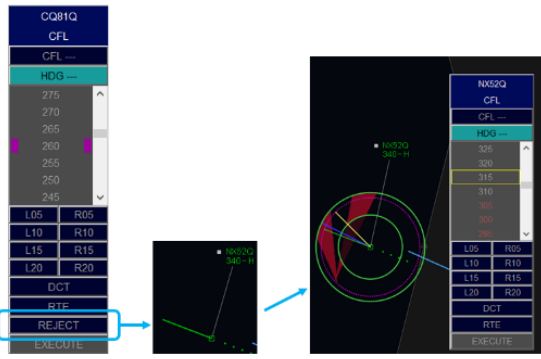
		<p>BRIEFING</p>	<p>When the system suggests a solution, the resolution advisory is indicated by a magenta vector line. In addition, when a resolution advisory is suggested, the system presents the solution in a circle Diagram for the selected aircraft. The Diagram becomes visible when the resolution is inspected in the flight label. In the Text window, the conflict detection and resolution system will provide a text message with information about the proposed solution.</p> <p>In the EVENT window, the system notifies you of a detected conflict by using a red coloured box. It says “conflict detected between QUT47 and REG05. In this conflict, the system proposes to turn QUT47 to the right heading 295. In the EVENT window, the system argues to “turn QUT47 right to aim at 10nm separation.”</p> <p>[Skip if T1 has been completed]</p> <p>The Diagram only presents information related to the horizontal relationship between aircraft. It does not present information about the vertical relationship between aircraft. The Diagram consists of a circle diagram around a selected aircraft. The inner diameter represents the lower boundary of the selected aircraft’s speed envelope. The outer diameter represents the upper boundary of the speed envelope. Other aircraft are represented by triangles inside the Diagram, which are conflict zones showing all heading and speed combinations that will result in a loss of separation with another aircraft. Hovering the mouse cursor over a conflict triangle highlights its corresponding aircraft on the radar screen in red. Each aircraft will have its own unique triangle inside the diagram. Any “hole” in the circular Diagram represents a “go” area, or a potential solution, to resolve a conflict. Resolving a conflict involves <u>aiming the tip of the speed trend vector outside of any triangle</u>. Note: a speed vector can intersect a triangle, <u>but as long as the tip of the vector is located outside the triangle, the conflict will be resolved.</u></p>		
			<table border="1" style="width: 100%;"> <tr> <td data-bbox="965 1254 1525 1342">Solution Space Diagram (SSD)</td> <td data-bbox="1525 1254 2112 1342">Circle Diagram</td> </tr> </table>	Solution Space Diagram (SSD)	Circle Diagram
Solution Space Diagram (SSD)	Circle Diagram				

				<p>Clicking the left mouse button on the heading label item opens the clearance menu in conjunction with the SSD + proposal.</p> <p>Select a longer vector range to increase the size of the SSD.</p> <p>Conflicting headings are shaded in red and match with SSD conflict zones. The red triangle indicates the relative position of other aircraft. If the system avoids these, there will be no conflict.</p> <p>By hovering over red triangle, the aircraft that it is in conflict with is shown in red colour.</p> <p>TIP: Use the VECTOR (right in radar menu) to “blow up” the SSD, and see in more detail where the vector is aimed inside the SSD.</p>  <p>Proposal is marked with magenta bars.</p>
--	--	--	---	---

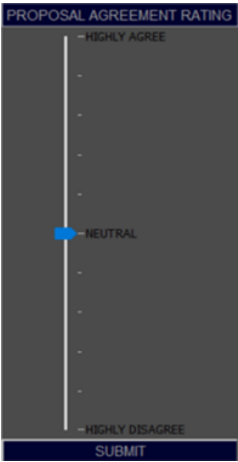
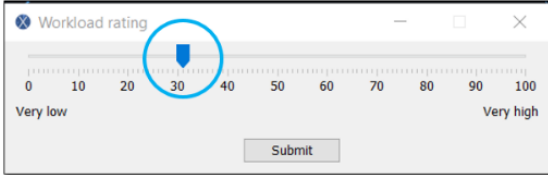
				<p>Hovering with the mouse cursor over other headings shows the corresponding YELLOW preview vector in the SSD.</p>
--	--	--	--	---

		<p>Interaction with proposals</p> <p><i>Demonstrate this on conflict QUT47 & REG05 (time: 03:20)</i></p>	<p>ACCEPT</p> 	<p>Accepting proposals</p> <p>For this conflict pair, the system suggest to solve the conflict by interacting with flight QUT47.</p> <ol style="list-style-type: none"> 1. Hover the mouse over the label of QUT47, the label is extended and magnified. 2. Left-click on the magenta HDG label item, a Clearance (CLR) menu opens, and advisory value is shown in magenta bars. In this scenario, it is H295. 3. Left-click on magenta HDG. The cursor automatically moves to EXECUTE button and the selected “HDG 295” is filled out in
--	--	---	--	--

				<p>the HDG clearance box on the top. This is the heading that you have entered in the system – it is the same as the one proposed by the system as indicated in the magenta bars.</p> <p>4. Press EXECUTE.</p> <p>We will also train how to modify and reject the proposal, so DO NOT PRESS THE EXECUTE BUTTON now.</p>
			<p>AUGMENT</p>	<p>Modifying proposals</p> <p>If you want to modify the advisory on the selected aircraft QUT47, do the followings:</p> <ol style="list-style-type: none"> 1. Scroll up/down to the HDG you prefer. The yellow vector line visualizes the heading hovered over. 2. Select a new value. The cursor automatically moves to EXECUTE button and the selected “HDG --” is filled out in the HDG menu box on the top. Now, this is the heading that you have entered in the system. Press EXECUTE.

				<p>3. BUT If you also want to implement altitude solution together with HDG, select the HDG first (the selected “HDG --” is filled out in the HDG clearance box on the top). Then click on the CFL menu, select a level. Now you will have values filled in both CFL and HDG menu box. Press EXECUTE. Please try it.</p>
		<p>Demonstrate this on conflict LAC82 & WAD46 (time: 08:30)</p>	<p>REJECT</p> 	<p>Rejecting proposals</p> <p>If you instead <i>want to solve the conflict by interacting with the other aircraft (in this scenario, it's WAD46)</i>, do the followings:</p> <ol style="list-style-type: none"> 1. Press REJECT button, the proposal and CLR menu disappears and WAD46 will turn green, indicating it is now under your control. This means any type of clearance (CFL, HDG, combi) can be given to WAD46. 2. Click on HDG label item to open CLR menu. 3. For HDG solution, select a value. Note that when you scroll up/down on headings, the

				<p>yellow vector line visualizes the heading hovered over. Now click on heading 075. The cursor automatically moves to EXECUTE button and the selected “HDG 075” is filled out in the HDG menu box on the top. Now, this is the heading that you have entered in the system. Press EXECUTE.</p> <p>4. For level solution, click CFL menu and select level 350. The cursor automatically moves to EXECUTE button and the selected “CFL 350” is filled out in the CFL menu box on the top. Now, this is the level that you have entered in the system. Press EXECUTE.</p> <p>5. For combined strategy, select values for both HDG and CFL item. The selected values are automatically filled out in its relative box. Press EXECUTE. Please try it. Note that WAD46 remains green and thus under manual control for the rest of scenario.</p>
			Questionnaires	Once you have executed the solution, you'll be asked to answer the following questionnaires.
		Questionnaire must precede agreement rating	1. Post-solution questionnaire	Two items: conformance and understanding (1-6)

		<p><i>Prepare ATCo that sim will resume immediately after SUBMIT button is pressed</i></p>	<p>2. Agreement rating, onscreen prompt</p> 	<p>Agreement rating</p> <p>After accepting or rejecting proposals, a dialog pops up. This requires you to report your level of agreement (0-100 scale).</p> <p>Highly agree 100 Neutral 50 ← default selected value when dialog pops up Highly disagree 0</p> <p>After pressing the SUBMIT button, the simulation will resume.</p>
				<p>Simulation continues to end</p>
			<p>3. Workload rating, onscreen prompt</p> 	<p>When the scenario have ended, the workload prompt is shown.</p> <p>Indicate your workload (horizontal slider, scale 0-100). Think of your workload in terms of how difficult you found it to be to monitor the automation and specifically understanding the proposed solution.</p> <p>The initial position of the slider equals the reported workload rating from the previous</p>

				scenario. In this way, it is easier for you to judge if the current scenario was more difficult or easier than the previous scenario.
--	--	--	--	---

		End of TW	Do you have any questions? If not, end training and move to step 2.
--	--	------------------	---

2				STEP 2: PARTICIPANT SELF TRAINING	
Time	Playlist	INSTRUCTIONS to Instructor	INSTRUCTIONS to Participant		
00:05:00	TST2	Start self-training playlist TST2 . The participant plays two scenarios á 2.5 minutes each without any specific instructions. Answer questions as they come up.	You will now play 2 scenarios on your own. Each scenario is 2 minutes long. You can ask questions about the interface and system and I will answer them.		

3				STEP 3: TRAINING TEST	
Time	Playlist	INSTRUCTIONS to Instructor	INSTRUCTIONS to Participant	CHECKLIST	
00:05:00	TT2	Start playlist TT2 . The participant plays 2 scenarios á 2 minutes long. If the participant does not answer the question correctly, then provide instructions and reload the training test playlist.	You will now play 2 scenarios. The purpose is to test your knowledge of the system. We want to ascertain that you know how to use it well enough before we continue with the actual experiment. I will ask you questions and provide some simple instructions.	Mark with "X" when done or answered correctly.	

	Scenario 1	Ask the followings.	
1			Arrange the Message windows to your preference.
2			Rotate flight labels to your satisfaction.
3	Answer: 6.9NM		Use the VERA tool between FK72P and ZEJ99 , what is the CPA?
4			Remove the aircraft pair from the MTCD VERA window.
5	Answer: Heading 291		What is the HDG of NX52Q ?
6	Answer: FL340		What is FL of GK68T ?
7			When an advisory pops up, inspect the message in the Events window. What does the red-coded message suggest? What about the magenta-coded message?
8			What argument for the advisory is made in the Events window?
9			Accept the proposal for the conflict pair.
10			Rate your agreement with the proposal.
11			Rate your workload.
	Scenario 2	Ask the followings.	
1			Select the VECTOR length to 2 minutes for all aircraft (then adjust as preferred)
2			When an advisory pops up, inspect the Diagram. Set VECTOR length to 4 to increase the Diagram view.
3			What argument for the advisory is made in the Events window?

4		Modify the proposal for the conflict pair RS44C and VJ27D .	
5		Rate your agreement with the proposal.	
6		For the conflict pair JH24Z and NEH77 , Reject the proposal and implement your own solution.	
7		Rate your agreement with the proposal.	
8		Rate your workload.	
	END of TT		

B4. Post-solution questionnaire

For the solution you just saw, please indicate your agreement with the two statements below, on a scale of 1 (Disagree highly) to 6 (Agree highly).

The system solved the conflict the same way I would have.

1	2	3	4	5	6
Disagree highly			Agree highly		

I can understand why the system suggested that solution.

1	2	3	4	5	6
Disagree highly			Agree highly		

B5. Post-session questionnaire

B5.1 T₀ (VECTOR) condition

In all sessions today, you will be using an artificial intelligence (AI) agent that detects conflicts, and proposes solutions.

In this session, solutions were presented as a **heading vector**.

Please indicate your agreement with each statement below, on a scale of 1 (Disagree highly) to 6 (Agree highly).

The solutions were accurate.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were safe.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were efficient.

1	2	3	4	5	6
Disagree highly				Agree highly	

I agreed with the system's solutions.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were different than I would have chosen.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were better than I would have chosen.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions lowered my workload.

1	2	3	4	5	6
Disagree highly				Agree highly	

I trusted the solutions.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were presented too early.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were presented too late.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions helped me resolve conflicts quicker.

1	2	3	4	5	6
Disagree highly				Agree highly	

The system was easy to use.

1	2	3	4	5	6
Disagree highly				Agree highly	

The presentation format made it easy to understand the solution.

1	2	3	4	5	6
Disagree highly				Agree highly	

B5.2 T_1 (DIAGRAM) condition

In all sessions today, you will be using an artificial intelligence (AI) agent that detects conflicts, and proposes solutions.

In this session, solutions were presented as a **heading vector, combined with a solution diagram**.

Please indicate your agreement with each statement below, on a scale of 1 (Disagree highly) to 6 (Agree highly).

The solutions were accurate.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were safe.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were efficient.

1	2	3	4	5	6
Disagree highly				Agree highly	

I agreed with the system's solutions.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were different than I would have chosen.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were better than I would have chosen.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions lowered my workload.

1	2	3	4	5	6
Disagree highly				Agree highly	

I trusted the solutions.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were presented too early.

1	2	3	4	5	6
Disagree highly					Agree highly
1	2	3	4	5	6
Disagree highly					Agree highly

The solutions helped me resolve conflicts quicker.

1	2	3	4	5	6
Disagree highly					Agree highly

The system was easy to use.

1	2	3	4	5	6
Disagree highly					Agree highly

The presentation format made it easy to understand the solution.

1	2	3	4	5	6
Disagree highly					Agree highly

B5.3 T₂ (TEXT) condition

In all sessions today, you will be using an artificial intelligence (AI) agent that detects conflicts, and proposes solutions.

In this session, solutions were presented as a **heading vector, combined with a solution diagram, and a text message explanation.**

Please indicate your agreement with each statement below, on a scale of 1 (Disagree highly) to 6 (Agree highly).

The solutions were accurate.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were safe.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were efficient.

1	2	3	4	5	6
Disagree highly				Agree highly	

I agreed with the system’s solutions.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were different than I would have chosen.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were better than I would have chosen.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions lowered my workload.

1	2	3	4	5	6
Disagree highly				Agree highly	

I trusted the solutions.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were presented too early.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions were presented too late.

1	2	3	4	5	6
Disagree highly				Agree highly	

The solutions helped me resolve conflicts quicker.

1	2	3	4	5	6
Disagree highly				Agree highly	

The system was easy to use.

1	2	3	4	5	6
Disagree highly				Agree highly	

The presentation format made it easy to understand the solution.

1	2	3	4	5	6
Disagree highly				Agree highly	

B6. Debrief

1. How realistic were the simulation runs?
 2. Did you notice similarities between simulation runs?
 3. Did the system propose solutions that matched your own?
 4. Were the proposed solutions well-timed? Too early? Too late?
-

One of the goals in this research has been to create advisories that reflect how you as an individual prefer to solve conflicts. To do so, we investigate how you solve the same conflict repeatedly over time. In today's simulation, there were only two scenarios, each repeated nine times.

5. What factors of conflict solutions do we have to capture (e.g., timing, aircraft choice, target CPA, etc)?
-

Advisories were based on one of three systems. A third of the proposed solutions that were based on how you solved conflicts in the previous simulation - an individual profile. Another third of the solutions were based on what the group of all controllers implemented on average. Finally, a third was based on what an AI system computed to be optimal.

The advisories varied in terms of timing, aircraft choice, and separation distance aimed for.

6. Does the timing of advisories affect your decision to accept?
 7. Does the aircraft choice of advisories affect your decision to accept?
 8. Does the separation distance of advisories affect your decision to accept?
 9. What other aspects drive your decision making?
-

In the three sessions you saw one of three advisory visualisations: the vector line, the diagram, and the diagram with a text window. [Show the diagram]

10. Which one of these do you prefer? Why?
11. Were the proposed solutions safe? How did this differ across sessions?
12. Were the proposed solutions efficient? How did this differ across sessions?
13. Were the proposed solutions understandable? How did this differ across sessions?

14. If you had a system that made advisories like you saw today, would you accept that system? Why or why not?

15. Do you think that other controllers would trust a system like this? Would controllers ever trust an advisory system?

16. What is your general opinion about how much of the ATC task automation might be able to do in the future?

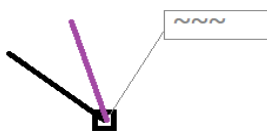
Tell participant:

“Please, do not tell the other participants that solutions were partly based on your own solutions from the previous trial. We are trying to keep that a secret....

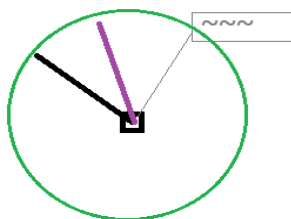
...Do you have any questions for us?

...Thank you for your participation.”

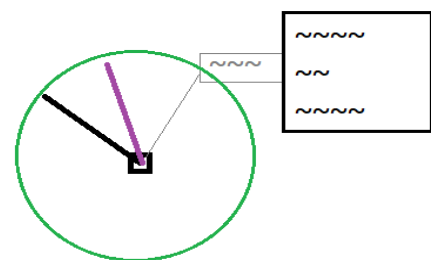
VECTOR



DIAGRAM



TEXT

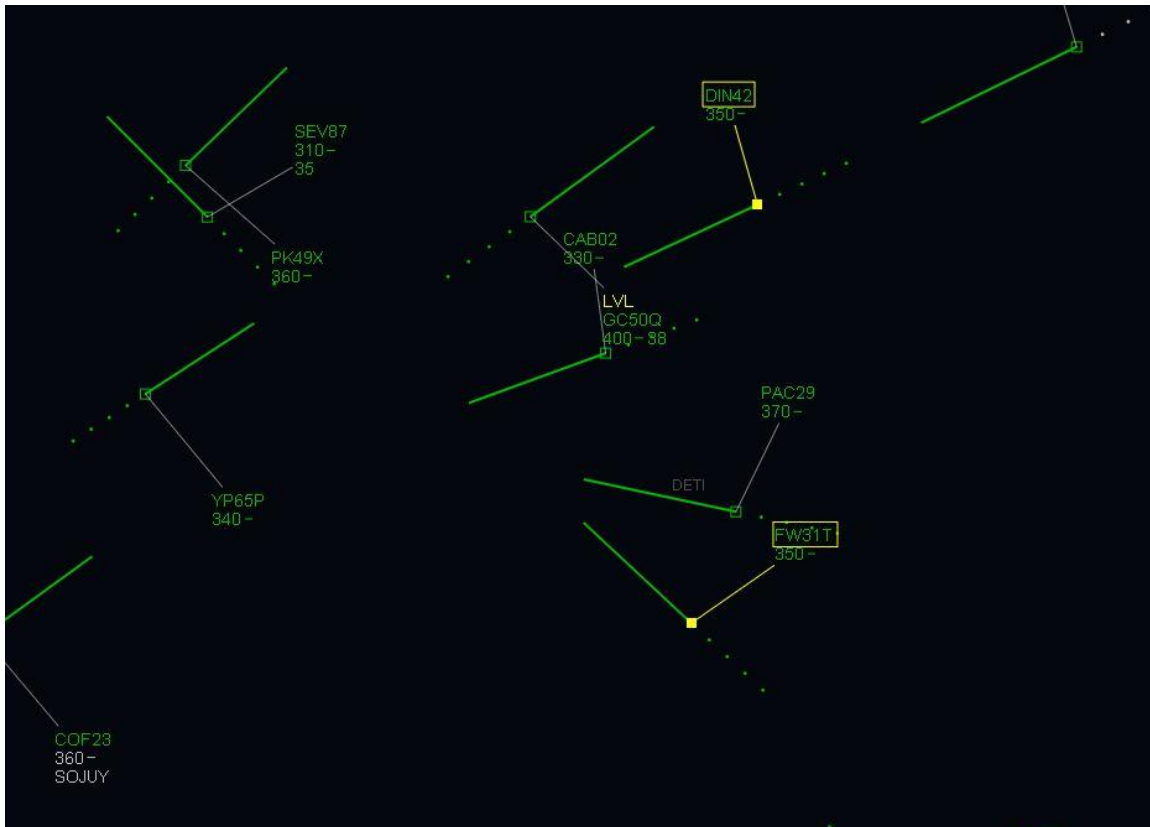


For items 1-7 below, please indicate your agreement with each statement, on a scale of 1 (disagree highly) to 6 (agree highly).

1. I accepted resolution advisories even though I did not agree with them.						
	1	2	3	4	5	6
	Disagree highly				Agree highly	
2. I accepted resolution advisories without inspecting the conflict.						
	1	2	3	4	5	6
	Disagree highly				Agree highly	
3. In the future, computers will do more and more of the controller’s job.						
	1	2	3	4	5	6
	Disagree highly				Agree highly	
4. In the future, computers might be able to perform ATC conflict resolution as well as I can.						
	1	2	3	4	5	6
	Disagree highly				Agree highly	
5. A system like this would make my job less rewarding.						
	1	2	3	4	5	6
	Disagree highly				Agree highly	
6. There is generally more than one acceptable solution to an air traffic conflict.						
	1	2	3	4	5	6
	Disagree highly				Agree highly	
7. Controllers will not accept a system like this.						
	1	2	3	4	5	6
	Disagree highly				Agree highly	

For each of the final two items (8 and 9), think about the presented traffic conflict, and how you would solve it. Then answer sub-items a, b, and c.

8. Consider the conflict between the two aircraft (DIN42 and FW31T) in the picture below. The aircraft are approaching at an angle of 68 degrees with a closest point of approach (CPA) of 0.0 nm.



8a. How would you prefer to solve the conflict using heading? (altitude is not an option)?

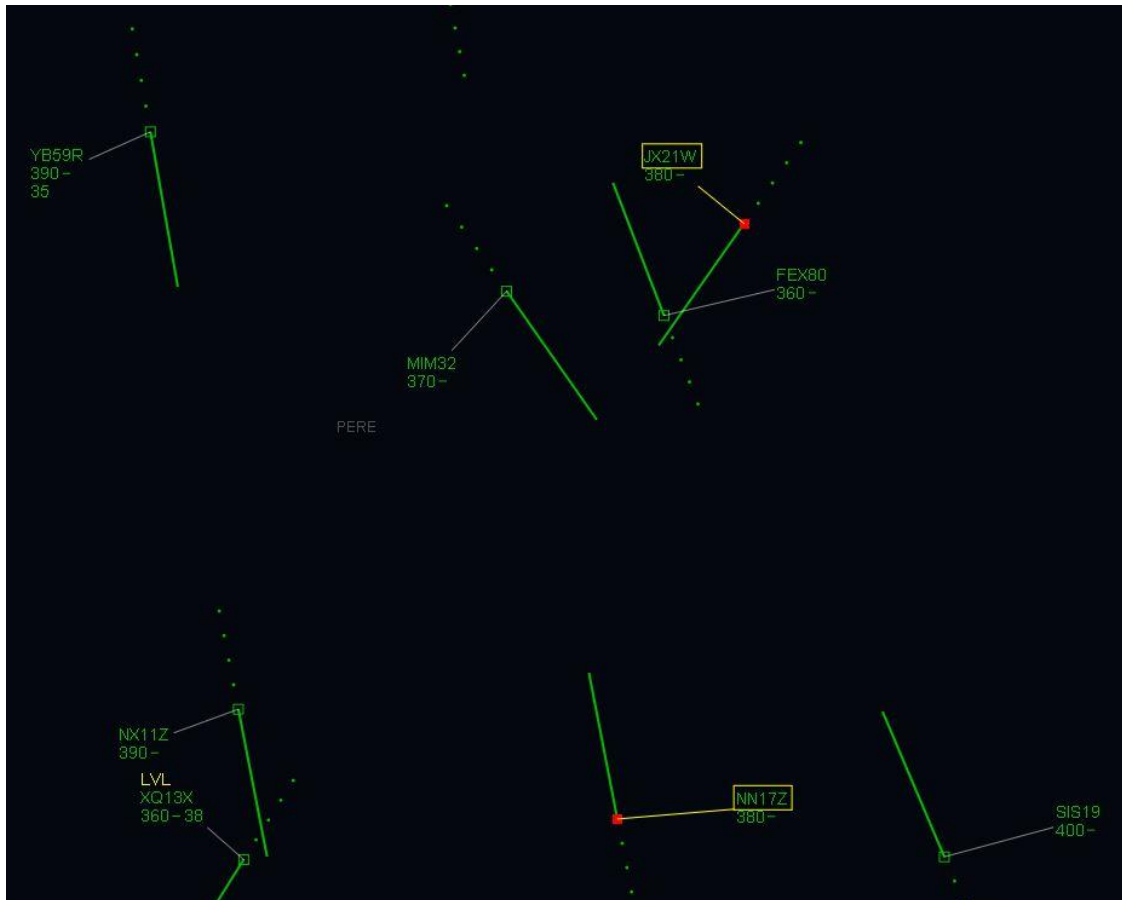
-

8b. Can you provide an explanation for why you choose this solution?

8c. If this conflict were repeated several times, would you solve it the same way every time?

1	2	3	4	5	6
Disagree highly			Agree highly		

9. Consider the conflict between the two aircraft in the picture below (JX21W and NN17Z). They are approaching at an angle of 134 degrees with a closest point of approach (CPA) of 0.0 nm.



9a. How would you prefer to solve the conflict using heading? (altitude is not an option)?

-

9b. Can you provide an explanation for why you choose this solution?

9c. If this conflict were repeated several times, would you solve it the same way every time?

1	2	3	4	5	6
Disagree highly			Agree highly		

ANNEX C: TRAINING PRE-TEST RESULTS

The training pre-test data was analysed in two steps: First, the conflict solution data for all participants was analysed to investigate the variability between controllers across all six scenarios. The results from that analysis were used to determine which scenarios to include in the main experiment. As stated earlier, scenario A and scenario B were chosen. Second, the conflict solution data for scenarios A and B were analysed for each individual participant to determine conformal models.

The bar charts in the following figures shed light on the variability in how participants solved the conflict in scenarios A and B for SIM2A and SIM2B.

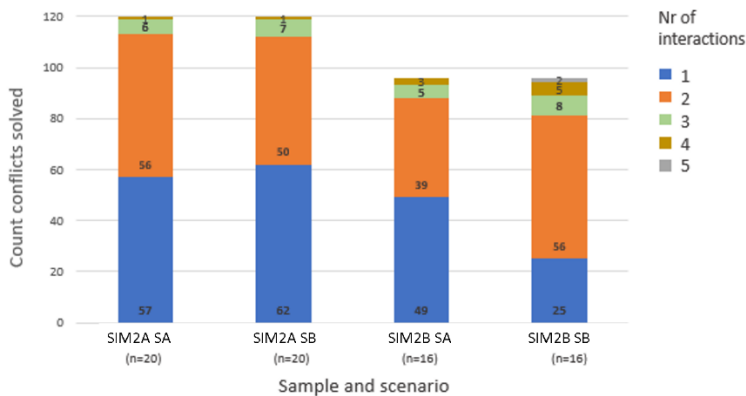


Figure C1. Number of flight interactions, by simulation group and scenario

Single interactions to solve the conflict (i.e. only changing heading or altitude once) accounted for about 50% of all solutions implemented to solve the conflict in scenario A and B, with the exception for scenario B in SIM2B, where only 25 out of 96 solutions consisted of single interactions. Almost equally frequent, and more so in scenario B in SIM2B, were solutions that required two interactions. Participants stated that two interactions were often required because of how close the two aircraft were to each other when the conflict was detected in the scenarios. Normally, a conflict would be detected earlier than was achieved in the scenarios. A limitation of the conformal model was that it proposed a single interaction to solve the conflict.

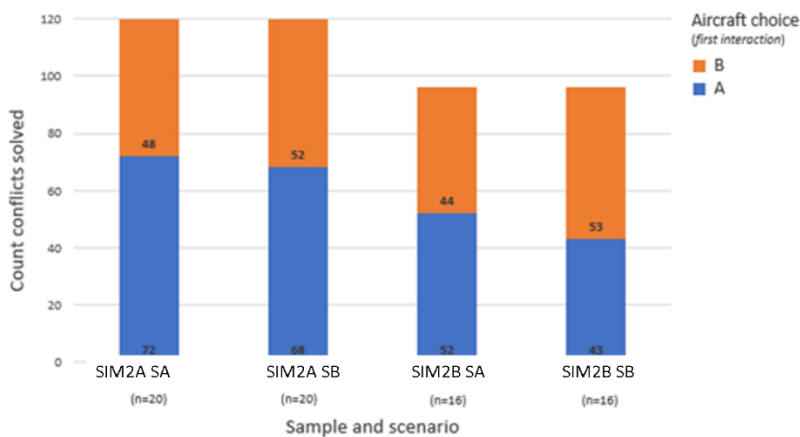


Figure C2. Aircraft choice, by simulation group and scenario

When looking at aircraft choice (only first interaction considered), aircraft A was slightly more often interacted with in both scenarios, except for scenario B in SIM2B where aircraft B was slightly more often interacted with. However, the nearly equal balance across scenarios indicate that the scenarios do not bias interaction with a specific aircraft.

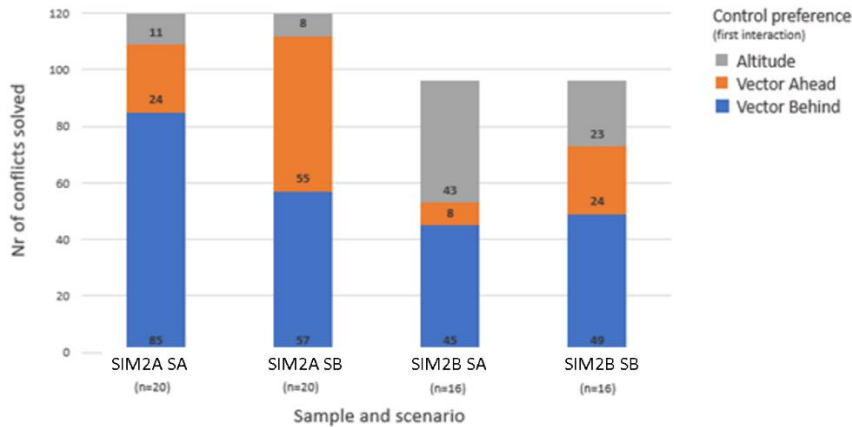


Figure C3. Control preference, by simulation group and scenario

The control preference reflects where spatially the aircraft interacted with is directed with relation to the aircraft it is conflict with. The bar chart shows that participants generally vectored the aircraft interacted with behind the other aircraft. In scenario B in SIM2A, vectoring ahead was implemented almost equally often. This result reflects the general rule-of-the-air that the turned aircraft should be vectored behind the other.

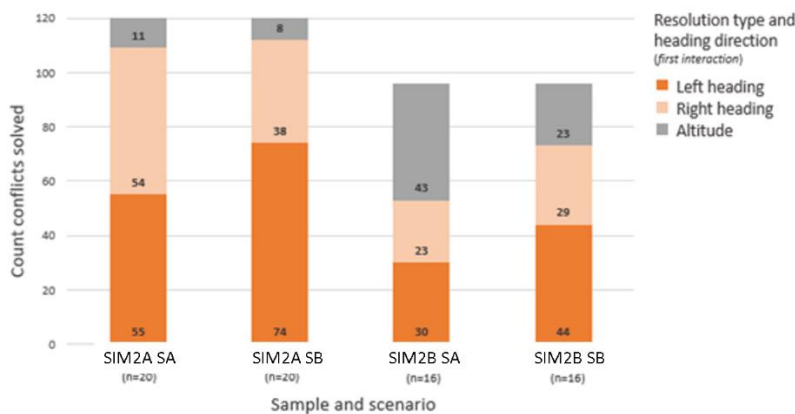


Fig C4. Resolution type and heading direction, by simulation group and scenario

When investigating resolution type, heading solutions were most common in both SIM2A and SIM2B. This is not surprising given that scenarios were designed to favour a heading solution. What is surprising is that so many solutions in SIM2B were level changes. This had a negative effect on the definition of conformal models, which did not consider altitude, and were created to reflect participants’ preferences for heading solutions. When looking at the direction, left or right, it can be seen that left heading changes were more common in scenario B in both samples, while left and right heading changes were more equally implemented in scenario A in both samples.

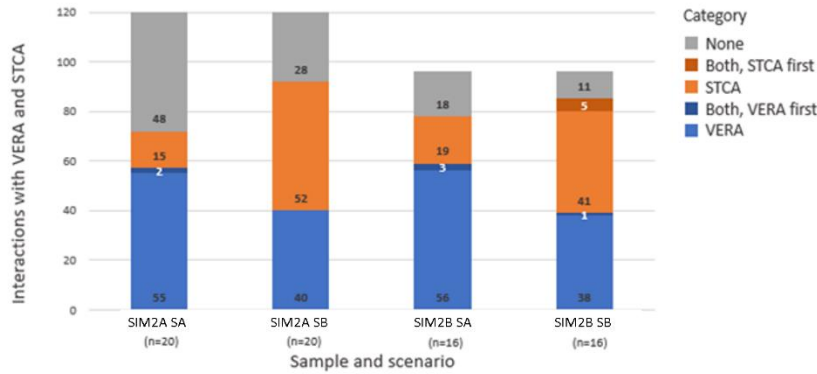


Fig C5. Pre-resolution VERA use and STCA activation, by simulation group and scenario

The bar chart of figure C5 shows the use of VERA and occurrences of STCA before first interaction taken to solve the conflict. STCAs were more common in scenario B than in scenario A. VERA was used more often in scenario A than in scenario B. The number of conflicts solved without use of VERA, or the triggering of STCA, was higher in scenario A. There are very few scenarios in which both VERA and STCA occur, which indicates that when VERA was used, an STCA did not occur, or when STCA was triggered, the VERA tool was not used. The fact that VERA was used more often in Scenario A suggests a proactive behavior where participants more often detected the conflict prior to the STCA warning. The fact that STCA was triggered more often in scenario B suggests reactive behavior where participants more often were notified of the conflict by the STCA being triggered.

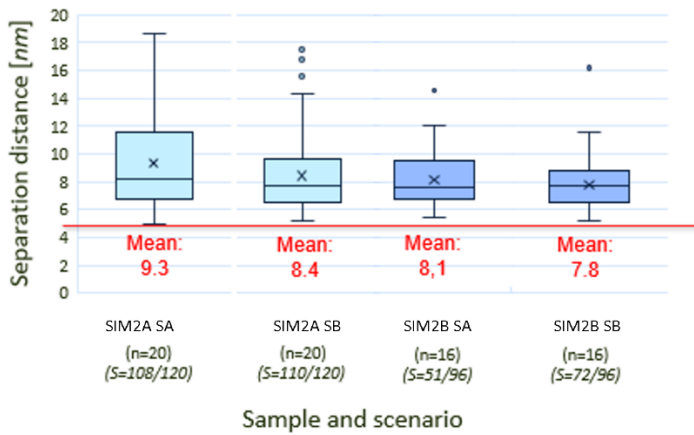


Fig C6. Achieved separation distance, by simulation group and scenario

The boxplot of figure C6 shows the separation distance achieved when solving the conflict in scenario A and B in both samples. The horizontal line in the box represents the median value, while the mean is indicated by an “x” (also annotated in red). The red horizontal line represents the 5 nm separation criterion. Note that only heading solutions are considered. For example, in scenario A of SIM2A, the boxplot reflects 108 solutions out of 120. The missing 11 solutions were eventually solved using altitude. The boxplot shows that participants generally vectored the aircraft in scenario B closer, achieving a mean of 8.4 nm in SIM2A and 7.8 nm in SIM2B.

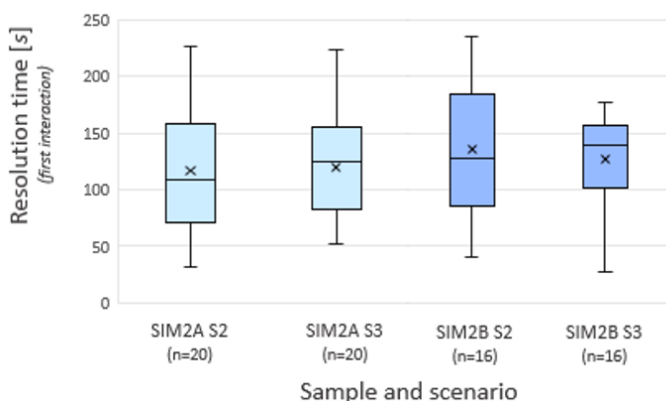


Fig C7. Resolution time, by simulation group and scenario

The resolution time shown in figure C7 represents the time after scenario start when the first interaction was taken to solve the conflict. The boxplot shows a similar resolution time for both scenarios. When looking at the median and mean values, it can be seen that participants interacted to solve the conflict in scenario A earlier than the conflict in scenario B. This is noteworthy given that the conflict in scenario B occurred earlier than that in scenario A (see table depicting scenario design). In other words, participants had interacted with the conflict in scenario B closer to the CPA.

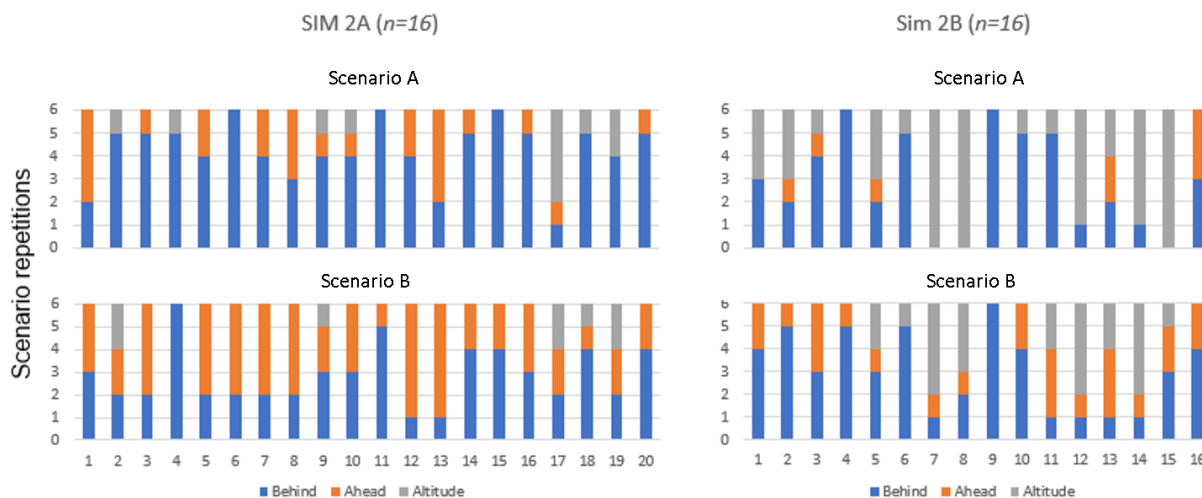


Fig C8. Resolution geometry, by simulation group and scenario

There was a clear preference for vectoring one aircraft behind the other in both scenarios in SIM2B and scenario A in SIM2B. In scenario B, SIM2A, there is a balance between vectoring ahead and behind. Overall this indicates that vectoring behind can be considered a more general preference. Note that the group model advisory, in for both samples and both scenarios, proposed to vector aircraft A in front of aircraft B. This is against what controllers did in scenario A, SIM2A and scenario A and B in SIM2B.

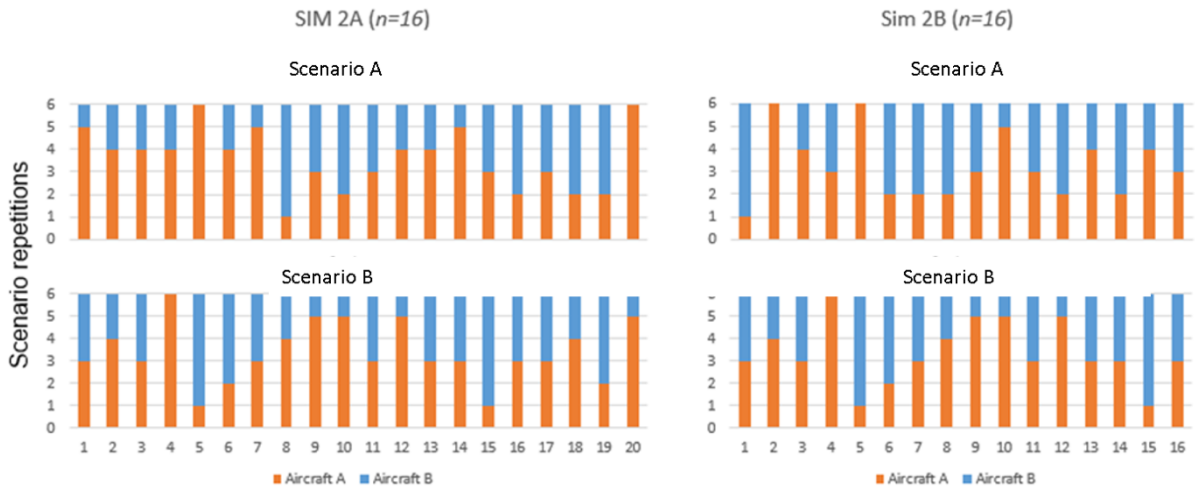


Fig C9. Aircraft choice, by simulation group and scenario

Note that very few participants consistently interacted with same aircraft in all scenario repetitions. There are also not that many that interacted with the same aircraft in five out of six repetitions. This indicates that aircraft choice may not be that important, which could itself be an artifact of using collision (CPA=0) scenarios, which do not present a preferred aircraft to maneuver.

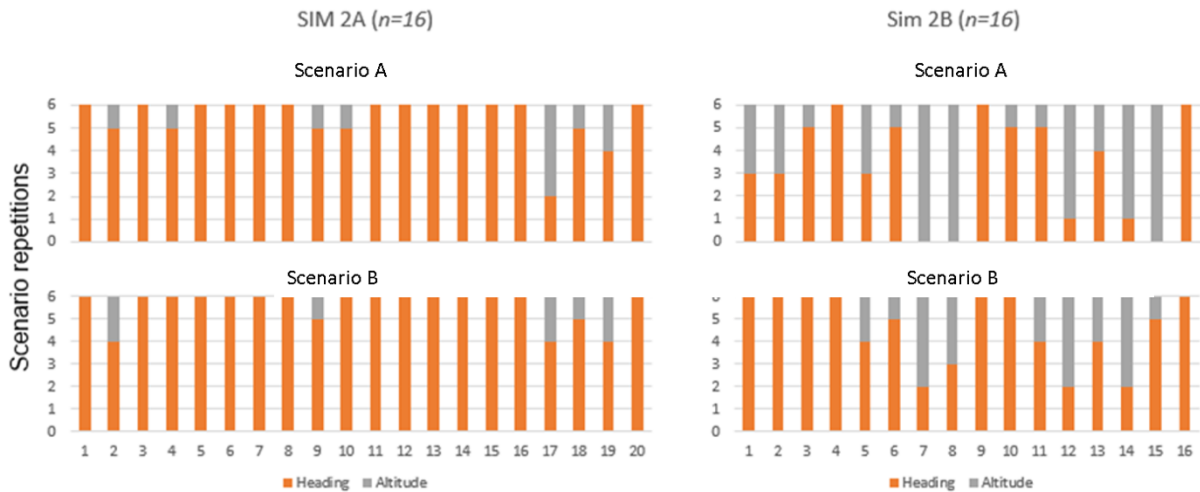


Fig C10. Resolution choice, by simulation group and scenario

Note that several participants in SIM2B consistently solved conflicts with altitude in scenario A. It can also be seen that altitude was used by the same participant in scenario A and scenario B, indicating that resolution type is an important individual difference.

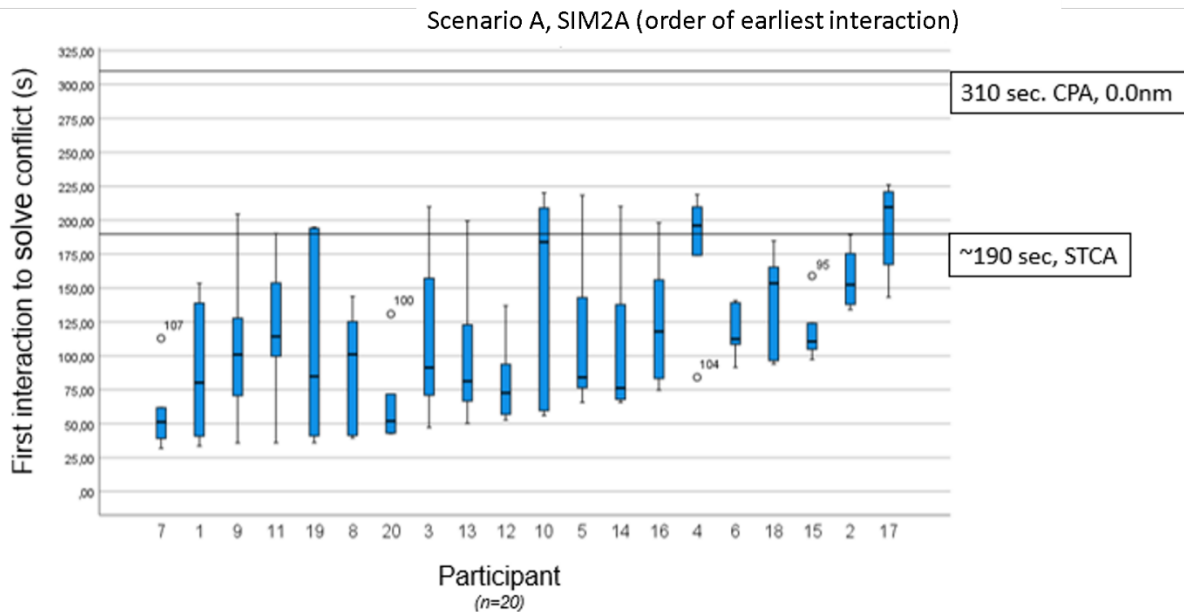


Fig C11. Time to first interaction, SIM2A Scenario A.

The plot shows that participants worked more proactively to solve the conflict, with conflict detection triggered by them finding the conflict on their own, and often using VERA.

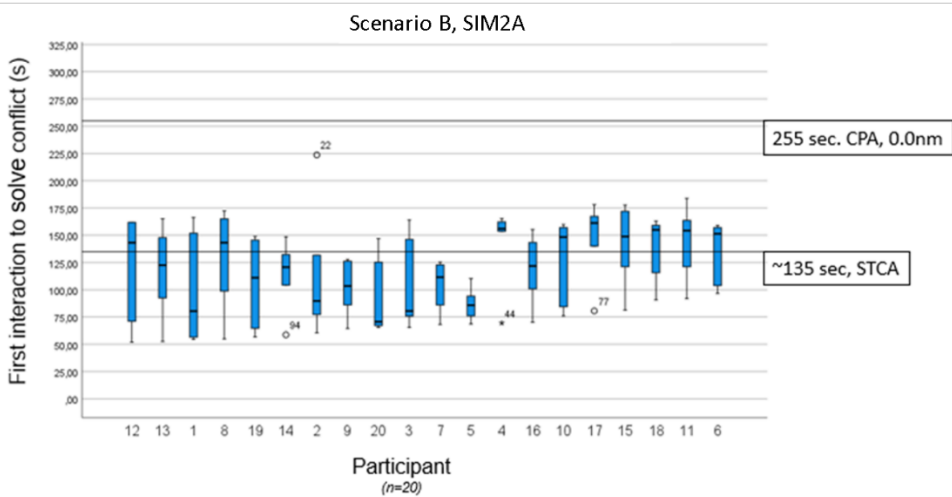


Fig C12. Time to first interaction, SIM2A Scenario B.

The plot of figure C12 shows that participants worked more reactively to solve the conflict, with conflict detection triggered by STCA.

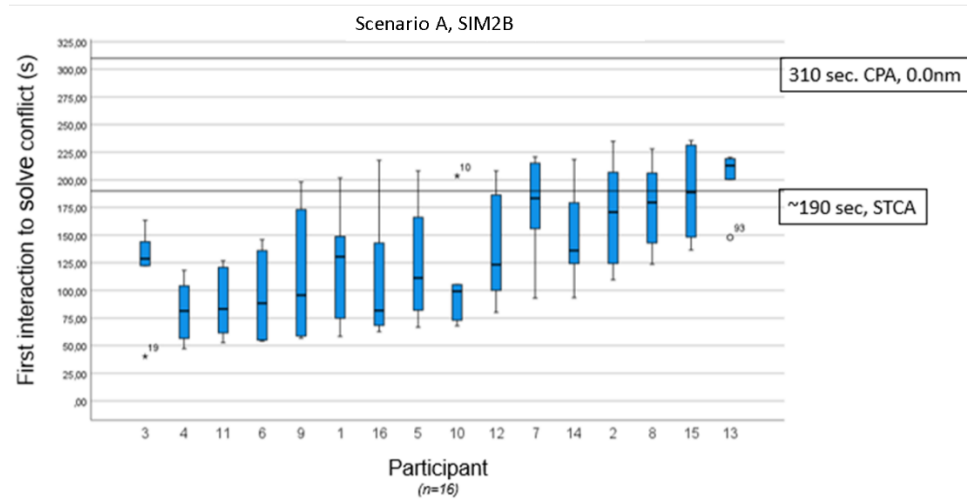


Fig C13. Time to first interaction, SIM2B Scenario A.

The plot of figure C13 shows that participants worked more proactively to solve the conflict, with conflict detection triggered by them finding the conflict on their own, and often using VERA.

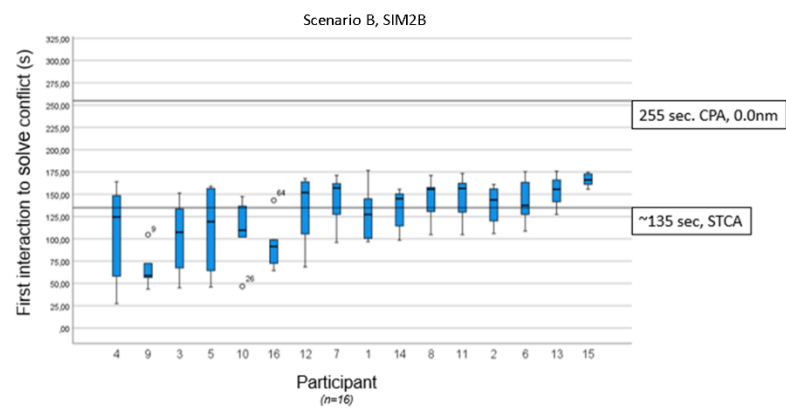


Fig C14. Time to first interaction, SIM2B Scenario B.

Figure C14 shows that participants worked more reactively to solve the conflict, with conflict detection triggered by STCA.

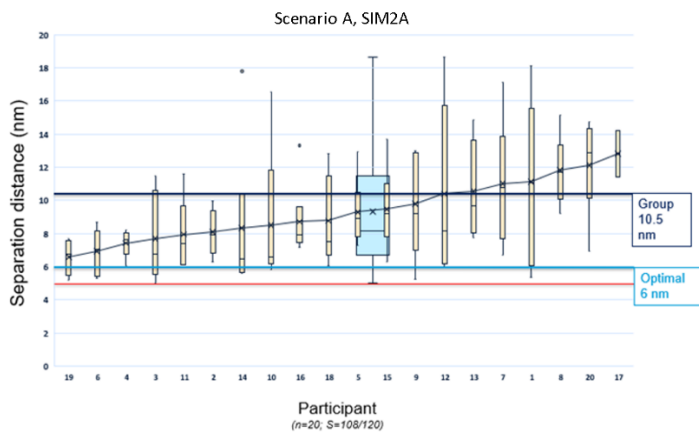


Fig C15. Separation distance (nm), SIM2A Scenario A.

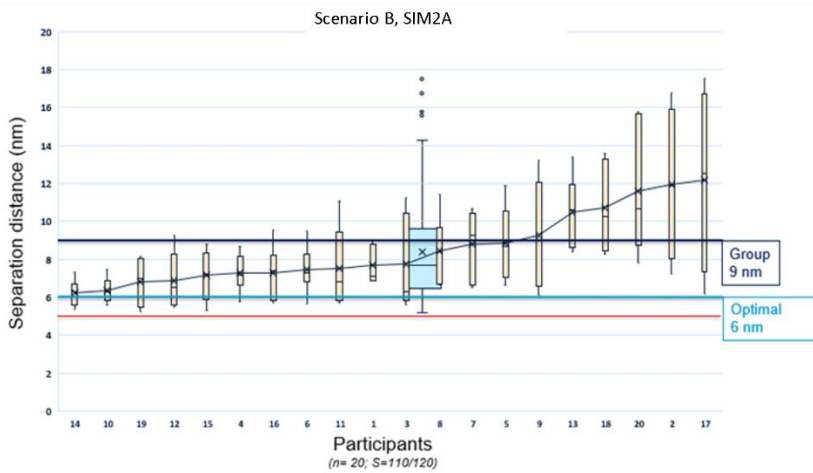


Fig C16. Separation distance (nm), SIM2A Scenario B.

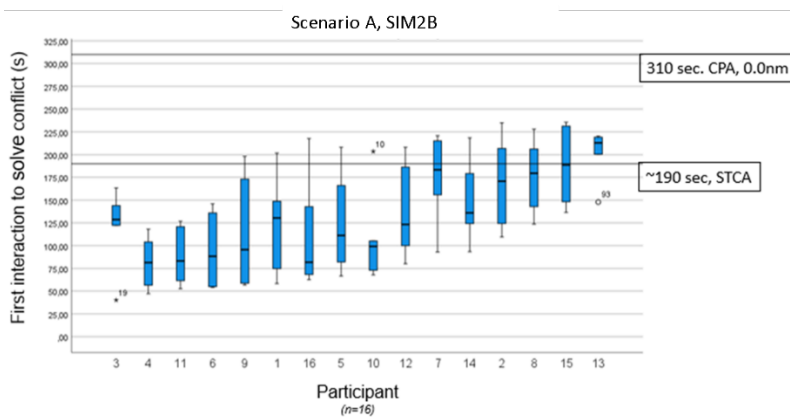


Fig C17. Separation distance (nm), SIM2B Scenario A.

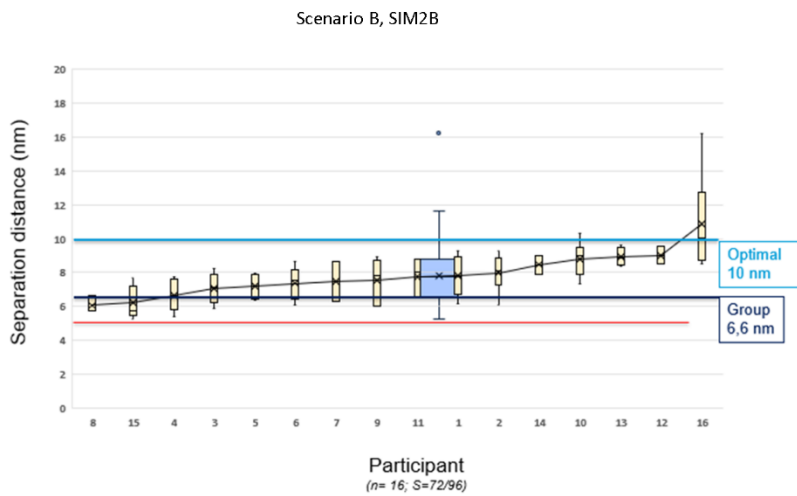


Fig C18. Separation distance (nm), SIM2B Scenario B.

The results presented in section 3 show that conformance and transparency had an effect on how groups responded to the optimal advisories in scenarios A and B in SIM2A, and scenario B in SIM2B. The general difference between the two groups is that, the group whose preferred separation margin was closer to the optimal, was less likely to modify the advisory (nudge, adjust, change, reject), had higher agreement ratings, smaller adjustments to separation margin (delta CPA) and rated advisories more similar to how they would have solved it themselves. These effects appear stronger for T1 (diagram) and T2 levels (text), with T0 (vector) levels being more equal between groups.