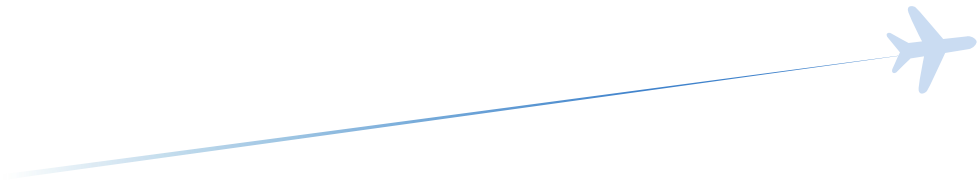

Experimental design document

Deliverable ID:	**D6.1**
Dissemination Level:	**Public**
Project Acronym:	**MAHALO**
Grant:	**892970**
Call:	**H2020-SESAR-2019-2**
Topic:	**SESAR-ER4-01-2019**
Consortium Coordinator:	**DBL**
Edition Date:	**4 Jan 2022**
Edition:	**00.02.00**
Template Edition:	**02.00.02**

Founding Members
EUROPEAN UNION EUROCONTROL
SESAR JOINT UNDERTAKING

Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
Carl Westin (LiU)	WP6 Leader	13/08/21
Brian Hilburn (CHPR)	WP2 Leader	09/09/21
Clark Borst (TUD)	WP5 Leader	17/09/21

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
Clark Borst (TUD)	WP5 Leader	11/10/21
Brian Hilburn (CHPR)	WP2 Leader	11/10/21
Carl Westin (LiU)	WP6 Leader	11/10/21
Magnus Bång (LiU)	WP3 Leader	11/10/21
Martin Christiansson (LFV)	Project Contributor	11/10/21

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

Name/Beneficiary	Position/Title	Date
Stefano Bonelli/DBL	Project Coordinator	14/10/2021
Stefano Bonelli/DBL	Project Coordinator	04/01/2022

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date

Document History

Edition	Date	Status	Author	Justification
00.00.01	13/08/21	Doc created	Carl Westin	Deliverable to Mahalo management
00.00.02	09/09/21	Internal review	Brian Hilburn	Deliverable to Mahalo management

00.00.03	17/09/21	Internal review	Clark Borst	Deliverable to Mahalo management
00.00.04	08/10/21	Internal Release	Brian Hilburn	Deliverable to Mahalo management
00.00.05	11/10/21	Internal Release	Matteo Cocchioni	Deliverable to Mahalo management
00.01.00	14/10/21	Final Release	Stefano Bonelli	Deliverable approved for submission to the SJU
00.01.01	10/12/2021	Internal Release	Brian Hilburn	Reopen for revision
00.02.00	04/01/2022	Final Release	Stefano Bonelli	Deliverable approved for submission to the SJU

Copyright Statement

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 Experimental Plan, deliverable D6.1 of the MAHALO project. D6.1 captures the research team's planned approach to conducting WP5 integration trials (Sim 1, WP5 Task 5.2), as well as WP6 simulations 2A and 2B. Notice that this D6.1 experimental plan feeds WP5 activities (ML and E-UI integration, along with integration trials of WP5 Task 5.2) both iteratively and interactively. That is, the experimental plan drives the integration of ML and interface, but WP5 integration activities have also fed the development of the current D6.1.

This experimental plan follows the structure suggested in SJU's *SESAR 2020 Experimental Approach Guidance ER* document. This D6.1 is based, in part, on the previously submitted D2.2 Concept Report, which outlined the operational concept underlying MAHALO. The present report is the validation plan for the broader MAHALO concept, and is distinct from the previously submitted D4.2 E-UI Validation Report, which limited itself to evaluation of the SectorX user interface.

This D6.1 document first provides an overview of the MAHALO project, including its research questions, the relationship between experiments and project objectives, and the inter-relationships between experiments.

Second, it covers the objectives, general approach, and methodology underlying the experimental plan.

Third, it covers the experimental approach itself, including specific testable hypotheses, dependent and independent variables, external validity generalizability considerations, choice of statistical tests, and all procedures associated with data collection and analysis.

Fourth, this report reviews data and software input requirements, including data types, sources, and formats. This also includes data reduction and postprocessing requirements, and specifying the link between specific results and the research hypotheses of the experiments.

Fifth, and finally, this report considers research coordination and development. This includes procedures for data storage, security and access (e.g. open source / source code availability, as appropriate), steps to ensure adequate methodological specification to facilitate experimental replication by outside researchers, if desired.

Table of contents

1. Overview of the Experimental design document

The MAHALO project has 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, focusing on such constructs as mental workload, acceptance, trust, reliance, and human – machine system performance. To achieve these two ambitious goals, the MAHALO project started from a clear specification of its concept of operations (ConOps). This ConOps was presented in the previous D2.1 *Concept of Operations* report.

This document is the Experimental Plan, deliverable D6.1 of the MAHALO project. D6.1 captures the research team's planned approach to conducting WP5 integration trials (Sim 1, WP5 Task 5.2), as well as WP6 simulations 2A and 2B. The aim of the current document is to present a detailed experimental plan that specifies the project's research questions and hypotheses, experimental design, data collection methods, and data analysis protocols. This Experimental Plan follows the structure suggested in SJU's *SESAR 2020 Experimental Approach Guidance ER* document. This D6.1 is based, in part, on the previously submitted D2.2 Concept Report, which outlined the operational concept underlying MAHALO. The present report is the validation plan for the broader MAHALO concept, and is distinct from the previously submitted D4.2 E-UI Validation Report, which limited itself to evaluation of the SectorX user interface.

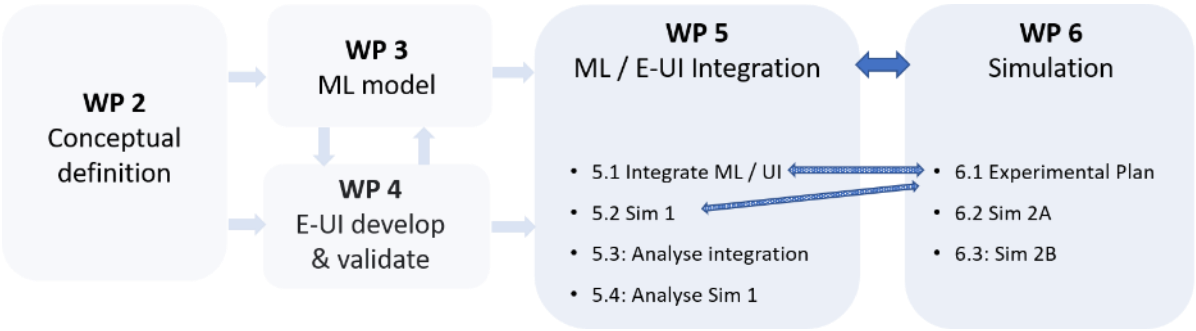

Fig. 1.1. WP6 (including WP6.1 *Experimental Plan*) within the MAHALO technical work package flow.

1.1 Goals of the Experimental Plan

D6.1 elaborates the following general elements of the MAHALO experimental plan:

- Research questions and hypotheses—what general questions does MAHALO intend to ask via human-in-the-loop (HITL) simulation (e.g., "How does the strategic conformance and transparency of a ML-based CD&R advisory system affect controllers' understanding, trust, acceptance, etc.?"), and what specific and testable hypotheses will we test (e.g., "controller self-reported acceptance of advisories will be higher if those advisories are based on solutions that conform to the controllers preferred solution in the same context").
- Experimental design— at the heart of this experimental plan is experimental design, which specifies: levels and ranges of independent variables, operationalised and measurable dependent variables, and a scheme for assigning participants (ATCOs) to experimental conditions; methods for experimental control; and general analysis approach.
- Data collection methods— what procedures, schedule, test scenarios, and materials will be used to conduct HITL simulations?
- Data analysis methods—what pre-processing, data reduction, analysis tools, statistical tests, and decision criteria will be used to analyse HITL data?

1.2 Report structure

The following four chapters (2-5) are organised as follows:

Chapter 2 outlines the objectives, general approach, and methodology for MAHALO, the projects broad research question. Next, chapter 3 covers the heart of the project's experimental approach, including

- hypotheses,
- variables (both independent- and dependent variables),
- test matrix,
- validation objectives,
- test scenarios (reference versus solution scenarios),
- participants and experimental assignment,
- data collection methods (including equipment, materials, and procedures), and
- data analysis methods (candidate statistical tests, decision rules, etc.).

Next, chapter 4 covers data format issues and the project data protection approach, including data file structure and contents. Finally, chapter 5 summarizes the project's research coordination and development considerations, including methods for dissemination, experimental replication assurance, and issues around source code / simulation platform ownership and distribution.

2 Objectives, General Approach, and Methodology

The MAHALO project started from simple questions: In the emerging age of Machine Learning, should we be developing automation that is conformal to the human, or should we be developing automation that is transparent to the human? Do we need both? Further, are there trade-offs and interactions between the concepts, in terms of operator trust, acceptance, or performance?

To answer these questions, the MAHALO team has been, first, defining an ATM Concept of Operations and User Interface on which to base this work (see deliverable D2.2 Concept of Operations report, earlier in this series); Second, the team has been developing an automated conflict detection and resolution (CD&R) capability, realised in a prototype Machine Learning (ML) hybrid system of combined architectures.

2.1 Research Question

The aim of the current work, reported here, is to use these foundations to now address specific research questions, as originally laid out by MAHALO, and to experimentally evaluate, using HITL simulations, the relative impact of *conformance* and *transparency* of advanced AI, in terms of e.g. controller trust, acceptance, workload, and human/machine performance. The broad research question to be addressed is:

> 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?

2.2 Simulation planning

MAHALO plans to conduct the following three human-in-the-loop (HITL) experiments:

- **Simulation 1:** The first simulation (currently planned for October 2021) will use novices (e.g., university students). This is a developmental simulation, aimed at testing the fully integrated ML CD&R system, and its ability to provide conformal and transparency advisories. This is also the first experiment for testing the scenarios, data collection protocols, experiment procedures, questionnaire and debriefing materials, and data analysis procedures. Simulation

1 is aimed at validating our simulation and analysis procedures, not at answering the MAHALO research hypotheses. These will be addressed in simulations 2A and 2B.

- **Simulation 2A:** The second simulation (planned for Dec 2021-Jan 2022), hosted by DBL and ANACNA in Italy, will involve 20 ATCOs as participants. The two phases (Conformance Pre-Test and Main Experiment, as discussed later) are scheduled for early Dec 2021 and late Jan 2022.
- **Simulation 2B:** The third simulation (currently planned for early 2022, also in the same two phases), hosted by LFV Sweden, will involve 16 ATCO participants. This will replicate the 2A simulation, with a different controller cohort.

3 Experimental Approach

3.1 Hypotheses

Building on the research questions posed in the previous chapter, the MAHALO project states the following testable hypotheses, that will be addressed in simulations:

Hypothesized main effects

- Hypothesis 1: controller self-reported acceptance of advisories will be higher if those advisories are based on solutions that conform to the controllers preferred solution in the same context;
- Hypothesis 2: controller self-reported acceptance of advisories will be higher if those advisories are presented in a high transparency display format, as discussed in section 3.3.2;
- Hypothesis 3: Conformal solutions (as provided by the SL system) will be associated with higher acceptance/agreement than will optimized solutions (as provided by the RL / algorithmic system[1]). This assumes that conformal and optimal solutions differ (see also section 3.2.1);
- Hypothesis 4: Conformal solutions will be associated with higher reported trust, than will optimized solutions;
- Hypothesis 5: Both transparency and conformance manipulations will be associated with a decrease in reported workload.

Hypothesized interaction effects

- Hypothesis 6a: Under low transparency, conformal advisories will be more accepted/agreed upon than will optimized advisories;
- Hypothesis 6b: Under high transparency, this trend will be less pronounced, and the difference in acceptance/agreement between conformal an optimal advisories will be smaller.

[1] An algorithmic (deterministic) CD&R system is being developed, as a possible adjunct to RL modelling.

3.2 Variables

Experimental design distinguishes *independent* and *dependent* variables. The former are experimentally manipulated in a controlled way, whereas the latter are the measured results of experimental manipulations. That is, an experiment manipulates independent variables, and measures dependent variables. For the Main Experiment, there are two independent variables - Conformance and Transparency - as defined below.

3.2.1 Independent Variable 1: Conformance

Conformance will be experimentally manipulated in three non-orthogonal levels, characterised by underlying ML architecture (SL vs RL), and prediction model (individual vs group).

a. C: Conformal (Supervised learning ML, Personalized prediction model)
b. GC: Group conformal (Supervised learning ML, Group prediction model)
c. NC: Non-conformal (Reinforcement learning ML, Optimized reward prediction model, with possible algorithm augmentation)

The conformance of resolution advisories will be varied in three different ways: conformal to the individual (personalized prediction mode); conformal to the group (group prediction model), and nonconformal (optimized prediction model). The ability to provide conformal resolution advisories requires a three-step design process as illustrated in Fig. 3.1.

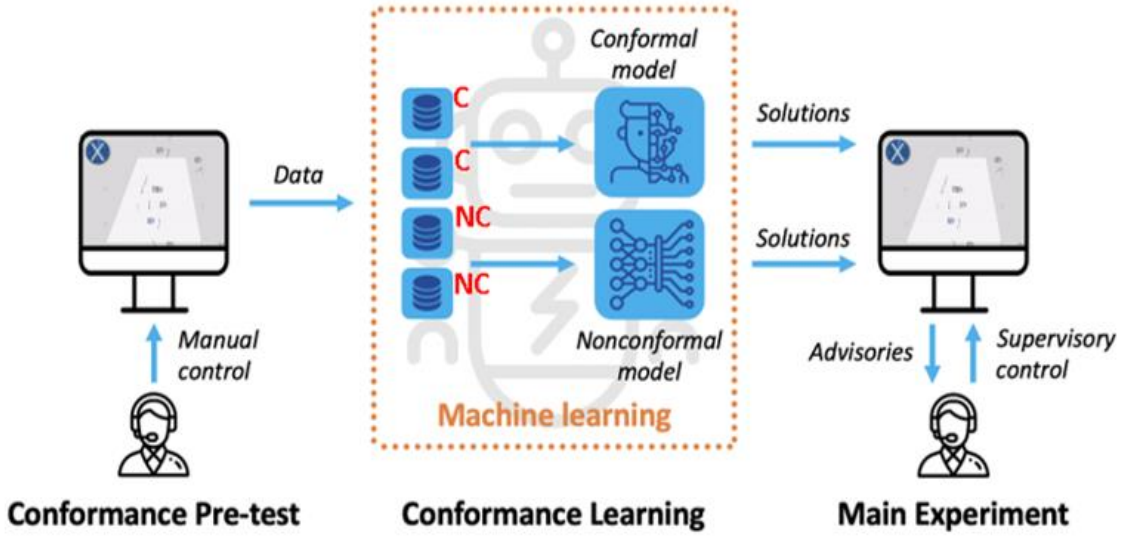

Fig. 3.1. Advisory conformance design process

It is important to note that each simulation will consist of two parts: a Pre-Test, and a Main Experiment. Only in the Pre-Test will controllers manually resolve pending conflicts. The Main Experiment will, in essence, be a supervisory control task. Notice that the dependent measures to be used in the Pre-Test and Main Experiment will therefore differ. As discussed later, the Pre-Test will collect objective behavioural such as response time, type and degree of chosen resolution manoeuvre, etc. The Main

Experiment, on the other hand, will provide information on such measures as acceptance, agreement, workload, etc.

The first simulation, the **Pre-Test**, is conducted to collect data on how different individuals solve different conflicts. At this stage, individuals solve conflicts 'manually' (i.e., no decision support is provided). Fig 3.2 overviews the data collected for each solution generated by participants.

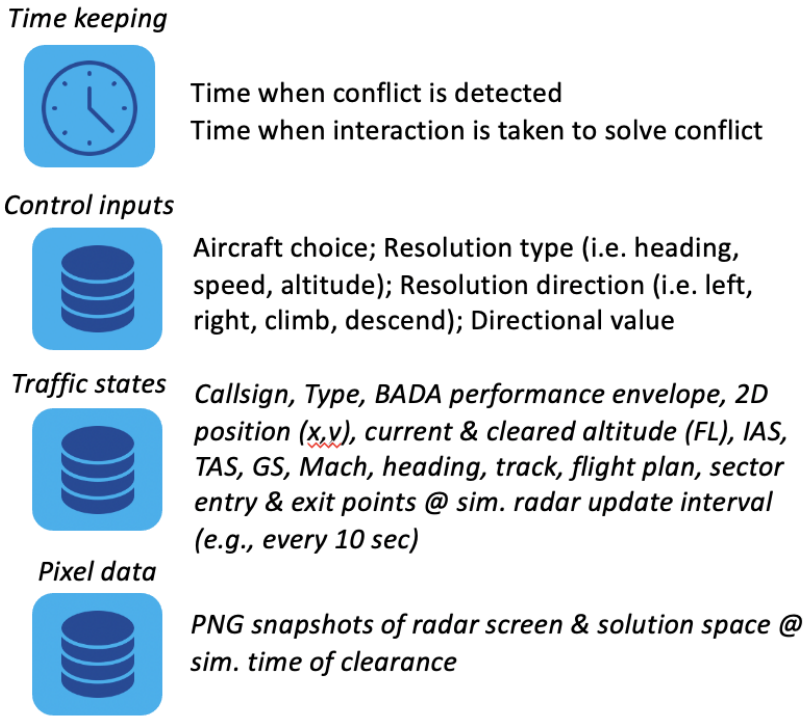

Fig. 3.2. Data collected in Pre-Test

Precisely determining conflict detection time and reaction time to presented conflicts, can be challenging. Response time can be inferred, only if a controller takes some action. Notice also that absence of an overt response does not necessarily indicate a lack of response -- a controller might perceive a conflict, and decide to take no action at the moment. Similarly, identifying at what point in time the controller actually perceived the conflict can be challenging. Eye tracking measures (e.g., eye point of gaze) allow determination of when on-screen elements were fixated, and thereby allows a partial inference about conflict detection time.

The need to determine conflict detection time and reaction time underlies the timing of resolution advisories on-screen. Advisories that are presented too early can increase workload and confusion, and advisories that are too late are likely of no use (since the controller will probably already have devised their own solution). Luckily, MAHALO can rely on previous research from the MUFASA project into the timing of advisories, which showed that the required accuracy in identifying the time when a conflict was detected is rather low (uncertainty is expected to be in range of several seconds). The main purpose of identifying the time for conflict detection in the Pre-Test is for all the ML models to provide conflict resolution advisories before the participant detects the conflict in the Main Experiment runs. Ideally, the resolution advisory is provided before the participant has come up with a solution for the conflict and implemented it.

Conformance is defined as the similarity between two conflict solutions. But how, exactly, should that similarity be defined? If a controller solves a conflict with a heading solution, but automation solves it using an altitude solution, those are clearly not conformal solutions. But what if the controller turns a given aircraft 10 degrees right, and the automation turns it 15 degrees right? Or turns the other aircraft left? Clearly, conformance is not a binary 'yes-or-no' dimension. The MUFASA project [1] proposed a conflict resolution framework (table 3.1) that classifies conflict solutions in a hierarchy where each step considers the solution in more detail. For example, the first stage considers the governing solution strategy (solution parameter hierarchy, control problem, or solution geometry). In the control preference strategy, the conflict is viewed as a control problem, focusing on the control action required to solve the conflict (e.g., vector aircraft ahead or behind). This preference considers that conflicts only can be solved by one aircraft going behind, in front, above, or below the other [1].

In the geometry preference strategy, solutions are based on the desired spatial relationship between the involved aircraft. It acknowledges that a solution is based on the spatial orientation between two or more aircraft and their constraints as they evolve over time, rather than on discrete information about aircraft state and position [1]. A more detailed consideration of solutions could include consideration of aircraft choice (e.g. aircraft A or B) and Resolution type (e.g. heading or altitude). The conformance levels of solutions that participants implement can then be classified according to the conflict resolution framework of Figure 3.3.

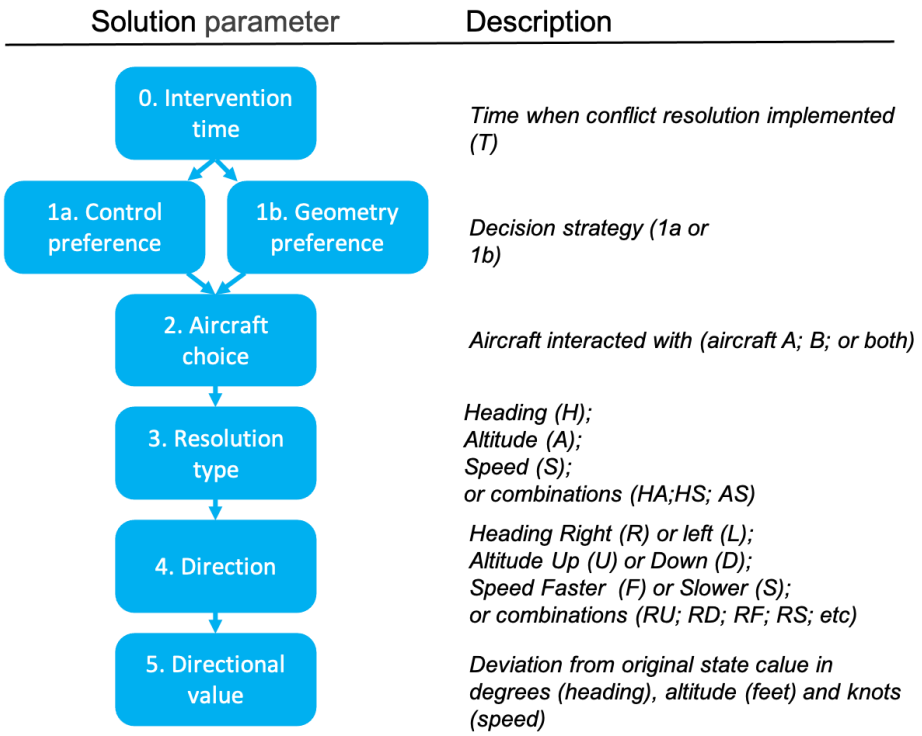

Fig. 3.3 Conflict resolution framework (adapted from [1])

It is not guaranteed that the conformance ML models used as a conformance variable will provide solutions according to the highest level of conformance (level 5). The conformance level that a conformal ML model can support depends on how consistent the participant is in solving similar conflicts over time. The objective is to support the highest level of conformance possible that participants in our sample support. In general, the conformance level of the conformance variable will

depend on the participant that is least consistent. If all participants perform consistently according to level 5 conformance, we will be able to provide conformance at that level. Based on previous experience, however, we expect participants' performance in solving conflicts to vary - some are more consistent than others. We will control the variable by establishing a conformance level that applies to all participants, for which the least consistent participant will be limiting. We can also elect to discard a participant, if that participant's performance turns out to be an outlier. Alternatively, we could also to divide participants in groups depending on how consistent they are (and that the sample size allows for doing).

We also acknowledge that there may be other aspects of CD&R performance that reflect individual differences, such as preferences for the distance between aircraft at their closest point of approach. Some controllers may aim for at least 6NM separation while others aim for 10NM. MAHALO will investigate these and other candidate conformance parameters in participants' CD&R performance.

The solution data collected in the Pre-Test is used train the ML system. This **Conformance Learning stage** (the middle step pictured in figure 3.1) will take place between the pre-test and main experiment phases. During this conformance learning stage, different ML models will be trained according to the three conformance levels (personalized, group, and optimized prediction models). All models must learn from the conflict resolution data what characterises a conformal solution for each individual participant. This means that the ML agent has to build a conformal model for each participant (who took part in the Pre-Test). As such, a model will be created for each individual. Note that, in contrast, there will only be one group conformance model for each pooled group of participants (i.e., Italian controllers and Swedish controllers). Furthermore, there will only be one optimized ML model. Importantly, the optimized ML model will suggest conflict resolutions that are non-conformal to the individual's conformal model (assuming ML-based and controller-generated solutions differ, as discussed in the following subsection). An example of a conformal and nonconformal solution is shown in Fig. 3.4. Here, conformance is defined as level 1 conformance. In this example, the conformal solution is similar to the preferred solution: aircraft A is vectored to the right behind aircraft B. The nonconformal solution aircraft B is deviated to the left, behind aircraft A. The nonconformal solution consists of vectoring aircraft B to the left behind aircraft: another aircraft is chosen for resolving the conflict.

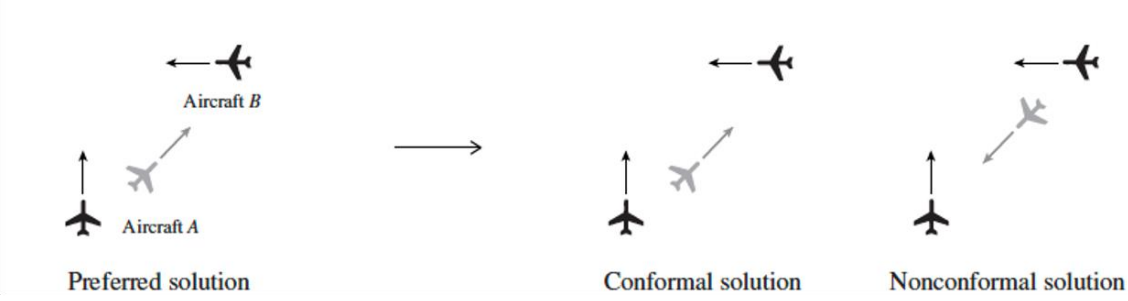

Fig. 3.4. Example of conformal and non-conformal solutions

In the final **Main Experiment**, the same participants who took part in the Pre-Test play the identical simulator and scenarios. Only this time, they will receive CD&R advisories from the ML agent. Only solutions provided by the ML agent with a personalized prediction model will be conformal. Solutions provided by the ML agents with a group prediction model and optimized prediction model will be non-conformal.

3.2.2 Independent Variable 2: Transparency

Automation transparency is achieved by providing additional layers on the ecological user interface with information about the underlying rationale for the solution recommended by the ML agent. Transparency will be varied in three levels:

T0: **No transparency.** Advisory condition where the ML agent recommends a solution to a conflict (e.g. in terms of a vector) without providing any underlying rationale.

T1: **Domain transparency.** How the ML agent has derived a particular solution. The visualized solution spaces can be used to present their recommended solutions. And highlight what information was considered that led to the advice.

T2: **Agent transparency/Conformance rationale.** The ML agent can present why it considers a solution conformal or nonconformal by explaining why a particular solution matches or deviates from the individual's preferred solution (i.e., why it is/is it not strategic conformal). For T2 we will explore ML interpretability methods (see D2.1 for examples) for identifying the relationship between input data and output solution.

According to our model of automation transparency (figure 3.5), these three experimental levels correspond to baseline (zero transparency), domain transparency, and agent transparency, respectively. Notice that domain and agent transparency (termed T1 and T2 in our experiments) are *nonorthogonal* in statistical design terms. The two are stacked, and agent transparency assumes domain transparency. As a result, they can never be compared in isolation.

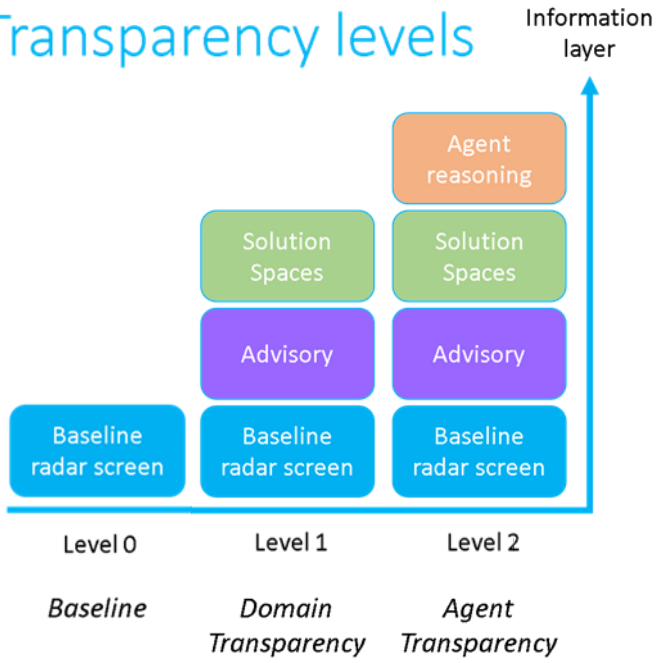

Figure 3.5. Levels of automation transparency

Test matrix, Main Experiment

For the Main Experiment, we plan a fully crossed factorial design, with three levels of each of the two independent variables. Factors are to be presented 'within subject,' meaning that each ATCO will be presented all nine experimental conditions. We intend to follow a Latin square experimental design, which is a useful way to control for confounding variables, by ensuring that treatment condition appears only once in each row and each column, as shown in figure 3.7.

Transparency IV

Conformance IV	Solution	Domain	Agent
Low	A	B	C
Medium	D	E	F
High	G	H	I

Figure 3.7. Test matrix of experimental conditions.

3.2.3 Dependent measures

The heart of the MAHALO experimental plan is the Main Experiment, when we will actually evaluate the impacts of transparency and conformance manipulations. The Pre-Test is used only to identify controller's individual (and group) preferred solution strategies, so as to enable the conformal manipulation in the Main Experiment.
The Main Experiment will collect the following data:

- Acceptance of solution (binary yes or no performance measure, converted to accept/reject ratios)
- Response time to solution (ms from onset)
- Agreement with advisory, self-reported (onscreen scale 0-100)
- Understanding of advisory, self-reported (questionnaire items, based on [11] [12][13])
- Workload, self-reported (questionnaire 0-100 scale))
- Trust elements, subjectively reported (questionnaire 0-100 scale, based on [2][3][4][5])

Several checklist and survey instruments are available to assess trust in computers and automation, including: the Checklist for Trust Between People and Automation [2]; Trust in Automation Scale [3]; Human Computer Trust (HCT) scale [4]; and the SHAPE Automation Trust Index, SATI [5]. The team decided to use the items from SATI, which was developed by EUROCONTROL as part of its SHAPE program to evaluate the workload, trust, and situation awareness impacts of ATC automation.

3.3 Validation Objectives

The table 3.1 below lists the six main Validation Objectives that MAHALO aims to pursue.

Table 3.1 Validation Objectives

ID	Validation Objective	Methods & measures	Validation criterion and testable assertion
VO1	Assess the impact of different combinations of different levels of conformity and transparency on controllers' **acceptance** of the solutions	● Binary accept / reject performance data, per individual advisory	● Controller binary accept / reject rates will vary by conformance and transparency
VO2	Assess the impact of different combinations of different levels of conformity and transparency on controllers' **agreement** with solutions[2]	● Subjective agreement rating (0-100), per individual advisory	● Controller self-reported agreement (0-100) will vary by conformance and transparency
VO3	Assess the impact of different combinations of different levels of conformance and transparency on controllers' **understanding** of the proposed solutions	● Post-session Questionnaire	● Controller self-reported understanding of proposed solutions will vary by conformance and transparency
VO4	Assess the impact of different combinations of different levels of conformity and transparency on controllers' **trust** in the AI	● embedded questions in post-session questionnaire	● Controller self-reported trust in ML proposed solutions will vary by conformance and transparency
V05	Assess the impact of different combinations of different levels of conformance and transparency on **safety**	● Post-session Questionnaire ● Separation level	● Objective (binary loss-of-separation) and subjective (self-report) safety levels will vary by conformance and transparency

[2] Notice that acceptance and agreement are distinct. Acceptance is a binary performance measure (was an advisory accepted or rejected?), whereas agreement is a subjective assessment of the match between controller and advisory strategy. Acceptance and agreement can dissociate.

V06	Assess the impact of different combinations of different levels of conformance and transparency on controllers' **self-reported workload**	• Self-reported workload (0-100), embedded in post-session questionnaire	• Self-reported workload will vary by conformance and transparency

3.4 Reference and Solution Test scenarios

Figure 3.8 depicts SectorX and examples of the supported functionalities. The Reference (i.e. Pre-Test) and solution (i.e., Main Experiment) scenarios will be similar and matched analogues of one another. Basically, scenarios focus on en-route RVSM airspace (roughly FL290-370), with a mix of cruise and climbing/descending flights, and two-aircraft conflict situations. Each scenario begins with a pending conflict - if the controller takes no action, two aircraft will lose separation. Given technical limitations (only heading solutions are possible in the Main Experiment), we structure scenario traffic in such a way that we maximize the likelihood of heading solutions. For example, we include 'blocking traffic' above and below the ownship to limit vertical manoeuvrability. We cannot guarantee in advance that controllers will not use altitude solutions, but will allow the solution type (heading or altitude—speed is unlikely) to be chosen naturally.

On the basis of expertise within the MAHALO network (LFV in Sweden and ANACNA in Italy), specific sector parameters (e.g., sector size, traffic density, closure angles) are being finalized. Part of this finalization for the simulation 2A and 2B reference scenarios depends on output and analysis of the Simulation 1 results performed in October 2021, and iterative tweaking of the reference test scenarios. Final specification of the Sim 2A/2B conflict scenarios is pending feedback by the Advisory Board Workshop, scheduled 28 Oct 2021. The results that will be collected during the Workshop will be used to finalise the scenarios.

The Pre-Test will use short scenarios of approximately two minutes each. Pre-Test scenarios will run at 4X--i.e., four times faster than real-time. Six different conflict scenarios will each be repeated 10 times.

The Main Experiment will use only three of the six scenario types, and we will strive to select scenarios that (a) use heading solutions, and (b) show variability between controllers. Main Experiment scenarios will run at 2X. Each Main Experiment scenario will contain a **designed conflict**, similar to the conflict used in the Pre-Test. This is the conflict for which an advisory will be provided, and for which dependent measures will be collected. In the Main Experiment, we are interested in controllers' responses to presented solutions, not in having controllers generate their own solutions. However, the Main Experiment will permit controllers to reject a given proposed solution, and implement their own.

For details of how these scenarios will be used during the Pre-Test and Main Experiment, see section 3.6.3 on data collection procedures.

3.4.1 Main Experiment - Solution scenarios

In the Main Experiment, the nine experimental cells (or **blocks**) will each include three scenarios, yielding a total of 27 5-minute scenarios.

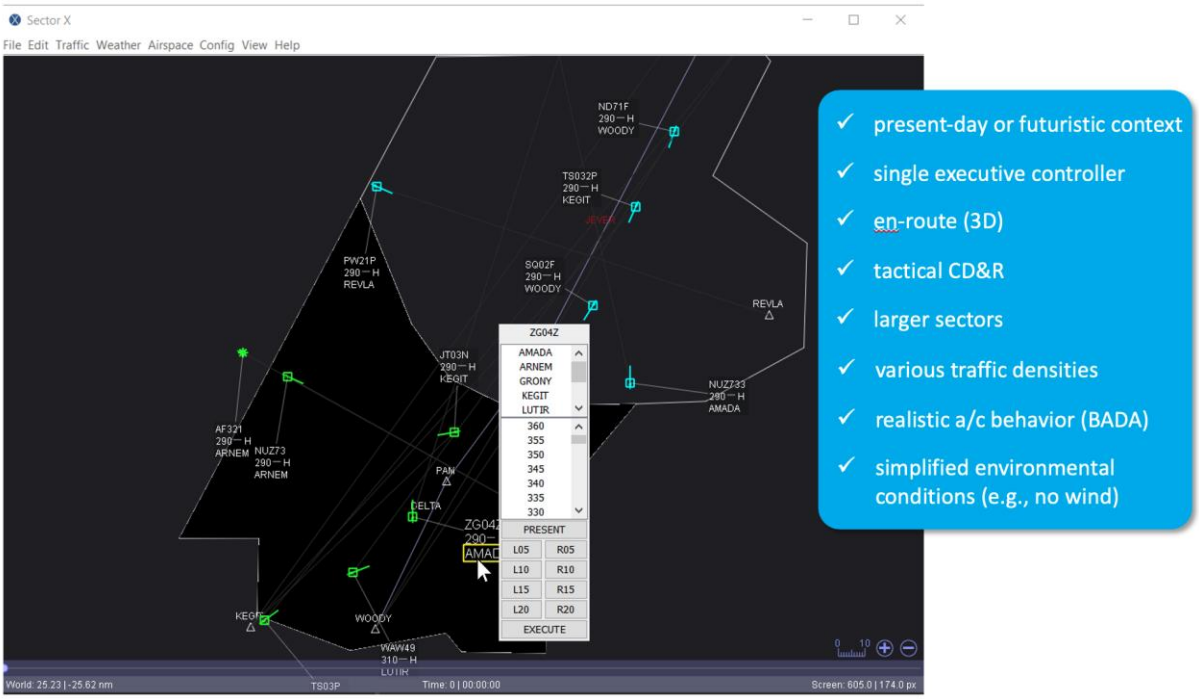

Fig. 3.8. SectorX platform, showing MUAC combined Delta & Jever sectors

3.5 Participants – sampling and assignment

As often happens with simulations that require air traffic controllers, operational realities generally limit how many participants are made available Notice that, according to the Latin square design shown in section 3.2.2, a multiple of nine participants will be required for a balanced design. Controllers will be randomly assigned to an experimental presentation order, and each controller will experience each of the nine experimental blocks only once.

Confirmed participant numbers are as follows:

- Sim 1 (The Netherlands): 4-6 students
- Sim 2A (Italy): 20 ATCOs
- Sim 2B (Sweden): 16 ATCOs

The current available sample size in Sweden (n=16) does not allow for a complete Latin squares design. We have several possible strategies for dealing with this including reducing the number of levels of the conformance independent variable (e.g. from 3 to 2 levels), and would explore options as required. Another important consideration regarding participants is that conformance in the Main Experiment is defined by an individual's own performance in the Pre-Test. This makes it essential that we have access to the same controllers in both phases.

3.6 Data collection methods

3.6.1 Equipment

3.6.1.1 Simulator Platform

Experiments will make use of the SectorX simulation platform, which can be run on a portable computer. Training of ML models will be performed offline, between the **Conformance Pre-Test** and the **Main Experiment** and will be done on a cluster of machines at project partner locations in The Netherlands and/or Sweden.

SectorX is to be run on a PC laptop (Intel Core i5 or better), connected to an external 28” 1920x1080 (or better) resolution display. User interaction will be via keyboard and mouse.

3.6.1.2 Eye Tracking Equipment

An eye tracker from Tobii is likely to be used for gathering gaze data. The wearable Tobii Pro Glasses 2 was used in the Validation simulation (D4.2). The system worked well and collected data with an acceptable accuracy. As configured, the Pro Glasses 2 uses a pupil centre corneal reflection technique, recording eye point of gaze (EPOG) at 50Hz. Small flashes on the glasses frame illuminate the eye with Infrared light to increase pupil and corneal reflections. Small cameras record the reflection (glint) on the cornea and pupil. Algorithms then triangulate EPOG and overlays it as a red circle on the video recorded scene. EPOG data would be analysed using the Tobii Pro Lab software.

A drawback of using a wearable eye tracker is the time and resource demanding mapping of gaze data with a static image representing the scene view. Because a static image must be used to map gaze data, dynamic elements in the interface and simulation are difficult to capture. Data that is not captured well (because of spatial resolution limitations) include aircraft movement, SSD interactions, and dropdown menus from the labels etc.

A promising alternative is the use of a screen based eye tracker, such as the Tobii Pro Fusion or Tobii Pro X3-120. This eye tracker is mounted under the display. This is a suitable alternative given that participants in the MAHALO simulations will work with only one display. Notice that the Pro Fusion and Pro X3-120 models have up to 5x the capture rate (50 vs 250 Hz) of the wearable Tobii system, allowing for richer data.

3.6.2 Materials (instruments and protocols)

The following participant materials are reproduced in Annex A:

- Informed consent (Pre-Test only)
- Post-block questionnaire (Main Experiment only)
- Post-session questionnaire (Main Experiment only)
- Debriefing (Pre-Test)
- Debriefing (Main Experiment)

The participant briefing documents (Pre-Test and Main Experiment versions) are not yet finalised.

In sims 2A and 2B, acceptance will be indicated (in the Main Experiment only) by the controller's objective performance (did they deploy the solution as proposed?), whereas agreement (also Main Experiment only) will be self-reported via an onscreen prompt incorporated in SectorX, using a continuous 1-100 slider scale.

3.6.3 Data collection procedures

Again, the Pre-Test (six conflict scenarios x 10 repetitions x two minutes per repetition) requires 120 minutes (2 hours) of run time. Allowing 30 minutes for familiarization, and a 10 minute break halfway through means that we should be able to finish each session within the allotted three hours.

The Main Experiment (three conflict scenarios x nine blocks x five minutes per repetition) requires 135 minutes (2:15 hours) of run time. Allowing 30 minutes for familiarization, and some additional minutes for agreement ratings, post-block questionnaires, and post-session questionnaire, should allow us to just finish within three hours.

Pacing and stop/start of scenarios will differ in the Pre-Test and Main Experiment. Because we do not intend to collect any data from the Pre-Test other than controllers' chosen solutions, the Pre-Test session will proceed as a single run of successive short (2 minute) scenarios. In the Main Experiment, the 27 scenarios will be somewhat self-paced. In the Main Experiment, simulation will pause at the time a solution is presented. Although this method seems to threaten realism, it also provides some experimental control and technical benefits (e.g., without pausing, a presented solution would only be valid for a short period of time). One benefit of this pause method is that it allows us to measure controller response time, and ultimately make inferences about the transparency manipulation.

In the Main Experiment, controllers would be free to accept or reject the presented solution. If rejected, they could implement their own solution. After completion of each of the 27 scenarios and not onscreen slider scale pops up to collect a 0 to 100 rating on agreement with that solution. The Main Experiment would include two types of questionnaires: post – block questionnaire (nine in all, one for each of the nine experimental blocks), and a single post-session questionnaire. These questionnaires appear in Annex A.

For the experiments we will use the ML hybrid (with possible algorithmic augmentation) approach for providing conformal and transparent advisories. The real time Hybrid ML system would be able to determine and provide conformal and transparent advisories during real time interaction with the system. Developing a well-functioning Hybrid ML system in the MAHALO project is ambitious. The possibility for using a Hybrid ML system in the experiments depends on the maturity of the ML system by the time for experiments, and the feasibility for travelling with this system (to simulations conducted in The Netherlands, Italy, and Sweden). Given the ambitiousness of the ML development effort in MAHALO, for the sake of contingency planning we have identified two fall-back augmentation options: algorithmic augmentation (as described earlier), and *Wizard of Oz* manual manipulation of transparency and conformance.

Training of ML models will be performed offline, i.e. between the Pre-Test and the Main Experiment. Training ML models is data and time consuming. It is expected that the training of the ML models will

take in the order of several hours. Therefore, it is important to allow for adequate time between the Pre-Test and the Main Experiment (i.e., several weeks).

Three types of ML models will be trained, which require different sets of training data:

- Personal prediction models for individual,
- Prediction models for the group, and
- Optimized prediction models

The personalised models use supervised learning approaches to train on the data generated by the ATCOs in the Pre-Test. The amount of data is therefore limited by practical considerations of available manpower and testing time per person. The training data for the personalised prediction models can be augmented with some artificial data, for example by mirroring measured data or by adding noise., but this will have an impact on the conformance level that can be obtained, so care should be taken in doing this. The optimized prediction models will be trained on artificial data, i.e., artificially generated scenarios, which means there is no real limit on the amount of training data that can be generated.

3.7 Data analysis procedures and general approach

Data analysis begins with data handling, including data logging, reduction, and pre-processing. This section sketches the procedures to be used for data handling. Further details can be found in section 4.

The SectorX platform is capable of real-time data logging, and writes time-stamped event codes to XML file. XML files are easily handled as ASCII text inputs to spreadsheet programs (e.g., Excel), commercial statistical analysis software (e.g., SPSS, Matlab, R), and custom data handling and pre-processing routines (e.g., in Python). To verify results, at least two researchers will statistically analyse results independently and in parallel. Example inferential tests include:

- Repeated measures Analysis of Variance (ANOVA) - for interval and ratio data, if normally distributed [6]
- Friedman test – and related non-parametric tests, for data not normally distributed [7]
- Cochran's Q-test – for binary data (e.g. acceptance) [8]

Analysis will aim, in the first instance, for the use of inferential statistical methods and customary decision rules (e.g. $p<.05$). Given the small samples available (a typical challenge in human in the loop simulations involving air traffic controllers) , analysis will as necessary adopt a more lenient probability value, and also explored descriptive statistics and trend level ($p>.05$) results.

Eye-tracking data, when available (see section 3.10.2), will be used as a dependent measure of the conformance and transparency manipulations. In the first instance, workload impacts will be inferred through analysis of pupil diameter, blink rate, visual saccade amplitude, and point of gaze dwell time. Qualitative and descriptive analysis will also be made of potential control strategy differences, as determined by trends in scan patterns.

3.8 Ethical considerations

Ethical considerations are handled separately in Deliverable 8.1 (PODP-Requirement No 4). There will be further reference to ethical considerations in Deliverable 8.2, to be delivered concurrently with the final experimental design.

3.9 Contingency planning and related considerations

3.9.1 COVID 19 contingency planning

It seems, as of this writing, that the ongoing COVID 19 pandemic will force some of the simulation activities to limit face-to-face contact. The team is planning for this contingency by:

(1) Developing a remote simulation capability, to enable participants to (largely) self-administer the simulation session/s, either at a central data collection site or separate (individual) sites;
(2) Developing a capability for realtime link between experimenters and participants, to allow (for example) tech support issues to be resolved quickly; and
(3) Creating a schedule to allow for minimal onsite staff and contact.

The MAHALO team assumes that, at least for Simulation 1, data collection will rely on a mobile platform such as TeamViewer or Zoom, which can allow for remote SectorX simulation.

3.9.2 Eye tracking logistics

The team has been exploring the use of eye tracking measures, both as a potential input to the supervised learning system but also as potential dependent measures suitable for human in the loop simulations. In the former case, as an input to the ML, eyepoint of gaze (EPOG) can serve as an indicator of attentional focus (i.e., where the controller is currently looking). This can be useful for directing the attention and action of the ML system. In the latter case, as a set of dependent measures, eye tracking can consist of for example blink rate, pupil diameter, and saccade amplitude (as measures of mental workload).

Given that the ongoing Covid pandemic is forcing a semi-remote testing protocol for at least simulation 2A (scheduled to take place in Italy Dec 2021-Jan 2022), eye tracking will not be available for this simulation. However, the option remains open for simulation 2B that will be run in Sweden, where logistics make the use of eye tracking equipment more practical.

3.9.3 Conformance vs Optimality

MAHALO defines an optimised solution as one based on RL/algorithmic model output (based on, for example, the *Deep Q-learning from Demonstrations* (DQfD) model's reward function). Notice however that these is not necessarily a single reward function. The function could, for example, prioritise fuel

burn, or time, etc. The exact tuning of the reward function is currently under development, and is based in part on controller expert input.

This experimental design assumes that conformal (i.e., matching the controller's own) solutions and optimized (i.e., per the RL / algorithmic function) solutions are different. This element of the experimental plan relies on controllers' actual performance. It is possible, therefore, that controllers might in fact choose solutions that match those of the ML system. In that case, conformal solutions are optimal solutions. If and when this turns out to be the case, analysis can be adapted. In the first instance, and assuming we see sufficient variation among controllers, analysis can focus on just the subset of controllers whose chosen solution differed from the optimal. This is not the preferred approach, as it will reduce the data set available for analysis, and possibly complicate factorial comparisons across experimental conditions. In the second instance, we can explore the presentation of alternative (but still 'optimised' per the system's reward function) solutions.

3.9.4 Controller performance consistency

Based on previous research within the team, MAHALO expects there to be some variation in controllers' preferred strategies. Notice, in fact, that the experimental design rests on this assumption – if all controllers end up choosing the same solutions, then there is no difference for example between an individually conformal solution and a group conformal solution. This eventuality could be addressed by manual scenario creation for the Main Experiment.

3.9.5 Insufficient training data

ML, in particular SL models, require enormous amounts of processed (labelled) data from which to learn. Collecting enough data from HITL data can be a challenge. We can simply not expect controllers to sit through thousands of conflict scenarios, to collect the data ML might need to learn controllers' strategies. It is possible that the MAHALO ML model will not initially demonstrate sufficient learning performance. If this is the case, we would intend to augment the training dataset via synthetic traffic generation [9]. This would basically involve cloning HITL data by adding stochastic noise (e.g., small random heading variations, or altitude changes) to the HITL seed data.

3.9.6 Matched controller samples across Pre-Test and Main Experiment

Because conformance in the Main Experiment is to find by each controllers own previous performance in the Pre-Test, it is essential that we have access to the same controllers across both phases. Given close organizational ties within the consortium however, we believe that this risk is fairly low. If this does turn out to be a problem, one possible solution might be arranging for a private makeup session. Because Covid has forced us to prepare for a semi-remote data collection, we are prepared to follow this path if necessary.

4 Data format and protection

This section provides a bit more detail on the data format and processing specifications. Some of this material, including postprocessing and statistical analysis methods, have already been covered in the preceding sections 3.1 – 3.6. Further details of the data formats will be presented in the MAHALO D5.1 report: Analysis and Report, Integration.

4.1 Simulation platform data outputs

Again, the SectorX platform writes to XML files in non-volatile memory. The platform writes two XML files, in particular:

- StateLog file: a timestamped encoding of traffic and display parameters, including everything necessary to reconstruct the static traffic geometry at a given time. These parameters include, for instance, aircraft call signs, altitudes and headings, and user interface settings. The platform logged all traffic states once every five seconds. As captured in the D4.2 E-UI Validation Report, an example of the StateLog.XML file would look thus:

```
<state realTime="22.585783" scenarioTime="45.171566" per
formanceScore= "81.54768">
<aircraft callsign="QL10X" radarStatus="active" icao="A3
88" isSelected="true" isAutomated="false" caution="false
" conflict="false" controlled="false" ownNavigation="tru
e" flightState="assumed" label_x="-99.452194" label_y="-
10.452847" x="8.60936"
y="20.421099" altitude="29000.0" heading="174.0" track="
174.0" bankAngle="0.0" acceleration="0.0" gs="387.1662"
tas="387.1662" ias="250.0" rocd="0.0" mach="0.6540285" i
as_min="215.68039" ias_max="340.0" tas_min="336.77472" t
as_max="513.8347" targetAltitude="29000.0" targetHeading
="174.0" targetIas="250.0"/>
```

- EventLog file: a timestamped encoding of all user inputs, including mouse-based interactions with the SSD and general user interface, and also keyboard inputs (e.g., label click and drags, clearances/commands, etc.). An excerpted example of the EventLog file is shown here, in which he flightlevel of aircraft CJ17T is changed to FL310:

```
 <event eventType="flightEvent" flightEventType="flightC
ommand" agent="human" callsign="CJ17T" flightCommandMode
="altitude" altitude="31000.0" scenarioTime="80.66174" r
ealTime="40.33087"/>
```

Together, the StateLog and EventLog files allow reconstruction of the overall user interface. The **EventLog** captures which aircraft received a flight command at what time. The **StateLog** can then be used to retrieve the controlled and observed aircraft states within the radar update that best fits the scenario time at which the flight command was given.The logdata can be used to access required timing information when action is taken to solve the conflict, aircraft choice, resolution type, resolution direction, and directional value. The SSD for a given aircraft can be shown in the replay of the scenario using the SectorX viewer mode. From the replay, the SSD for the aircraft controlled can be captured and saved for export, together with the logdata, for use as input to the training set for the SL models.

4.2 Eye tracker data outputs

As previously discussed in section 3.7.1.2, simulation activities assume the use of a Tobii system (either the Pro Fusion or X3-120) fixed base platform. The eye tracking system writes raw tabular data (time stamped with pupil diameter and X/Y coordinates based on calibration). It also incorporates Tobii's Pro Lab software which, among other things, provides a heat map visualization of visual fixations as shown in figure 4.1.

Eye tracking data is generally output in a tabular format. In its most basic form, information about recording timestamp is provided. For each timestamp the system categorises the eye tracking data as fixation, saccade, or eyes not found (neither fixation nor saccade) according to an algorithm. Different algorithms can be used to define what represents a fixations and saccade. When Areas of interest (AOIs) have been defined, an output can be generated for each timestamp that indicates in which AOI a fixations has been registered. Table 4.1 shows an extracted sample of tabular eye tracking data.

Table 4.1 Sample of tabular eye tracking data

Recording timestamp (milliseconds)	Mapped eye movement type	AOI 1	AOI 2	AOI 3	AOI 4	AOI 5	AOI 6	AOI 7
55510	Fixation	1	0	0	0	0	0	0
55530	EyesNotFoundMovement	0	0	0	0	0	0	0
55550	Fixation	0	1	0	0	0	0	0

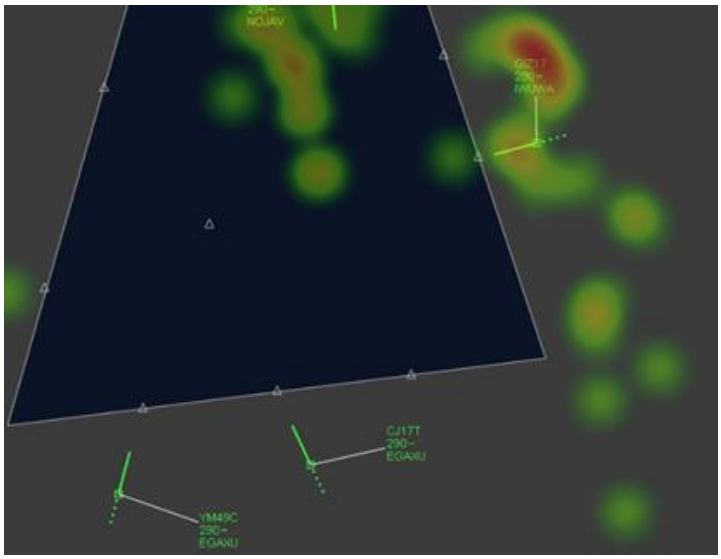

Figure 4.1 Heat map visualisation of visual fixations.

In addition to Tobii's proprietary data analysis software, the MAHALO team is exploring the use of the open source PyTrack software [10] for eye tracking analysis.

4.3 User behavioural outputs

SectorX Eventlog data will be pre-processed using bespoke Python routines, to extract time stamped controller actions of interest. This will lead to a number of derived performance measures, including the following:

- acceptance/rejection of a given advisory, which will be aggregated into acceptance rates
- response time to advisory onsets (notice that there is some subtlety involved in analysing response time, because lack of a response can indicate either no recognition response, or it can indicate a decision to wait and see, for example)
- individual clearances, which will be aggregated into interactions per aircraft averages
- chosen manoeuvre (e.g. heading, altitude) and value (e.g., degrees of heading change)

4.4 Data protection

As outlined in deliverable 8.1 PODP requirements No 4, MAHALO has procedures in place for complying with legislation and guidelines on the ethical protection of participant personal data.

These procedures embrace the rights of data subjects (e.g. the right to information, access, erasure, and rectification). These procedures also embrace such security measures as subject

pseudonymisation, data storage network disconnection, encryption, physical security, as well as planning for data destruction and personal data breach contingencies. Again, details of the intended data protection procedures can be found in D8.1.

5 Research Coordination and Development

Scientific research is built on the principle of replication. Scientific reporting should provide the interested reader sufficient information on the methods, equipment, and procedures, that the experiment or simulation itself can be reproduced. To this end, the MAHALO consortium is putting in place a data management plan for storing, securing, and accessing data over the project lifecycle. As discussed in other project deliverables, the MAHALO consortium is disseminating results via scientific publications and conference presentations. These dissemination activities, along with project deliverables, specified the project research methods, equipment, and procedures, insufficient detail for experimental replication by outside parties.

Having said that, aspects of the simulation platform remain proprietary. For example, although operation and dynamic demonstration of the simulation platform are planned (and indeed already available through project website and other dissemination avenues), some specific aspects of both the simulation platform and ML algorithms shall remain proprietary. In practice, this means for example that source code for the simulation platform shall not be made publicly available.

REFERENCES

[1] 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.

[2] Jian, J.Y., Bisantz, A.M. & Drury, C.G. (2000). Foundations for an empirically determined scale of trust in automated systems. International Journal of Cognitive Ergonomics, 4(1), 53-71.

[3] Merritt, S.M. (2011). Affective processes in human-automation interactions, *Human Factors*, 53(4), 356-370.

[4] Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. Proceedings of the 11th Australasian Conference on Information Systems. 6-8 December. Australasian Association for Information Systems. QUT, Brisbane, pp. 6-8.

[5] Dehn, D.M. [2008]. Assessing the Impact of Automation on the Air Traffic Controller: The SHAPE Questionnaires; Air Traffic Control Quarterly, Vol. 16(2) 127-146 (2008).

[6] Minke, A. (1997). Conducting Repeated Measures Analyses: Experimental Design Considerations.

[7] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association. **32** (200): 675–701.

[8] Cochran, W.G. (1950). The comparison of percentages in matched samples. Biometrika. 37 (3/4): 256–266.

[9] Rooijen, S.J. van, Ellerbroek, J., Borst, C. & van Kampen, E.J. (2019). Conformal Automation for Air Traffic Control using Convolutional Neural Networks. In Proceedings of the ATM Seminar, Vienna. June, 219.

[10] Ghose, U., et al. (2020). PyTrack: An end-to-end analysis toolkit for eye tracking. Behavioural Research, 52, 2588–2603. https://doi.org/10.3758/s13428-020-01392-6

[11] Kelly, C. et al (2003). Guidelines for Trust in Future ATM Systems: A Literature Review. EUROCONTROL EATMP document HRS/HSP-005-GUI-01.

[12] Cramer, H. et al. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. User Model User-Adaptive Interfaces, 18:455–496.

[13] Wang, N., Pyandath, D.V. & Hill, S.G. (2016). Trust calibration with a human-robot team: Comparing automatically generated explanations. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 109-116, doi: 10.1109/HRI.2016.7451741.

ACRONYMS

AI	Artificial Intelligence
ANOVA	Analysis of Variance
ATC	Air Traffic Control
ATCO	Air Traffic Controller
ATM	Air Traffic Management
CD&R	Conflict Detection and Resolution
DQfD	Deep Q-learning from Demonstrations
E-UI	Ecological User Interface
EID	Ecological Interface Design
FL	Flight Level
GS	Groundspeed
IAS	Indicated Airspeed
MAHALO	Modernising ATM via Human-Automation Learning Optimisation
ML	Machine Learning
MUFASA	Multidimensional Framework for Advanced SESAR Automation
RL	Reinforcement Learning
SATI	SHAPE Automation Trust Index
SL	Supervised Learning
SSD	Solution Space Diagram
SSR	Secondary Surveillance Radar
UI	User Interface

Annex A. Participant materials

This annex contains the following five documents, to be used by participants. Briefing documents for the Pre-Test and Main Experiment are still under development, and are not reproduced here.

- A1 Informed consent (Pre-Test only)
- A2 Post-block questionnaire (Main Experiment only)
- A3 Post-session questionnaire (Main Experiment only)
- A4 Debriefing (Pre-Test)
- A5 Debriefing (Main Experiment)

A1 Informed consent (Pre-Test only)

We have asked you to help us with a European Union funded research project called MAHALO. This project requires us to run two small simulations, one today and one a few weeks from now. You are being asked to participate in both of these simulations.

The aim of our research is to evaluate controller interaction with prototype air traffic control systems. Your participation today, and in the follow-up phase, is voluntary, and no compensation is offered.

Both today's session, and the follow-up phase in a few weeks, require a total of about three hours per day. In this first session, you will be controlling a number of short air traffic control scenarios. In the second session, in a few weeks, you will be doing the same, however you will be using a slightly redesigned interface, and we will be collecting some survey and other data from you during the follow-up session.

There are no known risks or discomforts associated with your participation in this simulation. You have the right to stop the simulation any time you wish. Finally, any data collected today or in the follow-up phase will be kept strictly confidential. You will never be identified in any way.

If you have any questions at all about your participation in this study, please feel free to ask one of the experimenters today.

By signing this Consent Form you confirm that you have read and understood its content, and you agree to voluntarily participate in this simulation. You may request a copy of this form.

Thank you for being a part of this simulation.

Date: _____

Your Name: _____

Your Signature: _____

A2 Post-block questionnaire (Main Experiment only)

[1] Workload: Please indicate the workload you felt, while doing the previous 3 scenario block. Indicate on scale of 0-100.

Please indicate how much you agree with the following statements, on a scale of 0-100. You can either draw a mark along the line, or by writing a number between 0 (total disagreement) and 100 (total agreement). Remember that we are only asking about the automated advisory tool (which popped up proposed solutions), not other aspects such as the traffic scenarios, simulation, interface, etc.

[2] The automation was useful.

[3] The automation was reliable.

[4] The automation worked accurately.

[5] I liked using the automation.

[6] The automation was easy to use.

[7] I could understand how the automation worked

[8] I know what will happen the next time I use the system because I understand how it behaves.

[9] Although I may not know exactly how the system works, I know how to use it to make decisions about the problem.

[10] The automation always proposed a safe solution.

[11] I sometimes accepted proposed solutions, even though I didn't agree with them

[12] I trusted the automation.

A3 Post-session questionnaire (Main Experiment only)

[1] Workload: Please indicate your general preference for using automation in your ATC job. Draw a mark anywhere along the line.

For the following items, please indicate, on a scale of 1-5, how much you agree with the statement.

[2] The traffic scenarios were realistic.

[3] The interface was similar to what I use on my job.

[4] The session was boring and too long.

[5] I recognised some of the traffic scenarios today, I saw them in the previous session.

A4 Debriefing (Pre-Test)

Thank you for your participation today. As you know, this was the first of two sessions. Your follow-up session is scheduled in a few weeks. Are you aware of the date and time, and can you already confirm that you'll be available then?

Because the two sessions of this experiment are dependent on each other, we can not provide much specific information yet on details of the study. However, we will provide you a thorough debriefing after your second session. What we can already session is that our aim is to examine how controllers interact with new prototype forms of automation. The scenarios that we ran today were obviously quite simplified. For example we had no wind effects. Also, traffic ran in short bursts, and was accelerated to four times normal speed.

Again, we can not provide much specific information today, we prefer to wait until after the second session, and provide a thorough debriefing then. However, if you have any specific questions, please ask them now.

Thank you again for your participation today. We look forward to seeing you again in a few weeks.

A5 Debriefing (Main Experiment)

Thank you for your participation today. We would like to provide you with some details about the research you have been helping us with.

Our research project is called MAHALO, which stands for Modern ATM via Human-Automation Learning Optimisation. Like so many other fields these days, ATM is looking at the potential for artificial intelligence to help with some aspects of the job. But before we start building artificial intelligence for air traffic control, we first wanted to ask some very basic questions. For example, does automation have to be designed in a way that the human can understand it? Does it have to solve problems the same way that a controller would? The task we focused on was conflict resolution in en route airspace.

Our experiment was designed to investigate two factors: CONFORMANCE and TRANSPARENCY. Conformance is how well the automation matches your own way of working. In other words, does it seem to think the same way that you do? Notice how we investigated this: during the previous session, we recorded your preferred solutions to conflicts. Between that first session and today, we actually took your solutions and use them to train and artificial intelligence automation program. In other words, we taught automation to solve airspace conflicts the same way you did. Today, during the second session, we then presented you with a number of scenarios. For some of these, solutions were provided by a system trained on your personal previous solutions. For other scenarios, solutions were provided by a system that was trained on other people's (different but still workable) solutions.

Transparency is just how understandable and explainable the automation is. We varied this today by giving you different types of explanations. Sometimes we gave you explanations about the traffic, and other times we gave you more information about specific considerations.

Our aim now in data analysis is to examine how these two factors that we varied actually impacted such things as controllers' workload, acceptance of solutions, etc.

We know that certain aspects of the simulation were unrealistic. For example, we did not simulate wind effects, nor encourage controllers to use altitude solutions. As you might have noticed, we placed blocking aircraft above and below the target aircraft, to try to discourage altitude solutions. You probably also notice that we ran the simulation at greater than real-time speed. And that we froze the display whenever conflict advisories were presented. We hope that these aspects of the simulation were not too troublesome.

At this point, we must ask you a favour. As you now see, it is critical that controllers who are helping out with the simulation should not know too much about our goals and methods, otherwise, they might bias our results. For this reason, we ask you to not discuss the simulation with other colleagues. At least until the end of this week, when we will have finished our data collection.

Thank you again for your help with the simulation. Do you have any questions for me?