EXPLORATORY RESEARCH

Evaluation Results

D3.2

MINIMA

Grant:	699282
Call:	H2020-SESAR-2015-1
Topic:	Automation in ATM
Consortium coordinator:	DLR
Edition date:	21 March 2018
Edition:	00.02.00

Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
Bruno Berberian BB/ONERA	WP1 Lead	2018-03-09
Jan Kraemer JK/DLR	Project Coordinator	2018-03-20
Oliver Ohneiser OO/DLR	WP2 Lead	2018-03-18
Peter Jänsch PJ/DLR	Project Associate	2018-02-19
Gianluca Di Flumeri GDF/BS	Third Party	2018-03-14

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
Francesca De Crescenzio FC/UNIBO	WP3 Lead	2018-03-20
Marcus Biella MB/DLR	Reviewer	2018-03-19
Jürgen Rataj JR/DLR	Reviewer	2018-03-18

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

Name/Beneficiary	Position/Title	Date
Jan Kraemer JK/DLR	Project Coordinator	2018-03-21
Oliver Ohneiser OO/DLR	WP2 Lead	2018-03-20
Francesca De Crescenzio FC/UNIBO	WP3 Lead	2018-03-21
Bruno Berberian BB/ONERA	WP1 Lead	2018-03-21
Gianluca Di Flumeri GDF/BS	Third Party	2018-03-21

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date

Document History

Edition	Date	Status	Author	Justification
00.00.01	2018-02-09	draft	BB	Initial structure proposition
00.00.02	2018-02-14	draft	JK	Data Analysis Input
00.00.03	2018-02-15	draft	OO	Complete Review and some Inputs
00.00.04	2018-02-22	draft	JK	Added section 3.5 "Technical issues and data exclusion"
00.00.05	2018-02-27	draft	OO	Second Review
00.00.06	2018-02-27	draft	JK	Completed Discussion and Conclusion
00.00.07	2018-02-28	draft	GDF	- Added input from BrainSigns - Added different paragraphs in Chapter 3 and 4
00.01.00	2018-03-02	final	JK	Final review and approval of all partners
00.01.01	2018-13-03	draft	BB	Revision of the document regarding SJU inputs
00.01.02	2018-03-14	draft	GDF	Revision of the document with respect to the SJU feedbacks
00.01.03	2018-03-19	draft	BB	Second revision
00.01.04	2018-03-20	draft	FC	Revision of the document regarding SJU inputs
00.02.00	2018-03-21	final v2	JK	Revision of the document with respect to the SJU feedback

MINIMA

MITIGATING NEGATIVE IMPACTS OF MONITORING HIGH LEVELS OF AUTOMATION

This deliverable is part of a project that has received funding from the SESAR Joint Undertaking under grant agreement No 699282[1] under European Union's Horizon 2020 research and innovation programme.

Abstract

An increase of automation in air traffic control can have negative effects on the air traffic controller's performance. The effects are known as out-of-the-loop (OOTL) phenomenon. The MINIMA Project developed a *Vigilance and Attention Controller* to mitigate these effects. A highly automated arrival management task served as a case study. Psychophysiological measurements like EEG were used to identify the state of the Air Traffic Controller and combined with adaptive task activation. This allowed for activating tasks based on the Air Traffic Controllers state to keep their performance on a high level and to ensure safe operations.

This report on Evaluation Results gives evidence on the MINIMA objectives with respect to the validation trials conducted from 06 – 17 November 2017 at the Virtual Reality Lab of the University of Bologna's premises in Forlì, Italy.

Table of Contents

List of Figures

List of Tables

List of Acronyms

Abbreviation	Description
AMAN	Arrival Manager
ANOVA	Analysis of Variance
ANSP	Air Navigation Service Provider
ATC	Air Traffic Control
ATCo	Air Traffic Controller
ATM	Air Traffic Management
BL	BASELINE
BS	BrainSigns
DLR	German Aerospace Center (Deutsches Zentrum für Luft- und Raumfahrt)
DSSQ	Dundee Stress State Questionnaire
EEG	Electroencephalography
ENAV	Italian ANSP (Società Nazionale per l'Assistenza al Volo)
FMS	Flight Management Systems
IAF	Individual Alpha Frequency
ID	Identifier
Kurt	Kurtosis
M	Mean
Max	Maximum (value)
Min	Minimum (value)
MINIMA	Mitigating Negative Impacts of monitoring high levels of automation
Mn	Median
ONERA	French Aerospace Lab (Office National d'Etudes et de Recherches Aérospatiales)
OOTL	Out-Of-The-Loop
ROCD	Rate of Climb/Descent
SD	Standard Deviation
SDD	Situation Data Display
Skew	Skewness
SO	SOLUTION

Abbreviation	Description
TE	Task Environment
TLX	Task Load Index
TMA	Terminal Manoeuvring Area
TRI	Task Related Interference
TRT	Task Related Thought
TTFF	Time-to-First-Fixation
TUT	Task Unrelated Thought
UNIBO	University of Bologna
VAC	Vigilance and Attention Controller

1 Executive Summary

1.1 Problem Area

Over the past few years, the global air traffic growth has exhibited a fairly stable positive trend, even through economic immobility, financial crisis, and increased security concerns. It is now clear that traffic flow patterns will become more complex, making conflicts and situations harder to identify for a human operator and will put immense pressure on the air traffic control system. In this context, several solutions have been proposed for modernizing air traffic control to meet the demands for enhanced capacity, efficiency, and safety. These different solutions rely on higher levels of automation as supported by both SESAR JU and HALA! Research Network.

On the one hand, implementing higher levels of automation can improve the efficiency and capacity of a system. On the other hand, it can also have negative effects on the performance of human operators, a set of difficulties called the Out-Of-The-Loop phenomenon (OOTL). In the current context of a continued increase in automation, understanding the sources of difficulties in the interaction with automation and finding solutions to compensate such difficulties are crucial issues for both system designer and human factor society.

While this OOTL phenomenon is considered as a serious issue in the human factors literature, it remains difficult to characterize and quantify. Detecting the occurrence of this phenomenon, or even better detecting the dynamics toward this degraded state, is an important issue in order to develop tools for evaluation and monitoring.

The general objective of MINIMA project was to improve our comprehension of the OOTL performance problem especially according to a future air traffic scenario. Further, MINIMA developed tools to detect and compensate the negative impact of this phenomenon and a carefully selected allocation of tasks between the human agent and the automated system for the use case of a highly automated Terminal Manoeuvring Area (TMA).

1.2 Description of Work

In this deliverable, the results of the *Evaluation Study* (T3.2) of MINIMA are reported. All results are based on the data recorded throughout the experiments conducted in accordance with the procedure documented in D3.1 (*Evaluation Plan*[2]). This deliverable concludes WP3 (Evaluation) of MINIMA.

In MINIMA, we aimed to compensate the negative impact of automation on human performance with a specific focus on the vigilance decrement observed in the OOTL phenomenon. In that sense, we proposed a tool which aims (1) to measure the current attention level and the attention focus of the human operator with the aim to detect or anticipate typical OOTL performance issues and (2) to adapt automation in case of vigilance decrement with the aim to compensate it. This tool was the Vigilance and Attention Controller[3] (*VAC*). A study to evaluate the MINIMA concept and to develop suggestions for further improvements has been used.

To ease comprehension of the results reported within this document, first the MINIMA concept is recapitulated. Afterwards, the conduction of the *Evaluation Study* is reported. Furthermore, the dependent variables are reported in detail. Along with each variable, hypotheses about assumed differences between scenarios and levels of vigilance are given. Finally, the results of all data analyses are reported in detail and afterwards discussed in a designated section. The document is closed by the conclusion section.

2 Introduction to MINIMA Concept

Increasing the level of automation in ATM is seen as a measure to increase the performance of ATM to satisfy the predicted future demand. This is expected to result in new roles for the human operator. Human operators will often work in a supervisory or control mode rather than in a direct operating mode. Operators will mainly monitor highly automated systems and intervene seldom. It can be expected that human operators in such a role will be affected by human performance issues like lack of attention, loss of situational awareness and de-skilling known as out-of-the-loop phenomenon. These problems are observed in other domains like flight-crew performance in the glass cockpit.

MINIMA addressed these performance issues. Its aim was to identify out-of-the-loop behaviour and to find solutions to minimize the negative impact of monitoring high levels of automation on the human operator's performance.

In this sense, MINIMA developed a dynamic adaptation of the task environment which is foreseen as a major requirement to keep the human 'in the loop', perfectly aware of the traffic situation. As a consequence of the developed concept, not all tasks potentially automated were automated every time. To trigger adaptations of the automation, MINIMA developed a real-time monitoring system that constantly measures the operators' vigilance and attention levels. This is called "Vigilance and Attention Observer" in MINMA. A component called "Adaptive Task and Support Activation", based on the measured vigilance and attention level, decided which adaptations of the task environment should be activated. An Overview of the MINIMA Concept is shown in Figure 1.

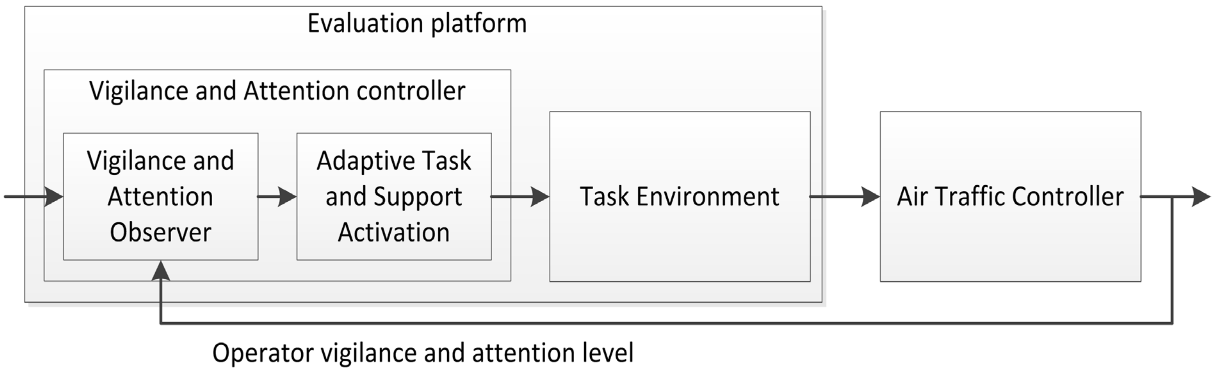

Figure 1. Evaluation environment for MINIMA and influences of components on others

In MINIMA, a highly automated Terminal Manoeuvring Area (TMA) has been selected as use case. This task environment represents an air traffic control task as it is expected for the future: Most of the interaction with the aircraft is automated. A principle assumption of MINIMA is that Air Traffic

Controllers (ATCos) are required to intervene in a few situations as error-free automation cannot be guaranteed.

In the MINIMA use case, the arrival management was highly automated. On-board Flight Management Systems (FMS) negotiated with an Arrival Manager (AMAN) on 4D-Trajectories automatically. However, these trajectories were only guaranteed to be conflict free at a merging point. Conflicts between arrival aircraft at other locations, conflicts between arrival and departures, and deviations form 4D-Trajectoreis were still possible - but seldom - and needed to be managed by the ATCos.

Several adaptations mechanisms were integrated into the use case. An adaptation mechanism changed the tasks that ATCos have to perform during operation, either by providing additional or by handing over task to automation temporarily. These mechanisms included different methods to guide the ATCo's attention and different tasks that can be activated dynamically during the simulation.

3 Method

In this chapter, a detailed description of the protocol used throughout the *Evaluation Study* is given. It follows the *Evaluation Plan* (D3.1[2]), but also includes details about data handling, deviations from D3.1 and reports about technical issues and how they were handled.

MINIMA's core intention is to prevent human operators from getting Out-of-the-Loop (OOTL) in highly automated task environments. Because vigilance decrement has been identified as the first concern and precursor of Out-Of-The-Loop episodes, our MINIMA concept was designed to provide the means to dynamically assign manual and automated tasks between the operator and the system when vigilance decrement was detected. The ATCo's vigilance measurement has been performed through neurophysiological signals, in particular his/her brain activity recorded by Electroencephalographic (EEG) technique and in a less extent ocular movements. Through these measures, the MINIMA concept shall keep operator vigilance at a reasonable and proper level at all times to ensure that they could intervene with the task in case a failure of the automation system made it necessary.

To evaluate the MINIMA tools, we aim to explore how vigilance decrement appears and how our tools can detect and compensate such vigilance decrement. It is noted that the main goal of the MINIMA concept is not to induce OOTL occurrences, but to anticipate them based on objective neurophysiological markers, namely vigilance measured through EEG. Therefore, what the *Evaluation Study*'s aims to show is that the MINIMA tools could detect vigilance decrement and compensate for it in order to prevent OOTL episodes.

3.1 Participants

Between November 6th and November 17th 2017, 15 voluntary subjects, ATCo from ENAV, participated in the study (14 male; average age of 45.0 years, SD = 7.5). The experiment took place at the Virtual Reality Lab of the University of Bologna, in Forlì (Italy). All participants were naive to the purposes of the study. Participants have been informed about the study's purpose after the experiment.

3.2 Procedure

3.2.1 Task

Participants were seated in a comfortable armchair with an appropriate height (see Figure 2). A simulated ATC task was presented in front of the subject by a 27-inch computer screen. The distance from the screen to the plane of the subject's eyes was 60cm.

Figure 2. General set-up used for the MINIMA validation trials.

Participants had to perform an air traffic control task. A highly automated Terminal Manoeuvring Area (TMA) had been selected as use case (see Figure 2). The subject was instructed to monitor arriving and departing traffic and to intervene only in cases of conflicts or emergencies.

3.2.2 Scenario

Three different scenarios were designed to evaluate the MINIMA concept: *BASELINE*, *MINIMA SOLUTION and TRAINING*. In each of them, approach and departure air traffic inside a TMA was simulated. Airspace was quite dense as in expected future scenarios automation will optimize traffic flow during normal operations. Each scenario lasted 45 minutes. Traffic was comparable between scenarios. There was a rate of roughly 30 arrivals per hour and runway as well as 15 departures per hour and runway. Scenario did not contain overflights. In total, about 100 aircraft were present in each scenario. The weight category of aircraft consisted of 10% "Heavy" and 90% "Medium". Typical callsigns of Munich airport were used, but changed between different scenarios to avoid learning effects. The starting points of aircraft were outside the TMA. These points were moved to different positions almost semi-circle-wise rotating around the runways due to airspace structure (again to avoid learning effects of study participants). Few conflicts were introduced in each scenario as MINIMA assumes a well-functioning automation with only very few necessary controller actions. Accordingly, participants were instructed to actively intervene only in case of danger of separation losses between aircraft.

3.2.2.1 Baseline Scenario

The *BASELINE* scenario served as a reference scenario for how air traffic control will be done more than 20 years in the future. While most of the work is left to a highly automated system, the human operator's role was reduced to that of a mere operator during this scenario. Only high level automation of our system (Level 2, see D2.2 *Task Environment* for detail [4]) was engaged in this scenario.

3.2.2.2 Solution Scenario

In the *MINIMA SOLUTION* scenario, the *VAC* (see D2.1 *Integrated Vigilance and Attention Controller* for details [3]) developed for MINIMA actively adapted the level of automation within the task environment, based on the subject's vigilance as measured via EEG data. An automated software-based vigilance check was performed every five minutes. When subjects showed mainly low levels of vigilance (Level 0) during the last time-frame and automation was high, the level of automation was lowered and vice versa. Different levels of automation were realised through different automation and attention guidance systems featuring different operational modes (see chapter 4 of D2.2 *Task Environment* [4]). Depending on the level of automation, controllers were either reallocated part of their operative tasks, i. e. manually manage traffic, or were provided with additional information such as unmonitored aircraft and potential separation losses. If controllers showed high levels of vigilance (Level 2), automation was set back to a higher level.

3.2.2.3 Training Scenario

The *TRAINING* scenario was used to introduce the subject controllers to the MINIMA concept. It aimed (1) to familiarise participants with the *VAC*, and (2) to cause subjects to trust the system and therefore increase their willing of using it during their work. During the TRAINING scenario, the automation level was manually altered by the experimenters. This served to provide subjects with a standardised introduction to the task environment in both low and high automation mode. TRAINING consisted of 15 minutes of high level automation, 15 minutes of low level automation and finally another 15 minutes of high level automation.

3.2.3 Sequence of events

The experimental protocol was developed along two sessions. In the first session, subjects were introduced with the experimental task and the automation systems. The second session was the actual experimental session during which subjects' performance with and without the MINIMA SOLUTION was compared.

3.2.3.1 Training Session

First, subjects received a briefing about the experimental procedure and signed an informed consent form. Afterwards, they received a thorough introduction to familiarise them with the task environment. This included a 'background-story' in which the future high-level automation scenario for which the MINIMA concept was developed was described. Furthermore, subjects were presented with the RadarVision display [5]. All instructions were given in Italian language and read aloud from a pre-printed sheet by the experimenter.

Following the briefing and familiarisation, subjects performed the 45 minutes TRAINING scenario during which automation levels were manually altered by the experimenters. Contrary to the planned sequence of events as reported in section 3.3 of D3.1 (*Evaluation Plan*[2]), only one TRAINING scenario with alternating automation levels was performed instead of 30 minutes of TRAINING with either one of both automation levels. It was decided that this direct comparison of both automation levels would improve the familiarisation process. Also, it was deemed unnecessary to complete the very same 45 minute scenario twice.

The training session was closed by a debriefing during which subjects were given the opportunity to give feedback on the procedure and ask any open questions.

3.2.3.2 Experimental Session

The experimental session was started with another briefing to allow for clarification of open questions. Afterwards, the EEG and eye-tracking systems were installed and calibrated.

Reference data of for the Vigilance Controller was gathered to calculate individual parameters as well as the thresholds for low and high vigilance for each subject. Originally, it was planned to use two short EEG reference scenarios, each designed to induce a constant level of either low or high vigilance (see section 3.2.2.4 of D3.1 Evaluation Plan for details). However, during the Calibration experiments (see section 3.4.1.1.1 for further details) it was found that these scenarios were not sufficient to reliably distinguish between low and high vigilance. Thus, they did not provide the necessary data to serve as a vigilance reference during BASELINE and SOLUTION scenarios. Also, from a scientific point of view, it is demonstrated that long monotonous tasks use to induce vigilance decreasing [6], with significant effects after 10 minutes [7]. As a countermeasure, a short 15-minutes scenario with the highest level of automation was used. Essentially, as for the BASELINE scenario, the ATCo had just to monitor the traffic. However, subjects were required to answer two standardised questions about their task situation in two different moments. Actually, the two questions were provided within the first five minutes, in order to induce “Relax” for the Scenario remaining period. This was supposed to induce a high level of vigilance and therefore serve as the ‘high level vigilance’ reference (see Table 1). Also during Calibration experiments, it was shown that data gathered from this method was effectively suited to distinguish between low and high vigilance levels (please see section for 3.4.1.1.1.2 further details).

Table 1. Control questions asked during the EEG reference scenario.

#	Time	Question
1	3 minutes	How many aircraft are under your control?
2	5 minutes	What is the altitude of the lower aircraft that is approaching the runway?

Following the EEG reference scenario, subjects completed the BASELINE and SOLUTION scenario. The order in which both scenarios were completed was randomised between subjects to control for sequence effects. EEG and eye-tracking data were gathered during both scenarios. After each scenario, subjects completed two electronic questionnaires. The first questionnaire was an adapted version of the Dundee Stress State Questionnaire (DSSQ [10]). The second questionnaire was the NASA Task Load Index (NASA-TLX [12]).

The experimental session was closed by a debriefing during which subjects were told about the experiment's actual purpose. Additionally, again subjects were given the opportunity to give feedback on the experiment and the MINIMA concept and to ask any open questions.

3.3 Material

All of the traffic scenarios were supported by an air traffic simulator that was responsible for proceeding radar tracks of each aircraft. This simulator also provided the aircraft behaviour triggered by automatically executed controller commands in all simulations. Those controller commands (e.g., DESCEND, REDUCE) were calculated by an Arrival Manager (AMAN[13]) and sent to the simulator on time. Nevertheless, the controller was still able to insert additional commands for each aircraft via the mouse interface of the radar display. Departure aircraft radar tracks were also generated by an air traffic simulator without following AMAN trajectory calculation respectively automatic commands. All scenarios ensured that they are almost free of conflicts except of those conflicts the controller should detect in very seldom cases.

3.3.1 The Arrival Manager (AMAN)

All trajectory planning was done by a software-based AMAN. The AMAN software consists of several modules: A lateral path predictor, an arrival interval calculator and a scheduler. In combination, these modules are capable of calculating arrival sequences for aircraft within a specified TMA. Aircraft movement was processed through a dedicated air traffic simulator for flight movements.

3.3.2 Radar display: RadarVision

Visualisation of radar data calculated within the simulation software was done via the RadarVision display (see D3.1 *Evaluation Plan* [2]). RadarVision visualizes static airspace dependent data as well as calculation results from the AMAN. The central view consisted of the Situation Data Display (SDD) that displays runways, TMA borders, routes, points, and aircraft. By using the "mouse over"-functionality on an aircraft icon corresponding data like the planned 4D-trajectory or weight category could be visualized in an extended label (for more details see D2.2 *Task Environment* [4]). A timeline was shown right of the SDD. Each aircraft had a label dedicated to a certain time and runway. All dynamic elements moved downwards as time went on.

RadarVision also served as a human machine interface as it allowed the controller to give clearances to the aircraft displayed within the TMA. In RadarVision, mouse control interfaces were used to give commands to the aircraft within the controller's area of responsibility. For instance, altitude clearances were given as follows: The controller selected the altitude label of aircraft DLH123 via mouse click. Then, a drop-down menu unfolds from which the desired altitude value was selected and confirmed by clicking a button labelled OK. The confirmed command was forwarded to a simulator control module of the AMAN from where it was sent to an air traffic simulator for flight movements.

3.3.3 Instruments of measurements

3.3.3.1 Electroencephalogram (EEG) recorder

The EEG device used in MINIMA Evaluation Experiments has been the g.USBamp (Guger Technologies GmbH, Austria), a wired EEG system able to record up to 16 EEG channels with a sampling frequency up to 1024 Hz. In MINIMA, the sample frequency was set on 256Hz. According to the conclusions of the preliminary Calibration Experiments (see section 3.4.1.1.1.3 for further information), 15 wet electrodes were used during these Evaluation Experiments in order to reduce the system intrusiveness. The electrodes were placed mainly on the prefrontal, frontal, and centro-parietal sites, in particular: Fpz (it has been used only for ocular artifacts rejection), AF3, AF4, AF7 AF8, Fz, F3, F4, F7, F8, CP3, CP4, Pz, P3 and P4, according to the 10-20 International System. A pair of electrodes on the earlobes has been used as reference, while the system ground has been placed on the left mastoid. The impedance of all electrodes was kept below 10 kΩ.

To perform online classification of the ATCo mental state, we used software developed by BrainSigns srl (*BS Recorder*), which allowed recording, processing and visualizing online EEG signals. Moreover, the computation and online classification of neuro-indexes of the investigated mental state and its dispatching (i.e. the online index) through a specific network protocol were also implemented. In the framework of MINIMA, it had been implemented in order to capture the vigilance level by means of an EEG based vigilance index. The different steps of the signal processing (pre-processing, Feature extraction and pattern classification) were detailed in chapter 3 of D2.1 (Integrated *Vigilance and Attention Controller*[3]). Therefore, the online EEG-based Vigilance index of the ATCo has been used to trigger the automation adaptation. The different adaptation solutions proposed were described in chapter 4 of D2.2 (*Task Environment*).

3.3.3.2 Eye-tracking recorder

Gazing behaviour was recorded using a Tobii Eye-Tracking System EyeX (https://tobiigaming.com/product/tobii-eyex/). The Tobii EyeX Controller uses near-infrared light to track the eye movements, the fixations and gaze point of a user. The device provides data at a time resolution of 60 Hz and can capture the human gaze pointing at a screen point up to a dimension of the screen of 27”. This eye-tracking system was set on the desk in front of the subject, between the subject and the screen.

Pre-processing of eye-tracking data recorded by the Tobii EyeX Controller was implemented into the RadarVision software. Fixations are detected when the captured gaze points are located within an area of around 0.2 percent of the screen for at least 20 ms. For each fixation, the software automatically recorded the relative x and y on-screen position, the type of object (aircraft/route point) looked at and its ID. Additionally, timeticks of start and end of each fixation were recorded. All data was written to the data base in a separate table. Object type and ID allowed for a definite assignment to all other data of the respective object such as its absolute position in airspace at the time of each fixation. It has to be noted that because the data base’s timeticks were based on seconds, it was not possible to further distinguish fixation durations below one second. However, it was possible to record multiple fixations occurring within one second and save them to the data base without loss of information.

3.4 Measures

In this section, the dependent variables used to quantitatively examine the hypotheses are reported along with their ways of determination. Particularly, they should serve to quantify differences in vigilance, attention and performance of subjects between BASELINE and SOLUTION scenarios. Measures were subdivided into two different groups: main results and additional results.

The main measures aim to prove the capability of the MINIMA's tool to detect and mitigate vigilance decrement in order to prevent OOTL phenomenon. It is the core of our analysis. They correspond to biomarkers of vigilance and attention which include both EEG and oculometric data.

Together with these main measures, additional measures have been added to our analysis. These additional measures include subjective and performance measure. They aim to go further in our comprehension of the impact of the MINIMA tools on human operators.

It has to be noted that the protocol has been particularly designed to perform statistical analysis regarding the main measures only, i.e. the biomarkers of vigilance and attention. In contrast, statistical significance is not targeted for the additional measures due to the small number of data collected. To carefully explore performance in such tasks, we need to include more conflicts. However, such a change would have been against our aim to create an ecological environment: operators are likely to encounter few events in automated environments. Moreover, OOTL phenomenon is known to appear in case of highly reliable system. Increasing conflict could impact human system interaction failure as it appears in ecological environments. At the same time, to manage both Mind Wandering and Mental Workload, we need to include more subjective reports which would have induced an increase in vigilance and thus be detrimental to the hypothesis under investigation. In that context, our additional measures are expected to indicate tendencies (i.e., results which not reached the significance threshold) but no generalizable results are expected. Finally yet importantly, it is widely demonstrated in scientific literature the higher sensitivity of the neurophysiological measures compared to traditional techniques such as questionnaires, i.e. the same effect needs larger subjects sample to be highlighted by traditional techniques[15].

3.4.1 Biomarkers of Vigilance and Attention

EEG and eye-tracking data were gathered to examine differences in vigilance and attention between BASELINE and SOLUTION scenarios. Vigilance data, in terms of scores, was primarily derived from the EEG system. Additionally, eye gaze behaviour served as an indicator of vigilance and attention in both scenarios. This included the eye fixations per second and the Time-to-First-Fixation (TTFF) for each aircraft presented during a scenario. The former was used as an indicator of general activity and therefore vigilance. The latter was used as an indicator of attention as a more attentive controller was hypothesised to earlier fixate on newly introduced aircraft than a less attentive one.

3.4.1.1 Vigilance

During the MINIMA evaluation, the Power Spectral Densities (PSD) of the EEG signal have been used as informative feature. In general, scientific literature demonstrated how the PSDs are a powerful method to investigate brain cortex behaviour with respect to specific cognitive phenomena. Also, specific changes of brain rhythms over particular cortical sites are related to vigilance increasing/decreasing (please see Deliverable 1.1 for a comprehensive Literature Review). However, the findings are very varied, sometimes also in contradiction, and usually they have been obtained in laboratory environments, thus with very controlled experimental settings.

For such a reason, the Consortium decided to arrange some preliminary experiments, hereafter named Calibration experiments, in order to develop and tune the vigilance measurements for the specific application investigated within the MINIMA project. Also, these experiments have been useful to test the whole experimental setup that would be used during Evaluation, in order to understand all the possible issues and criticalities.

The following paragraphs aim to describe the Calibration experiments and to present the results.

3.4.1.1.1 The Calibration Experiments

Five ENAV ATCos performed these experiments during July 2017, at the University of Bologna Virtual Reality Laboratory on a first release of the DLR TE simulator. They signed an Informed Consent before the experiments themselves. The experiments were conducted following the principles outlined in the Declaration of Helsinki of 1975, as revised in 2000, and received the favourable opinion from the Ethical Committee of University of Bologna. Note that the same applies to the actual Evaluation Study reported within this document.

3.4.1.1.1.1 Material and Methods

During these preliminary experiments the brain activity has been recorded by means of the BEPlus (EB Neuro Spa, Italy), a wired EEG system able to record up to 64 EEG channels with a sampling frequency up to 2056 Hz. In this case, the EEG signals have been recorded through 40 electrodes placed all over the scalp, with a sampling frequency of 256 Hz. A pair of electrodes on the earlobes has been used as reference, while the system ground has been placed on the left mastoid. The impedance of all electrodes was kept below 10 kΩ.

The ATCos, after a training period to take familiarise with the ATM Platform, performed two scenarios:

- A 15-minutes-long scenario, hereafter named "CALIBRATION SCENARIO", needed to calibrate the Vigilance observer as described in the section 3.2.3.2during this scenario two questions were asked to the subjects during the first 5 minutes, in order to induce a higher vigilance level at the beginning;
- A 20-minutes-long scenario, hereafter named "TESTING SCENARIO", very similar to the BASELINE scenario (see section 3.2.2.1), i.e. with the automation level constantly set on the maximum level (LEV. 2) along the whole scenario.

The EEG signals Power Spectrum Densities (PSD) for each electrode have been calculated in the three main frequencies of interest: Theta [IAF – 6 ÷ IAF – 2] Hz, Alpha [IAF – 2 ÷ IAF + 2] Hz and Beta [IAF + 2 ÷ IAF + 18] Hz, where IAF stands for *Individual Alpha Frequency* [8], estimated for each subject before the experiments.

The aim was to:

1. Identify the EEG brain features mostly related to vigilance variations induced by the CALIBRATION SCENARIO, hereafter labelled as "CALIBRATION";
2. Verify that the highlighted brain features are also sensitive towards spontaneous variations in vigilance, as those one expected during the TESTING SCENARIO, hereafter labelled just as "SCENARIO".

Finally, once defined the vigilance-related brain features, the BS recorder was tested in order to:

3. Verify its effectiveness in measuring different levels of vigilance during the scenario.

The BS recorder is based on the *automatic-stop-StepWise Linear Discriminant Analysis* (asSWLDA), a machine-learning technique developed and patented by BrainSigns[9] to measure human mental states on the basis of EEG activity.

3.4.1.1.1.2 Results

In order to meet the first two objectives, two segments of the respective scenarios (CALIBRATION and SCENARIO) have been selected, in particular the first and the last five minutes. In fact, because of the vigilance decreasing induced by monotonous tasks, the two segments within the same scenario would be different in vigilance:

- CALIBRATION: between the first (min 1 ÷ 5) and the last (min 10 ÷ 15) minutes the effect would be enhanced because of the questions at the beginning, in addition to the consequences of monitoring a long monotonous task;
- SCENARIO: between the first (min 1 ÷ 5) and the last (min 15 ÷ 20) minutes the effect would be lower but still present because of the task length.

Figure 3 shows the comparison of EEG PSDs for each band between the beginning (High Vigilance, *HIGH VIG* in the Figure) and the end (Low Vigilance, *LOW VIG* in the Figure) of both scenarios. For each brain rhythm (Theta, Alpha and Beta) two information are provided:

a) *"Maps"*: the scalp map represents the *t* of repeated t-tests for each electrode between the two conditions (HIGH vs LOW), with a colour-code varying from hot ($t > 1$, the PSDs in that band on that electrode during the first 5 minutes were higher than during the last 5 minutes) to cold colours ($t < 1$, the PSDs during the first 5 minutes were lower than during the last 5 minutes);

b) *"Features trends"*: the bar graphs the activation/deactivation trends for those areas that showed a similar behaviour for both the scenarios (CALIBRATION and SCENARIO), providing a complementary visualization arising from the maps, since the bars are calculated by using the PSDs values averaged on those specific brain areas.

Because of the low sample dimension (5 ATCos), any significant statistical analysis was not possible, however the trends seem to reveal a relationship between vigilance decreasing and decreasing in frontal and parietal Theta, increasing in centro-parietal High Alpha ([IAF ÷ IAF + 2] Hz), and increasing in frontal Beta.

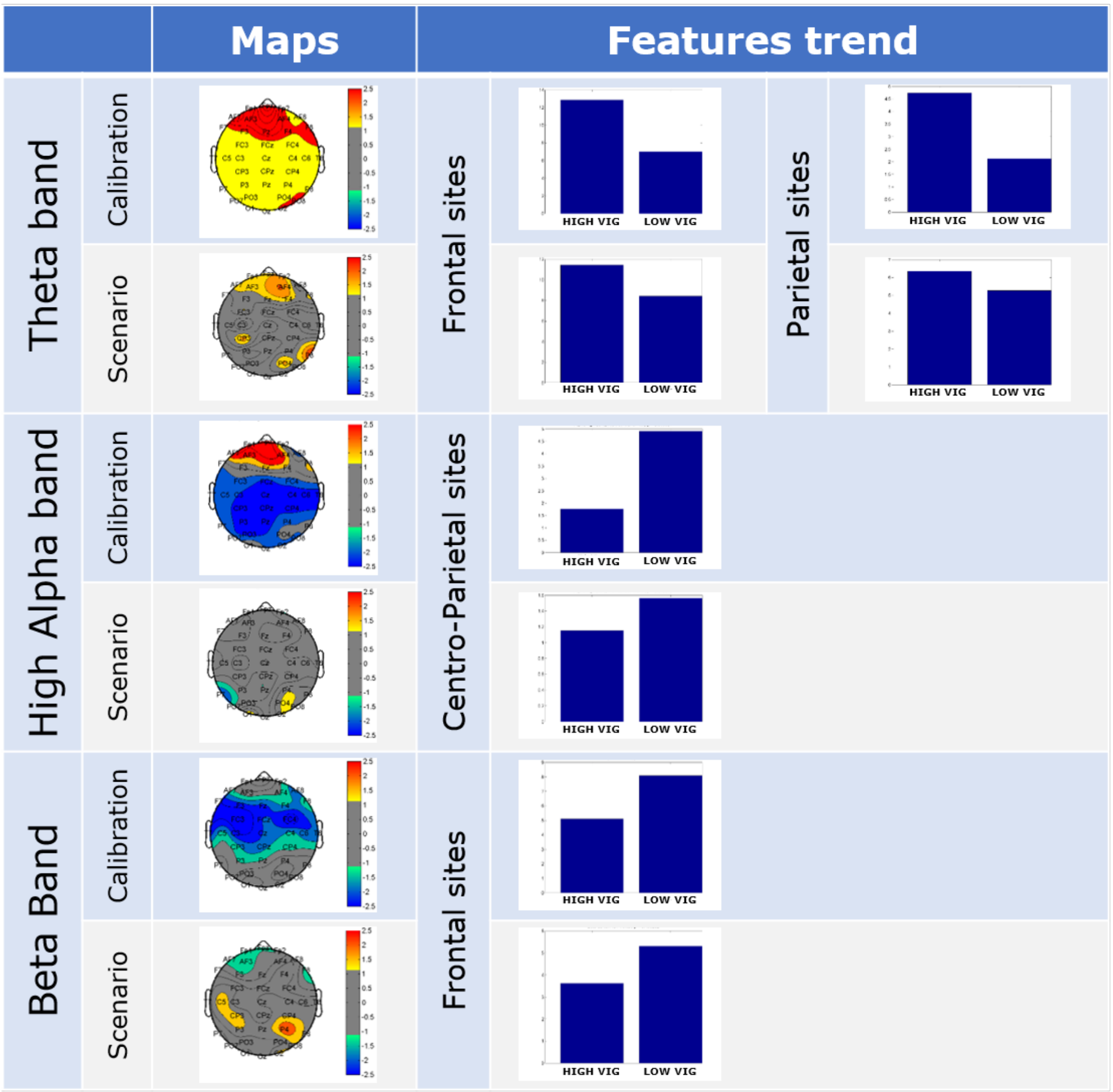

Figure 3. Comparison of EEG PSDs for each band between the beginning and the end (i.e. First 5 vs Last 5 minutes) of both the scenarios. The scalp map represents the t of repeated t-tests for each electrode, with a colour-code varying from hot (t > 1, the PSDs in that band on that electrode during the first 5 minutes were higher than during the last 5 minutes) to cold colours (t < 1, the PSDs during the first 5 minutes were lower than during the last 5 minutes). The bar graphs show the activation/deactivation trends for those areas that showed a similar behaviour for both the scenarios.

Therefore, these brain activity features have been selected to calibrate the Vigilance observer (3rd objective), in order to test its classification performance in distinguishing the two different segments of the TESTING scenario. By using all the electrodes, the Vigilance observer was able to achieve a classification accuracy of 82 ± 15 %. Also, it has been possible to reduce the electrodes number up to 14, by ensuring both the same classification performance and covering all the relevant areas of interest (frontal and centro-parietal). In particular, the following electrodes have been selected: AF3, AF4, AF7 AF8, Fz, F3, F4, F7, F8, CP3, CP4, Pz, P3 and P4. In Figure 4 the two EEG configurations (Calibration and Evaluation experiments) are showed.

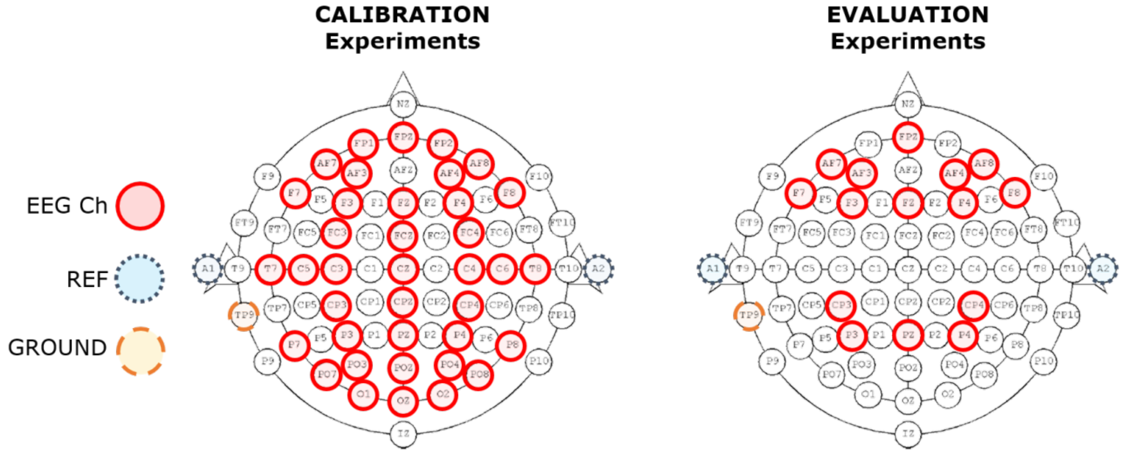

Figure 4. EEG configurations used during Calibration and Evaluation Experiments.

3.4.1.1.1.3 Conclusion

The Calibration experiments allowed to:

- Identify the brain activity features related to Vigilance, in particular frontal and parietal Theta, centro-parietal High Alpha and frontal Beta;
- Test the reliability of the Vigilance observer, that achieved a classification accuracy of about 82 % on average on 5 subjects;
- Reduce the electrodes number from 40 to 14.

3.4.1.1.2 The Vigilance Observer outline

- The following outline resumes the steps performed by means of the Vigilance observer in order to produce the Vigilance assessment during the experimental scenarios:15-channels EEG montage;
- 1 minute recording with closed eyes for the IAF estimation;
- 1 minute recording with open eyes for subjective parameters estimation;
- 15 minutes of CALIBRATION of the Vigilance observer on the individual brain activity;
- Online use on the experimental scenarios. The Vigilance observer was measuring Vigilance with a 30-seconds time resolution (one Vigilance score each 30 seconds); on the basis of these scores, each 5 minutes a classification (two classes: *Low* and *High*) of the ATCo recent Vigilance Level is performed, and during the MINIMA SOLUTION Scenario (see section 3.2.2.2) it is used to adapt the level of interface automation.

3.4.1.2 Eye Gaze Behaviour

Two dependent variables were derived from the eye-tracking data recorded in the data base of each subject: Fixations per second and TTFF of aircraft.

Fixations per second were calculated as the total number of fixations divided by the total duration of fixations. For example, if a total of 10,000 fixations with a total duration of 4,000 seconds were recorded for one subject, fixations per second were calculated as 10,000 fixations / 4,000 seconds = 2.5 fixations per second. As pointed by Jepma and Nieuwenhuis (2011), the adaptive regulation of the balance between exploitation and exploration is critical for the optimization of behavioral performance. In our case, less fixation could reflect to a better processing of these task-relevant stimuli (see Adaptive Gain Theory, Verney, Granholm, & Marshall, 2004; Wierda, van Rijn, Taatgen, & Martens, 2012). Therefore, it was expected that fixations per second would be lower during SOLUTION scenario.
TTFF was used as an indicator of controller attention. It was calculated as the difference in time between the appearance of an aircraft on the RadarVision display and the first fixation of the same aircraft. For example, if an aircraft appeared at timetick 50,000 and was first fixated by the controller at timetick 50,010, TTFF for that aircraft was TTFF = 50,010 – 50,000 = 10 seconds. A lower TTFF was considered to indicate a higher level of attention.

In addition to the two primary eye-tracking indicators described above, the general distributions of fixations throughout scenarios were used to evaluate general gazing behaviour. This was done to ensure that subjects did in fact monitor all of the TMA throughout each scenario.

3.4.2 Additional measures

3.4.2.1 Subjective assessment

In addition to the objective measurements gathered from EEG, eye-tracking and behavioural measures, subjective measures of mind wandering and workload were assessed using post-trial questionnaires. The former was assessed using an adapted version of the DSSQ, the latter using the NASA-TLX. Both questionnaires were prepared as electronic online questionnaires and presented using the online survey platform LimeSurvey.

3.4.2.1.1 Mind Wandering

The DSSQ has been used as a measurement of mind wandering episodes [10]. More precisely, it contains a 'Thinking Content' component which can be interpreted as an indicator of mind wandering experiences and has been used as such in the past[11]. This component further consists of two sub-scales: 'Task-Related Interference' and 'Task-Unrelated-Thought' (see Annex A). Both can be used as an indicator of mind wandering episodes. In addition to those two sub-scales, a number of items were developed by the MINIMA team which specifically ask subjects about 'Task-Related Thoughts' (see Table 2). Those items were designed to inquire to which extend subjects thought about anything related to their actual task. As the MINIMA concept aims to reduce mind wandering episodes and therefore keep subjects focused on their task, this dimension was deemed useful for overall evaluation.

The DSSQ items required subjects to rate how often they thought about different things during the last scenario they had completed. All items were given in a conjoint table starting with the phrase: "During the last scenario, I thought about…". Frequency of respective thoughts was rated on a 5-Point-Likert-Scale from 1 (Never) to 5 (Very often). Items were randomly arranged between subjects and scenarios to control for sequence effects.

Table 2. 'Task-Related Thought' items of the newly created 'Thinking Content' component of the DSSQ.

#	During the last scenario, I thought about…
1	…the current traffic situation
2	...something related to the task at hand
3	...the relevance of the system's decisions
4	...potential conflicts of aircraft
5	...how the system resolved potential conflicts

3.4.2.1.2 Workload

The NASA TLX questionnaire (NASA, 1986 [12]) was used to evaluate workload along six dimensions (see Table 3). Administration of the NASA-TLX was done in two parts as per its manual. First, participants were asked to rate the extent of each dimension during the last scenario they had completed. Ratings were given using a horizontal line, ranging from "Low" to "High" on a scale from 0 to 20 (see Annex A). In part two, all combinations of the six dimensions were presented to the subjects. For each pair, subjects should decide which of them they deemed more important to how demanding the last scenario was. Those pairwise comparisons were later used to weigh the ratings of each dimension and calculate an overall workload score.

Table 3. Dimensions of the NASA Task Load Index.

Dimension	Scale Endpoints	Description
Mental Demand	Low/High	How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical Demand	Low/High	How much physical activity was required? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand	Low/High	How much time pressure did you feel due to the rate of pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Effort	Low/High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration	Low/High	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?
Performance	Good/Poor	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

3.4.2.2 Behavioural measure

Behavioural measures were used to quantify the effect of the MINIMA SOLUTION. The lack of operator involvement in supervisory modes and passive information processing was expected to contribute to critical human cognitive errors. Amongst others, literature has shown that the OOTL phenomenon is characterized by difficulties to detect and to understand critical situations. Such critical situations were some separation losses occurring in the BASELINE and SOLUTION scenarios by intent. It was expected that during the SOLUTION scenario, subjects were able to prevent such separation losses better than in the BASELINE scenario as the MINIMA concept was designed to keep them in-the-loop and focused on the task.

Separation losses were calculated based on the radar data recorded in the SQL data base of each subject and scenario. As approach air traffic was simulated, minimum separation was set to three nautical miles horizontally and 1,000 feet vertically. A separation loss occurred whenever both thresholds were undercut by two aircraft. Departure aircraft and aircraft outside of the subjects' responsibility were excluded from such analyses.

3.5 Technical issues and data exclusion

Despite not affecting the achievement of the experimental aims, some technical difficulties occurred during the experiments which have to be considered for the data analysis. First, some crashes of the RadarVision software occurred throughout the *Evaluation Study*. Second, it appears that in some cases altitude and speed advisories given by the controllers were not properly implemented into the simulation environment. Both problems are reported in more detail, along with the ways they were handled prior to data analysis.

3.5.1 RadarVision software crashes

The RadarVision software used as the human machine interface in the task environment encountered some crashes throughout the experiments (see Table 4). Every time such a crash occurred, RadarVision had to be restarted to continue the scenario. As the traffic simulation itself was executed by a separate system, from a technical perspective this was a minor problem. RadarVision could easily be restarted to continue the scenario with only a few seconds lost in between. However, the experimental design aimed to induce OOTL phenomena through low vigilance levels caused by a monotonous task. As software crashes were highly salient, they obviously caused a major impact on the subjects' vigilance level at the time of the crash. Therefore, crashes were handled in two ways depending on the time they occurred. If they occurred early in the scenarios (less than 30 minutes into), the scenario was restarted completely if there was sufficient time to do so. If they occurred later than 30 minutes into the scenario, the scenario was not repeated and only the data obtained up to that point was included in the analysis.

As can be seen in Table 4, software crashes did not have a critical impact in most cases. Three out of 15 subjects encountered crashes during the experimental session. For three out of five crashes, complete data sets could still be recorded. In the two remaining cases, 31 and 37 minutes of data were recorded. As all measures described in section 3.4 either do not need standardisation or are standardised by duration anyway, the data was still included in the analyses.

While the software crashes could be dealt with in most cases, they still caused the drop out of one subject (ATCO_08). During the training session, the subject encountered several crashes which severely interrupted the whole process and led to high time pressure. It was decided to exclude the subject from the experimental session and therefore lose one subject. As the time schedule was very tight, no replacement subject could participate.

Table 4. Report of RadarVision software crashes in experimental scenarios (BL - Baseline, SO - Solution, TR - Training).

Subject	# of crashes	by Scenarios	Minutes of data obtained
ATCO_02	2	1x BL, 1x SO	45 (BL), 31 (SO)
ATCO_07	1	1x BL	45
ATCO_08	3	3x TR	Excluded from experiment
ATCO_11	2	1x BL, 1x SO	45 (BL), 37 (SO)

After the first crashes appeared, DLR tried to identify what caused them but without success. Based on the feedback of UNIBO, crashes occurred randomly at different scenarios and different times. No systematic errors could be found. As can be derived from Table 4, most subjects did not encounter any crashes at all.

3.5.2 Advisory implementation issues

It has been reported by the experimenters that during the *Evaluation Study*, participants sometimes found themselves in situations where their altitude and speed advisories were not properly implemented by the system. For instance, if a controller advised flight level 60 to an aircraft, the aircraft apparently did not follow the advisory. As continuous descent approaches were simulated, advisories were limited to 'descend' (altitude) and 'reduce' (speed). Continuous descent also means that aircraft continuously lost height throughout their approach to the simulated airport. Therefore, it is possible that advisories were mistakenly perceived as not being implemented because aircraft were reducing their height anyway.

Another issue with the height advisories was that BASELINE and SOLUTION scenarios were designed to have one separation loss between two aircraft. One aspect of the MINIMA concept's evaluation was to see if controllers were more likely to prevent the conflict in the SOLUTION scenario as their vigilance was supposed to be higher compared to the BASELINE scenario. However, as altitude advisories influenced the planned and then flown trajectory, controller interaction might have caused only very small effects from time to time. Controllers could have perceived this as an unreliable way to prevent the separation losses and thus may have reduced their interactions. However, further analysis of the recorded data (1) any such potential perception has not shown any obvious systematic bias (2) did show that all advisories were implemented correctly by the system. The latter was shown through analysis of changes in flight behaviour of aircraft before and between controller commands.

3.5.3 Other issues

In addition to the issues reported in the two foregoing sections, there was one more instance which led to the exclusion of one subject's data from the analysis. Namely, subject ATCO_10 was a substitute subject who readily participated to replace a last minute dropout. Unfortunately, the subject did not have sufficient time to complete all of the experiment. Furthermore, the training session was omitted. Although the controller had already participated in the preliminary trials in July and underwent the training session then, it was not enough for the standardised evaluation design. Therefore, the subject's data was also excluded from the analysis.

3.5.4 Summary of data exclusion

Originally, it was planned to collect 30 data sets (15 subjects x 2 scenarios). After the data exclusion reported in sections 3.5.1 – 3.5.3, a total of 26 data sets were collected (from 13 subjects). Data sets had an average duration of *M* = 43.75 minutes (30.90 minutes in the shortest scenario).

4 Results

In this chapter, the results obtained from the data gathered through the methods described above are reported. As previously introduced, our MINIMA tools aimed (1) to measure the current attention level and the attention focus of the human operator with the aim to detect or anticipate typical OOTL performance issues and (2) to adapt automation in case of vigilance decrement with the aim to compensate it. The analysis of our results is mainly driven to demonstrate the relevance of the MINIMA tools in regard to these two objectives. Additional measures are also presented.

4.1 Main measures

4.1.1 EEG and Classification of mental state

In Figure 5, the Vigilance score distributions measured during the two scenarios are showed. The two tails paired T-test highlighted a significant increasing ($p = 0.042$) of the overall vigilance scores during the MINIMA SOLUTION scenario, i.e. it kept the ATCos more vigilant.

In Figure 6, the time percentage of the scenarios classified as “Low Vigilance” is showed. The two tails paired T-test highlighted a significant decrease ($p = 0.0027$) of the time spent by the ATCo in a “Low Vigilance” condition during the MINIMA SOLUTION scenario, i.e. again it was able to keep the ATCos more vigilant. Taken together, these two results indicate that the MINIMA solution succeed to increase the operator vigilance.

Finally, Figure 7 shows the Vigilance scores evolution along the time (the scenario has been divided in 5 minutes long windows to facilitate the representation), averaged between the subjects for both the scenarios. It is impossible to perform any statistical analysis, since (i) the subjects number (13) should be at least one magnitude order higher than observations; and (ii) probably each subject experienced vigilance decreasing in different moments of the task, thus it is difficult to select two segments to compare. However, it is evident the higher decreasing trend of the Vigilance scores for the baseline scenario.

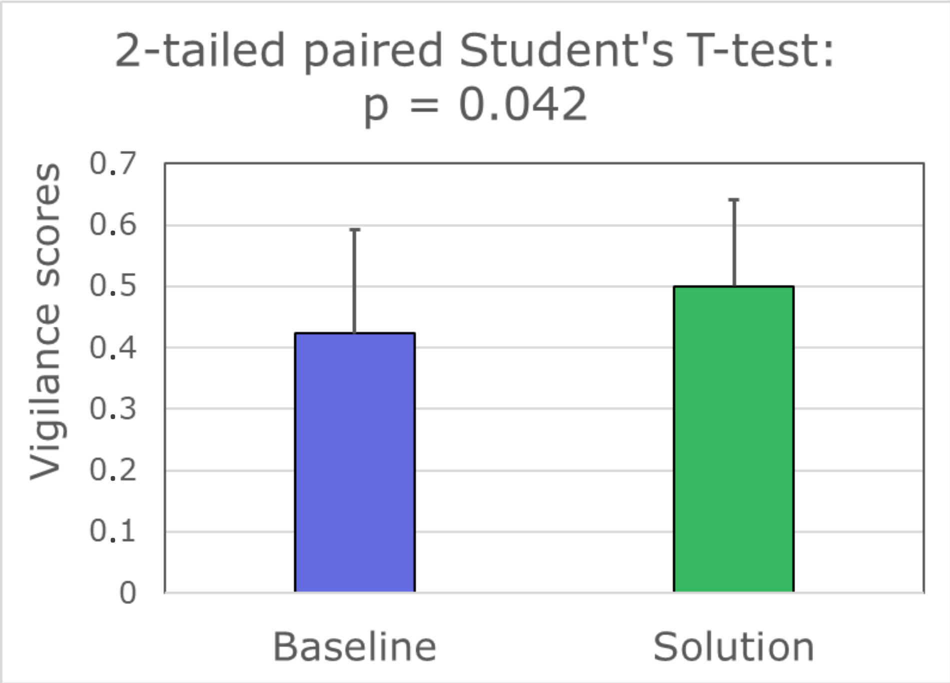

Figure 5. Two-tailed paired Student's T-test between the distributions of the EEG-based Vigilance scores measured during BASELINE and MINIMA SOLUTION scenarios.

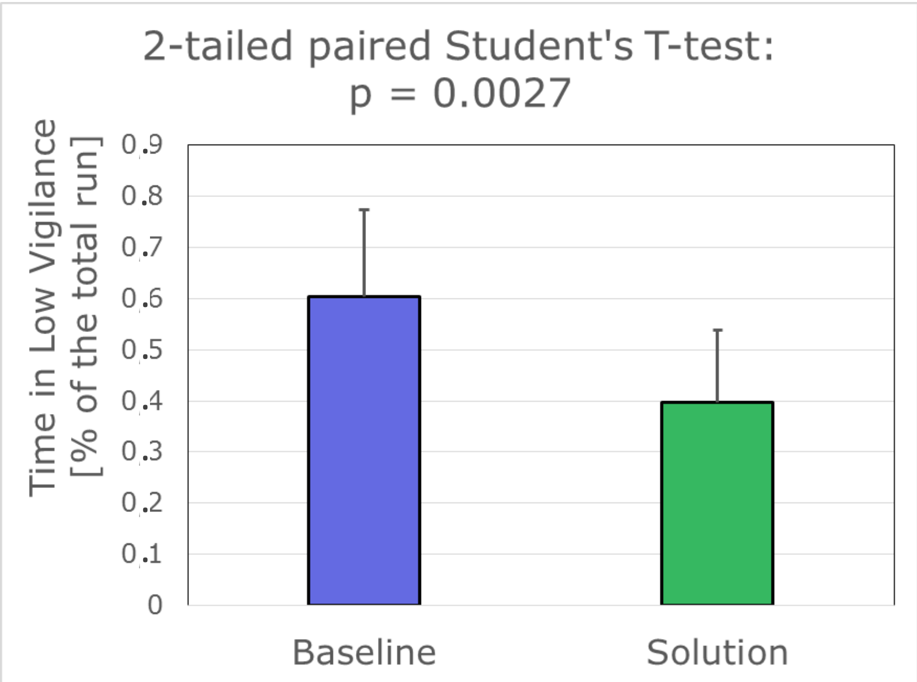

Figure 6. Two-tailed paired Student's T-test between the distributions of the time durations (in % related to the whole scenario duration) of the run periods classified as "Low Vigilance", for both the BASELINE and MINIMA SOLUTION scenarios.

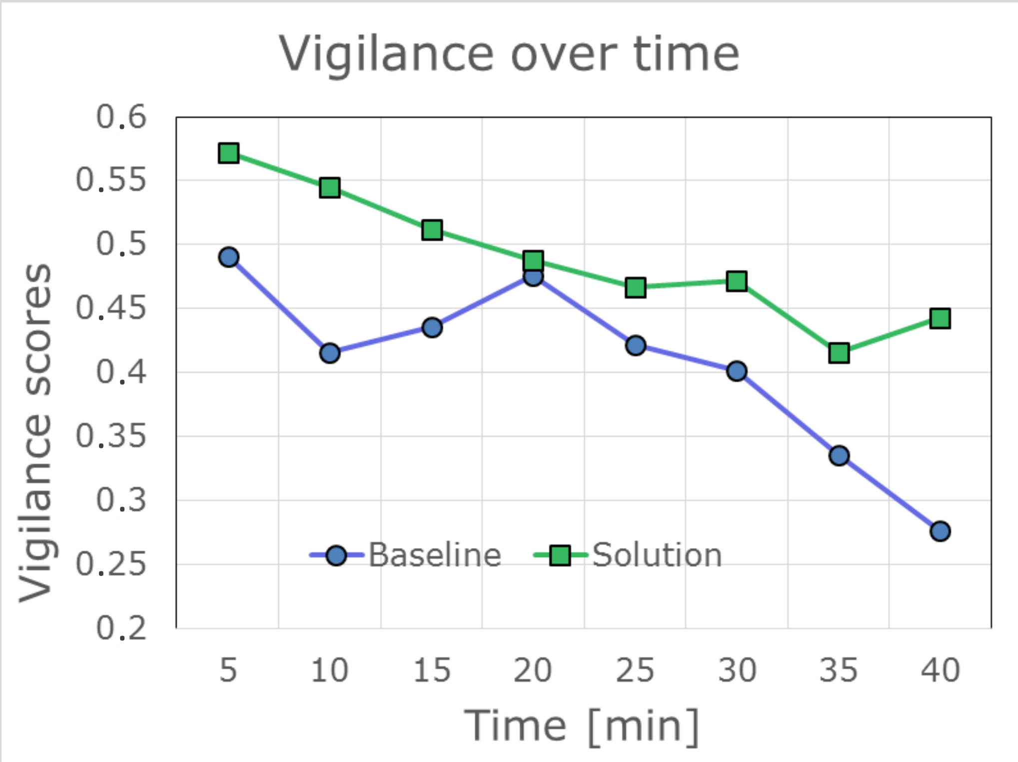

Figure 7. EEG-based Vigilance scores trends along the time, averaged between the subjects for both the scenarios.

4.1.2 Oculometric measures

4.1.2.1 Eye gaze distribution

In Figure 8, the heat maps of eye gaze distributions are shown, separated by scenarios and vigilance levels. All four diagrams show no noticeable 'hot spots' in any condition. However, distributions between SOLUTION (SO) and BASELINE (BL) scenario show that subjects did fixate less on the southern part of the TMA during BASELINE condition. This was consistent with the shift in responsibility areas during low vigilance level episodes of the SOLUTION scenario. To some extent, it can also be seen that the distribution of high vigilance level episodes in the SOLUTION scenario was a bit more dense in the southern TMA portion compared to low vigilance level episodes in the same scenario. This was also consistent with the adaptive automation as the TMA was only extended during low level automation episodes (which induced high vigilance levels).

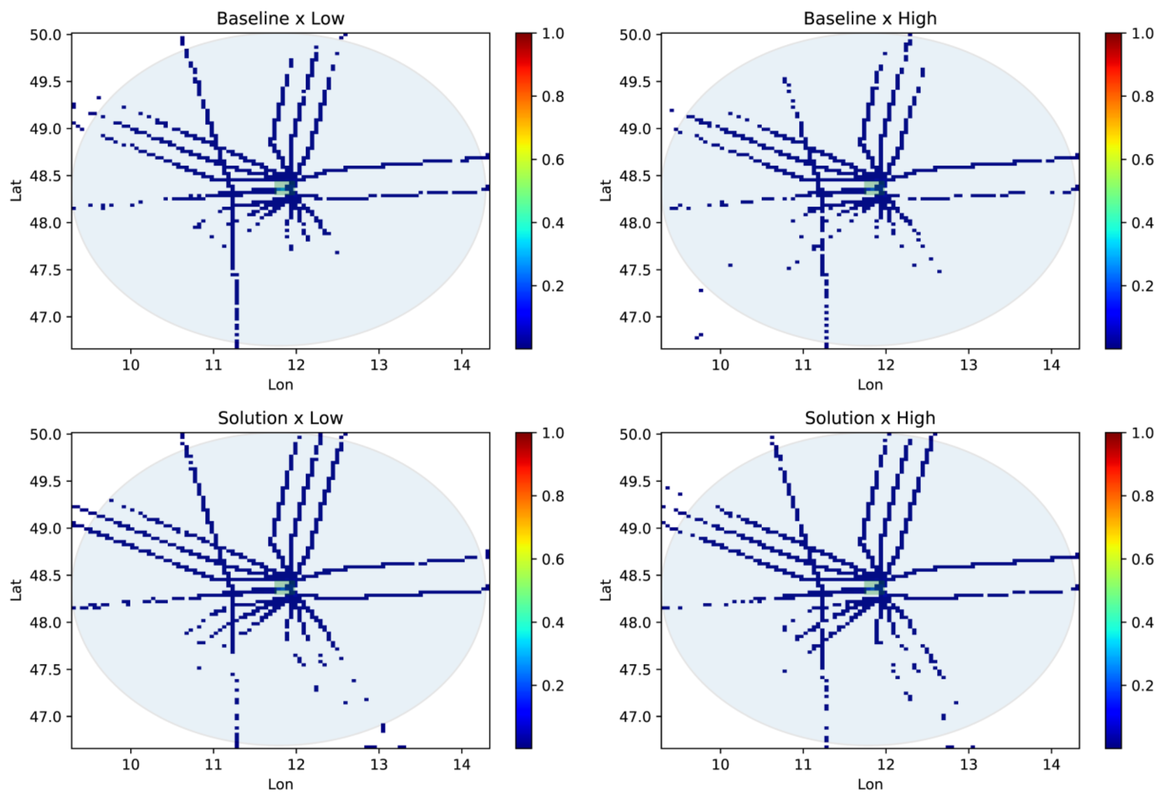

Figure 8. Gaze distributions between scenarios (Baseline vs. Solution) and vigilance levels (Low vs. High)

4.1.2.2 Fixations and Time-to-First-Fixation

As described in section 3.4.1.2, eye gaze fixations were normalised as fixations per second. Descriptive statistics of fixations per second between scenarios and vigilance levels are given in Table 5. As three of the four variables showed high skewness (Skew) and kurtosis (Kurt), they were considered non-normally distributed.

Table 5. Descriptive statistics of eye fixations per second.

Scenario	Vigilance	M	Mn	SD	Min	Max	Skew	Kurt
BL	Low	5.01	3.79	3.00	2.91	12.51	1.85	3.03
	High	4.17	3.76	1.47	2.14	6.82	0.88	-0.26
SO	Low	4.44	3.82	1.92	2.94	9.75	2.29	5.48
	High	4.23	3.85	1.71	2.54	9.27	2.66	8.20

As normal distribution was not given, non-parametric tests were performed to analyse differences in fixations per second between scenarios and vigilance levels. Wilcoxon-tests were calculated between conjoint variables of scenarios (BASELINE vs. SOLUTION) and vigilance levels (Low vs. High). Significantly less fixations per second where found in the SOLUTION scenario compared to the BASELINE scenario (Z = -2.12, p = .034). Such results seem indicate that our MINIMA solution could optimize performance in the current task by promoting exploitation behaviour.

Descriptive statistics of Time-to-First-Fixation (TTFF) between scenarios and vigilance levels are given in Table 6. TTFF in BASELINE scenario and low vigilance showed a high skewness and kurtosis and therefore could not be considered to be normally distributed.

Table 6. Descriptive statistics of Time-to-First-Fixation.

Scenario	Vigilance	M	Mn	SD	Min	Max	Skew	Kurt
BL	Low	390	394	129	65	596	-1.41	4.57
	High	354	314	194	84	730	0.56	0.02
SO	Low	378	382	93	244	528	-0.10	-1.10
	High	249	259	92	53	401	-0.65	1.04

A 2x2 Analysis of Variance (ANOVA) was calculated to test differences in TTFF between scenario and vigilance levels for statistical significance. The results showed a significantly lower TTFF in the SOLUTION scenario ($F(1, 10) = 5.27$, $p = .045$, $\eta = .35$). As depicted in Figure 5, TTFF was lower in SOLUTION scenario when vigilance was high which indicate a better performance when MINIMA solution is activated.

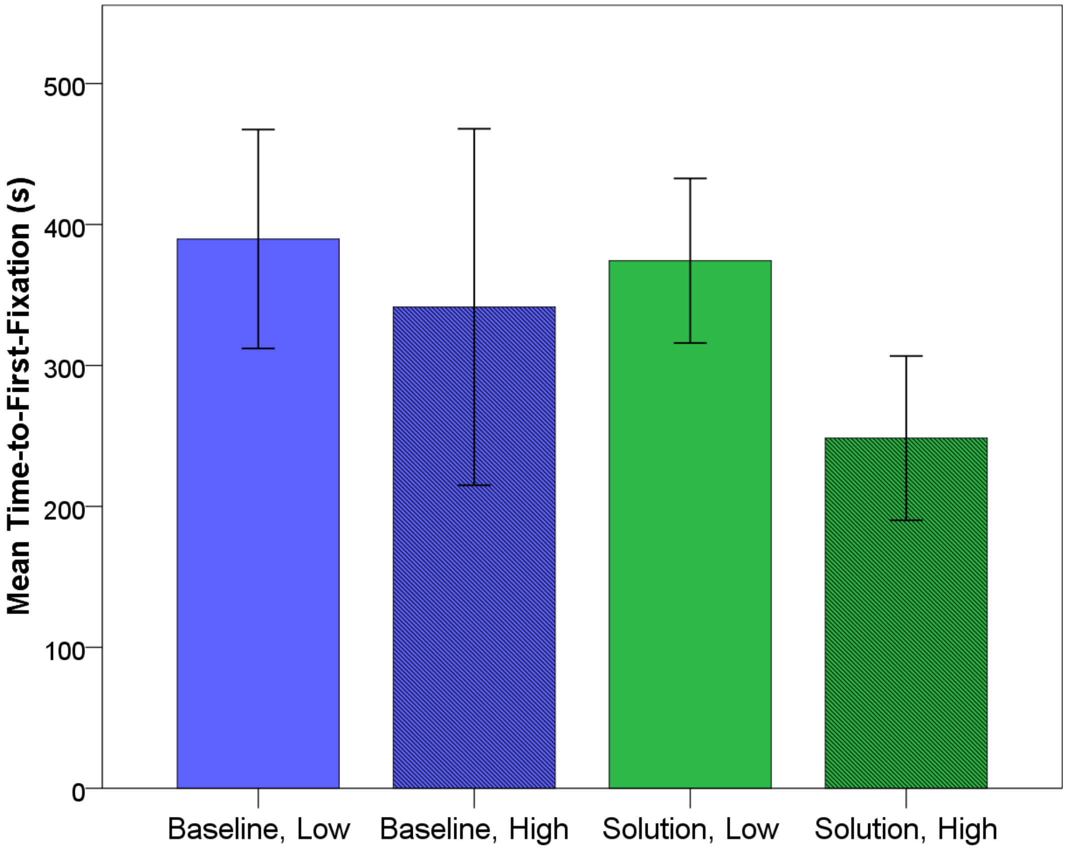

Figure 9. Mean Time-to-First-Fixation by Scenario (Baseline vs. Solution) and Vigilance Level (Low vs. High).

4.2 Additional measures

4.2.1 Performance measures

As described in section 3.4.3, separation losses were used as performance indicator. First, two sub-parameters of separation losses were considered for analysis of performance differences between scenarios and vigilance levels. First, the overall number of separation losses per scenario and vigilance level was calculated. This calculation only included unique conflicts, i.e. each pair of conflicted aircraft was only counted once. Second, the total overall duration of separation losses was calculated by summing the durations of each individual conflict.

As vigilance levels were dependent on individual differences, durations of both low and high vigilance episodes were not standardised among subjects. Such differences would have resulted in a systematic bias of both separation loss parameters. More precisely, having one separation loss in ten minutes is completely different from having one in 100 minutes. Likewise, a total separation loss duration of ten minutes in a ten minute scenario is completely different from ten minutes in a 100 minute scenario. Such bias was prevented by standardising the total duration of separation losses in each combination of scenario and vigilance level by the respective total duration of the respective combination. For example, if a controller had ten minutes of high vigilance in the SOLUTION scenario and within those ten minutes separation losses totalled to 6 seconds, the performance indicator was calculated as 6 seconds divided by 600 seconds (ten minutes x sixty seconds). Performance of this controller in terms of conflict time was 6 / 600 = 0.01 (1 %). Thus, performance took into account the number and duration of separation losses during each combination of scenario and vigilance level, as well as the total duration of each combination. Finally, to prevent potential confusion, performance was calculated as the relative portion of conflict-free time by subtracting the former indicator value from 1 (Performance = 1 – (6 / 600) = 0.99 [99 %]). This way, a higher value actually indicated better performance, which appeared more reasonable for interpretation.

Descriptive statistics of conflict-free time portions are given in Table 7. As three out of four variables showed high values of either skewness or kurtosis, variables were considered non-normally distributed.

Table 7. Descriptive statistics of portion of conflict-free simulation time.

Scenario	Vigilance	M	Mn	SD	Min	Max	Skew	Kurt
BL	Low	.89	.89	.09	.76	1.00	0.13	-1.58
	High	.97	1.00	.07	.78	1.00	-2.11	3.83
SO	Low	.91	.99	.13	.60	1.00	-1.48	1.25
	High	.82	.79	.15	.53	1.00	-0.24	-0.53

Wilcoxon-Tests were calculated to test differences in performance between scenarios and vigilance levels. No significant differences were found between scenarios (SOLUTION – BASELINE, $Z = -1.96$, $p = .050$) or vigilance levels (High – Low, $Z = -0.24$, $p = .814$). However, it must be noted that the p-value for the difference between scenarios lay exactly on the brink to statistical significance. Moreover, mean performance ranking was higher in the BASELINE scenario (see Figure 10).

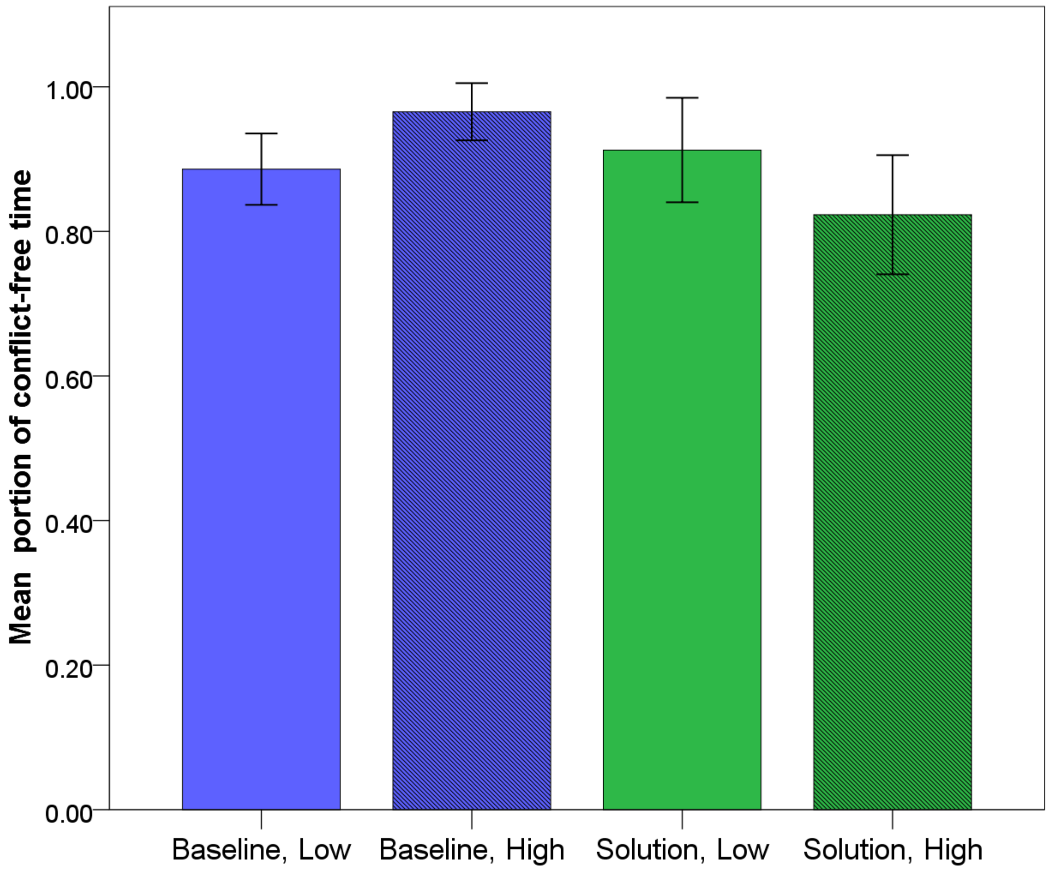

Figure 10. Mean portion of conflict-free time by scenario and vigilance level.

4.2.2 Subjective measures

4.2.2.1 Workload

Descriptive statistics of the NASA-TLX scores are given in Table 8. As a number of variables showed had very high values of skewness and kurtosis, those variables were considered to be non-normally distributed.

Wilcoxon-Tests were calculated to test differences in the NASA-TLX dimensions between BASELINE and SOLUTION scenario. The results of these tests are given in Table 9. None of the tests revealed any significant differences in the different workload scales. However, it has to be noted that all differences, although not statistically significant, showed the desired directions. Demands, Effort and Overall Workload had higher absolute median values in the SOLUTION scenario while Frustration and Performance had lower (= preferable) median values in the SOLUTION scenario (see also Figure 11).

Table 8. Descriptive statistics of the NASA Task-Load Index subscales (weighed values, range 0 – 33) and overall workload (sum of weighed values, range 20 – 100).

Dimension	Scenario	M	Mn	SD	Min	Max	Skew	Kurt
Mental Demand	BL	7.50	5.83	7.16	0.67	24.00	1.02	0.50
	SO	10.51	7.00	9.31	0.67	30.00	0.89	-0.13
Physical Demand	BL	1.83	1.17	2.42	0.00	8.00	1.65	2.23
	SO	2.80	1.33	4.09	0.00	14.00	2.09	4.37
Temporal Demand	BL	5.45	2.83	5.51	0.67	17.33	1.06	0.01
	SO	4.64	4.00	2.93	0.67	9.00	0.17	-1.56
Effort	BL	5.86	3.50	5.50	1.33	19.00	1.61	1.78
	SO	7.08	5.00	6.66	0.00	18.67	0.71	-1.19
Frustration	BL	9.12	3.83	11.19	0.00	33.33	0.97	-0.19
	SO	8.97	1.00	13.44	0.00	33.33	1.20	-0.33
Performance	BL	7.93	8.00	5.75	0.00	21.67	0.76	1.28
	SO	9.10	6.67	5.10	3.33	17.00	0.31	-1.70
Overall Workload	BL	37.69	33.67	22.05	5.00	93.00	1.39	2.51
	SO	43.10	39.33	19.33	21.33	80.33	0.73	-0.57

Table 9. Results of Wilcoxon-Tests for NASA Task-Load Index scales (BASELINE - SOLUTION)

Dimension	Mental Demand	Physical Demand	Temporal Demand	Effort	Frustration	Performance	Overall Workload
Z	-.736	-.736	-.524	-.338	-.365	-.338	-.593
p	.462	.461	.600	.735	.715	.735	.553

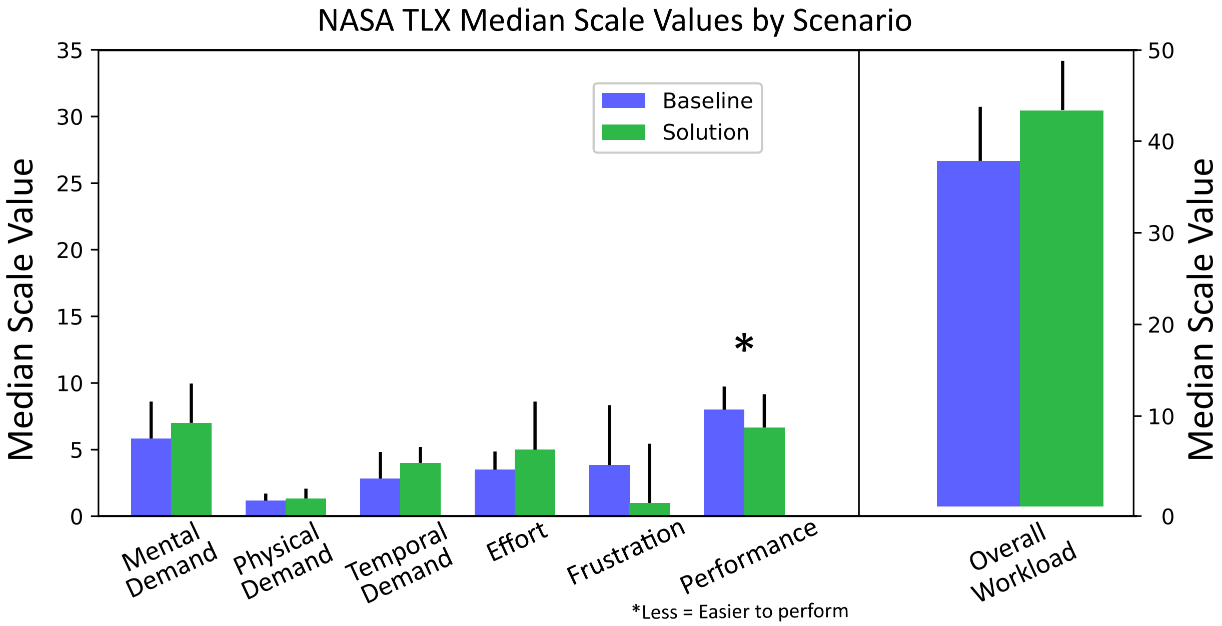

Figure 11. Median values of NASA Task-Load Index scales for BASELINE and SOLUTION scenarios.

4.2.2.2 Mind Wandering

Descriptive statistics of the DSSQ scales 'Task Related Interference', 'Task Unrelated Thought' and 'Task Related Thought' are given in Table 10. All variables showed reasonable values of skewness and kurtosis and were therefore considered normally distributed.

Paired-t-tests were calculated to test differences in the mind wandering dimensions between BASELINE and SOLUTION scenario. No significant differences were found for any of the scales ($0.59 < t < 1.91$; $.098 < p < .576$). However, as observed for Workload index, it has to be noted that all differences showed the desired directions. Indeed, we observe a decrease in non-relevant thought (task related interference and task unrelated thought) in the Solution scenario whereas task related thoughts seems equivalent between the two conditions.

Table 10. Descriptive statistics of the DSSQ subscales.

Dimension	Scenario	M	Mn	SD	Min	Max	Skew	Kurt
TRI	BL	2.28	2.13	0.65	1.25	3.25	0.22	-1.08
	SO	2.00	2.00	0.52	1.38	3.25	0.98	0.78
TUT	BL	2.02	1.75	0.74	1.00	3.25	0.27	-1.45
	SO	1.63	1.43	0.62	1.00	2.88	0.83	-0.52
TRT	BL	3.16	3.20	0.71	1.80	4.40	-0.12	-0.16
	SO	3.12	3.20	0.71	1.75	4.20	-0.33	-0.61

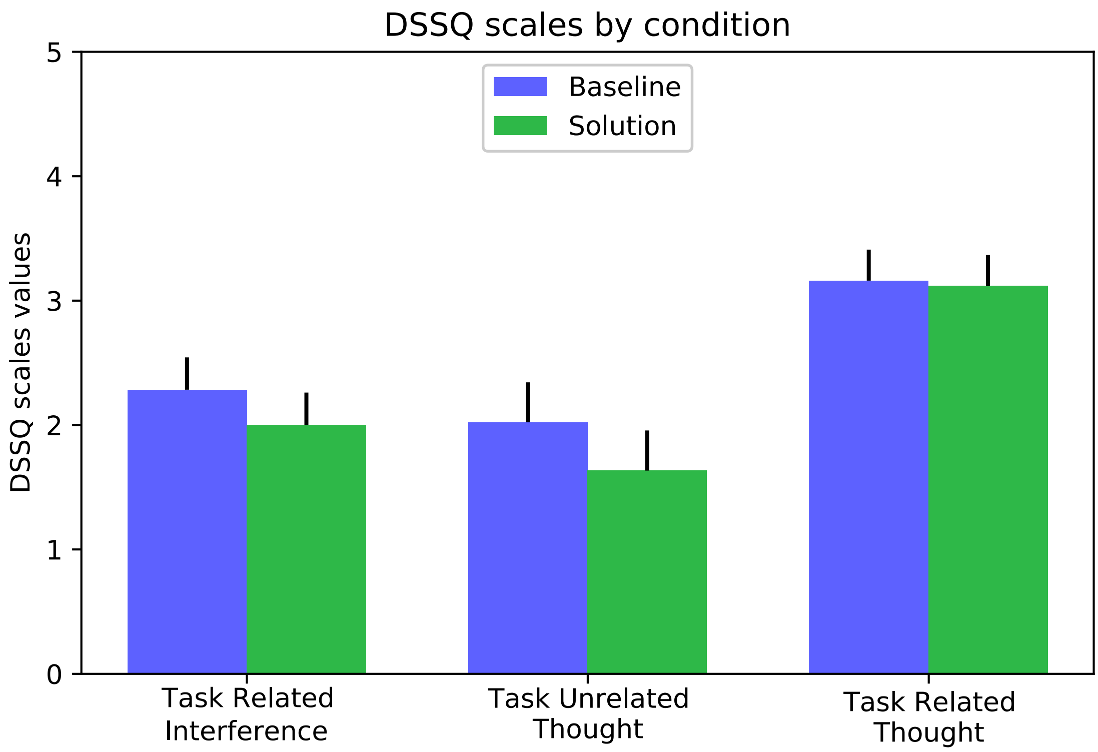

Figure 12. Mean scale values of the DSSQ subscales in BASELINE and SOLUTION scenarios. Error bars equal 2 standard errors of mean.

5 Discussion

The *Evaluation Study* reported here aimed to show to which extent the MINIMA concept is suitable to prevent OOTL phenomena in a highly automated ATC task environment. A key aspect of this evaluation was to show that 1) A continuous high level of automation (most of the ATCos operative actions are automatized within the interface) will lead to a significant decrease in vigilance over time; and 2) Use of the *VAC* developed in MINIMA, e. g. dynamically adapting the level of automation with respect to the ATCo's vigilance level, can prevent such decrease in vigilance. Furthermore, the impact of the MINIMA concept on ATC performance and system usability has been explored.

The results reported in this document show that the MINIMA concept, namely the *VAC*, did work as intended. EEG data show that vigilance was higher when the *VAC* was used, i.e. when adaptive automation helped to keep the ATCo *In-the-Loop*. Furthermore, it was shown that vigilance in the BASELINE scenario did continuously decrease over time (with a slight exception).

In addition to the EEG data, several behavioural variables revealed significant differences between BASELINE and SOLUTION condition, favouring the latter. Finally, subjective ratings on Workload and Mind Wandering given by subjects showed that the higher activation was also consciously perceived by the subjects.

5.1 MINIMA's impact on controller vigilance

The results of the EEG data analysis have shown that vigilance in fact decreased over time when controllers were not actively involved in the task. This was indicated by a significant decrease of mean vigilance during the BASELINE scenario. Consequently, as automation was constantly kept at a high level at all times, vigilance kept decreasing over time. Although a recovery phase was indicated in the middle of the scenario, vigilance continued to decrease after that, and further below the vigilance level as it was prior to the recovery phase. However, since each subject probably experienced vigilance decreasing in different moments of the scenario, fluctuations along the time are less important that the overall trend. Actually, in addition to the decrease in vigilance observed during the BASELINE scenario, vigilance was also found to decrease during the SOLUTION scenario when automation level was set to high. This further supports the hypothesis on the impact of high automation levels on controller vigilance. However, while the trend of Vigilance scores during BASELINE scenario is monotonously decreasing, during the SOLUTION the decreasing trend seems to be moderated and resulting in a plateau, theoretically because of the effect of adaptive automation. Both scenarios were virtually equal when automation was high. Therefore, it can be assumed that differences in controller vigilance did not stem from systematic bias due to difference in traffic, but actually resulted from lack of active involvement. The results in terms of scenario time percentage classified as "Low vigilance" supported these conclusions: in fact, the time spent by the ATCos in a "Low vigilance" condition was higher during BASELINE than during SOLUTION scenario. This is a clear consequence of the fact that, while during BASELINE in case of "Low vigilance" detection nothing happened, during MINIMA SOLUTION the system effectively reacted to a "Low vigilance" state occurrence increasing again the ATCo's vigilance level.

5.2 MINIMA's impact on controller monitoring behaviour

It was hypothesised that a decrease in controller vigilance would be reflected by a change in their gazing behaviour. Namely, it was assumed that controllers would less actively monitor their TMA during episodes of low vigilance. Monitoring behaviour was assessed using different dependent variables, mainly by the eye-tracking data collected through the Tobii EyeX system. Indicators of monitoring behaviour were: General gazing distribution, Time-to-First-Fixation of aircraft, and number of fixations per second.

5.2.1 Eye gaze distribution

In short, the results reported in section 4.2.1 indicate that controllers do not fail to monitor all of their area of responsibility even if their vigilance decreases over time. As the heat maps of Figure 4 show, controllers' eye gazes were evenly distributed all over their TMA in both scenarios and during both low and high level vigilance episodes. As per expectation, the heat maps clearly reflect the TMA as it was during each combination of scenario and vigilance level. Moreover, no hot spots (e. g. red areas with the colour maps used) can be seen even during low vigilance episodes. This indicates that controllers did not reduce their general monitoring behaviour, e. g. they did not neglect part of their TMA and only 'stared' at one or few parts throughout the respective episodes.

5.2.2 Time-to-First-Fixation

Another indicator of controller monitoring behaviour was the Time-to-First-Fixation (TTFF), e. g. the difference in aircraft onset time and the time at which it was first fixated by controllers. A higher TTFF meant that it took the controller longer to recognize the respective aircraft entering the TMA. It was hypothesised that controllers would show higher TTFF during low vigilance episodes as they would less actively monitor their TMA. Results reported in section 4.4.2 support this assumption. TTFF were found to be significantly lower during the overall SOLUTION scenario. Furthermore, TTFF was lowest during SOLUTION scenario when vigilance was high. Although this TTFF was not significantly different from TTFF in low vigilance episodes of the SOLUTION scenario, they show what is commonly called 'a trend' towards lower TTFF in the SOLUTION scenario when vigilance was high.

5.2.3 Fixations per second

A commonly used indicator of gazing behaviour is the number of fixations per second. It is calculated as the total number of fixations throughout the scenario divided by the scenario's duration. At first, it appears obvious to interpret a higher number of fixations per second to represent a higher level of monitoring activity. However, regarding adaptive gain theory and more particularly the Exploration-Exploitation trade-off, we assumed that a decrease in the number of fixations per second could be evidence that controllers actually carefully process what they look at (i.e., exploitation phase). As reported in section 4.4.2, fixations per seconds were indeed significantly lower during the SOLUTION scenario without a further effect of vigilance level. Given the way fixations per second were calculated, it is clear that if total numbers of fixations are equal, longer mean duration results in a lower number of fixations per second and vice versa. As discussed in section 5.2.1, controllers evenly monitored their TMA regardless of scenario or vigilance level. Hence, it can be assumed that although eye gazes, although not necessarily equal in number, were sufficient to cover the area of responsibility. Therefore, there is no reason to assume that controllers aimlessly looked around the area in either of the scenarios or vigilance levels. With that, the lower number of fixations per second during the SOLUTION scenario can be interpreted to represent a more conscious processing and thus a more careful monitoring behaviour. Additionally, TTFF were significantly lower during SOLUTION scenario and lowest when vigilance was high. This also indicates that controllers more carefully monitored their TMA, as incoming aircraft were recognised earlier compared to BASELINE scenario.

In conclusion, the results of the eye-tracking data analysis indicate that the MINIMA concept used during the SOLUTION scenario induced a more active monitoring behaviour of controllers. Although general gaze behaviour was equal among scenarios and vigilance levels, TTFF and fixations were found to be lower during the SOLUTION scenario. Therefore, incoming aircraft were recognised earlier and controllers more carefully processed information during fixations. This is consistent with the higher average level of vigilance in SOLUTION scenario as shown by the EEG results. Therefore, the eye-tracking data show that the neurophysiological reactions to lack of involvement also result in observable changes in controller behaviour.

5.3 MINIMA's impact on workload and mind wandering

In addition to the objective measures, questionnaires were used to get a subjective insight on how controllers perceived their thinking in terms of workload and mind wandering. It was hypothesised that controllers would perceive their workload to be lower when vigilance was low and vice versa. Likewise, it was expected that controllers would more likely perceive episodes of mind wandering when vigilance was low.

5.3.1 Workload

Analyses of the NASA-TLX data assessed through the MINIMA questionnaire revealed very promising results. As opposed to most studies on ergonomics, the MINIMA concept aimed to increase workload through higher task involvement. Therefore, it was hypothesised that all 'demand scales' would be higher in the SOLUTION scenario compared to the BASELINE scenario. Results reported in section 4.4.1 support this hypothesis.

Median values of mental, physical and temporal demand were higher in the SOLUTION scenario. This indicates that during the SOLUTION scenario controllers perceived the task to require them to think more, act more and do so in a more time critical manner. In total, controllers felt like there was 'more to do' during the SOLUTION scenario.

Concerning the remaining sub-scales of the NASA-TLX (Effort, Frustration, Performance), the results were also in accordance with our hypotheses. Usually, it is expected that increasing demand results in more effort, more frustration and less performance. If the MINIMA concept aimed to result in more effort as it puts the human operator back in the loop, it also aims to lower frustration stemming from the lack of involvement and increase performance.

The results of the remaining sub-scales show that during the SOLUTION scenario, controllers did indeed perceived the task to require more effort. They also rated the task to be less frustrating during the SOLUTION scenario. Finally, stated that it was easier for them to achieve good performance during the SOLUTION scenario. It is likely that those results stem from the higher degree of active task involvement during the SOLUTION scenario as their role was shifted from a mere passive monitor to an actively involved controller.

Finally, the overall workload calculated as the sum of the weighted NASA-TLX sub-scales was taken into account. Results showed that overall workload was higher during the SOLUTION scenario. While the difference was not significant, there is still one important factor to consider regarding this difference. As mentioned above, an increase in mental/physical/temporal demand is usually accompanied with more effort, more frustration and less performance. The MINIMA concept however, induced a decrease in frustration and an increase in performance. While both results were desirable, they still put a bias on the overall workload scale. More precisely, the difference in overall workload would have been even higher, was it not diminished by that bias.

In conclusion, the NASA-TLX results show that controllers perceived the SOLUTION scenario to be more demanding, less frustrating and easier to achieve good performance in. Those are promising results in favour of the MINIMA concept.

5.3.2 Mind Wandering

The Out-of-the-Loop phenomenon is commonly associated with mind wandering episodes [see D1.1 [14]]. Therefore, it was expected that controllers would perceive mind wandering to be more frequent during episodes of low task involvement. In terms of the mind wandering sub-scales used in the DSSQ, this means that during the BASELINE scenario, controllers were expected to experience more task related interference (TRI), more task unrelated thought (TUT) and less task related thought (TRT). As a reminder, the latter (TRT) was not part of the original DSSQ but designed for this *Evaluation Study* to inquire to which extent controllers actually thought about their task.

As with the NASA-TLX results, the results of the mind wandering analysis were again promising. During the SOLUTION scenario, controllers reported to have experienced less TRI and less TUT compared to the BASELINE scenario. Therefore, controllers were less likely to be distracted by other matters beside their task. Regarding TRT, mean values of BASELINE and SOLUTION scenario were virtually equal. This indicates that controllers equally thought about things related to their task no matter how actively they were involved in it.

5.4 MINIMA's impact on controller performance

Interpretation of the controller performance indicator used for this *Evaluation Study* turned out to be complicated. As reported in 3.5.2, controllers got the impression that at times their advisories were not properly implemented into the simulation. Although analyses have shown that this was not the case, it remains unclear to which extent controllers reduced their advisory activity or stopped issuing advisories altogether. A reliable indicator of such behaviour could not be derived from the data, let alone one which could be linked to the controllers' subjective perception.

Despite the issues described above, it was decided to analyse performance as the portion of conflict-free simulation time by scenario and vigilance level. Results reported in section 4.3 show no significant differences in performance between scenario or vigilance levels. At first, this means that controller performance was equal among the four combinations. However, looking at the absolute means of all four combinations, performance was lowest during SOLUTION scenario when vigilance was high. This result is contrary to MINIMA's concept which aims to prevent such decrease in performance in the first place. However, due to the issues discussed above, it must be interpreted with great caution. First, although mean performance was lowest in this combination, it was still above 82 % and not significantly different from mean performance in any of the other combinations.

In conclusion, the results of the performance analysis should be interpreted with great caution. More likely, they should be completely omitted as it could not be determined to which extent controllers actually continued to interact with the system if they felt being unresponsive to their advisories. Therefore, there most probably was an important bias in at least part of the data which cannot be quantified. Such bias might have been compensated for by examining a larger sample. Given the small sample size examined in this study, it is likely that such bias had a great impact on the data.

5.5 Summary

In summary, the results of this *Evaluation Study* support the assumptions made within the MINIMA project. The core results can be summarised as follows.

1. **Controller vigilance decreased when they were not actively involved with the task but only acted as a mere monitor.**
2. **Without any changes to the automation system, vigilance would continue to decrease over time.**
3. **The decrease in vigilance was reversed through the Vigilance & Attention Controller.**
4. **Usage of the Vigilance and Attention Controller in the SOLUTION scenario caused**
    a. Controllers to recognise incoming aircraft earlier
    b. Controllers to more carefully process information
    c. Controllers to perceive their task to be:
        i. More demanding and effortful
        ii. Less frustrating and easier to perform good in
    d. Controllers to be less likely to get distracted by task related interference or task unrelated thought
5. **Usage of the Vigilance and Attention Controller in the SOLUTION scenario did not cause**
    a. Controllers to perform better or worse compared to the BASELINE scenario in terms of conflict-free simulation time
    b. Controllers to think more or less about task related matters

6 Conclusion

The results reported in this document predominantly show that the MINIMA concept worked as intended. First, it was shown that a continuous decrease in vigilance occurs when controllers are practically uninvolved with their task. Then, it was shown that such decrease could be anticipated and reversed using the *VAC* developed within MINIMA. Therefore, it can be concluded that the MINIMA concept of mitigating OOTL through an EEG-based adaptive automation system was successfully implemented.

In the future, as by its generic nature, the MINIMA concept could be implemented in various other domains other than ATC. For example, it could be used in cockpit environments where high level automation is also an important aspect, especially in case of system failure where pilots have to take over. Furthermore, the MINIMA concept can be applied to the automotive sector, where automated driving is a long-term goal of high public interest. Theoretically, the MINIMA concept can be applied to pretty much every domain in which human operators interact with highly automated systems. Of course, such applications will require some technical adjustments as the technical implementations used in MINIMA were especially designed for ATC. However, the EEG-based Vigilance Observer does not necessarily need to be used on ATCos, but can most likely be applied on human operators regardless of domain.

An important issue to address in future studies is to schedule more thorough training of the subjects to help them familiarise with the task environment. Such training could help prevent such misconceptions regarding advisory implementations as they were experienced by the subjects in this study. Actively guiding them through a standardised learning process could help make subjects more sensitive to changes within the system. This would help ensure that controllers have experienced actual changes to the system and how great and small they can appear at first.

In total the evaluation of the MINIMA concept can be considered successful. Conducted as planned and on schedule, it resulted in the collection of an invaluable set of expert subject data. This data strongly supports the assumption that the *VAC* is suitable to prevent human operators from getting OOTL in highly automated task environments. Especially with its broad applicability in other domains in mind, the MINIMA concept holds great potential for future research and, some day, operational usage.

Finally, the MINIMA project team would like to use the opportunity to express their gratitude to all the ATCos of the Italian air navigation service provider ENAV who voluntarily participated in this *Evaluation Study*. Without their help, the MINIMA project would not be where it is today.

7 References

[1] MINIMA (699282) – Grant agreement.

[2] MINIMA (699282) – D3.1 - Evaluation Plan, Ed. 00.02.00, 26 October 2017.

[3] MINIMA (699282) – D2.1 Integrated Vigilance and Attention Controller, Ed. 00.01.00, 31 July 2017.

[4] MINIMA (699282) – D2.2 Task Environment, Ed. 00.01.00, 31 July 2017.

[5] Ohneiser, O. (2010): RadarVision – Manual for Controllers. DLR (IB112-2010/54).

[6] Thackray, R. I., Jones, K. N., & Touchstone, R. M. (1974). Personality and physiological correlates of performance decrement on a monotonous task requiring sustained attention. *British Journal of Psychology*, 65(*3*), 351-358. doi: 10.1111/j.2044-8295.1974.tb01409.x

[7] Loh, S., Lamond, N., Dorrian, J., Roach, G., & Dawson, D. (2004). The validity of psychomotor vigilance tasks of less than 10-minute duration. *Behavior Research Methods, Instruments, & Computers*, 36(*2*), 339-346. doi: 10.3758/BF03195580

[8] Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews*, 29(*2-3*), 169-195. doi: 10.1016/S0165-0173(98)00056-3

[9] Aricò, P., Borghini, G., Di Flumeri, G., & Babiloni, F. (2017, March 22). EP 3143933 A1, BrainSigns Srl. Method for estimating a mental state, in particular a workload, and related apparatus.

[10] Matthews, G., Szalma, J., Panganiban, A. R., Neubauer, C., & Warm, J. S. (1999). Profiling Task Stress With The Dundee Stress State Questionnaire. In L. Cavalcanti & S. Azevedo: Psychology of Stress (pp. 49 - 91). New York City, USA: Nova Science Publishers, Inc.

[11] Smallwood, J., Fitzgerald, A., Miles, L. K., & Phillips, L. H. (2009). Shifting moods, wandering minds: negative moods lead the mind to wander. *Emotion*, 9 (*2*), 271-276. doi: 10.1037/a0014855

[12] NASA (1986). Nasa Task Load Index (TLX) v. 1.0 Manual.

[13] Helmke, H., Hann, R., Uebbing-Rumke, M., Müller, D., & Wittkowski, D. (2009): Time-Based Arrival Management for Dual Threshold Operation and Continuous Descent Approaches. In: Proceedings of the 8th USA/Europe ATM Research and Development Seminar.

[14] MINIMA (699282) – D1.1 State of the Art report, Ed. 00.02.00, 1 December 2016.

[15] Aricò, P., Borghini, G., Di Flumeri, G., Colosimo, A., Pozzi, S., & Babiloni, F. (2016). A passive brain–computer interface application for the mental workload assessment on professional air traffic controllers during realistic air traffic control tasks. *Progress in brain research, 228*, 295-328. doi: 10.1016/bs.pbr.2016.04.021

8 Annex

Annex A: Dundee Stress State Questionnaire

Table 11. Items of the Dundee Stress Test Questionnaire. Subjects were asked to rate the frequency of each thought on a 5-point Likert scale from 1 (Never) to 5 (Very Often).

Component	Item-Phrase
	During the last scenario, I thought about...
Task Related Interference	...how I should work more carefully
	...how much time I had left
	...how others have done on this task
	...the difficulty of the problems
	...my level of ability
	...the purpose of the experiment
	...how I would feel if I were told how I performed
	...how often I get confused
	During the last scenario, I thought about...
Task Unrelated Thought	...members of my family
	...something that made me feel guilty
	...personal worries
	...something that made me feel angry
	...something that happened earlier today
	...something that happened in the recent past
	...something that happened in the distant past
	...something that might happen in the future