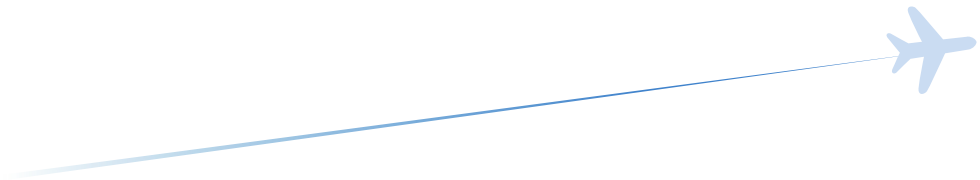

D3.2 - Principles for Transparency in AI/ML automation in ATM

Deliverable ID:	**D3.2**
Dissemination Level:	**PU**
Project Acronym:	**TAPAS**
Grant:	**892358**
Call:	**H2020-SESAR-2019-2**
Topic:	**SESAR-ER4-01-2019 Digitalisation and Automation principles for ATM**
Consortium Coordinator:	**CRIDA**
Edition Date:	**13 June 2022**
Edition:	**00.03.00**
Template Edition:	**02.00.02**

Founding Members

EUROPEAN UNION EUROCONTROL

sesar JOINT UNDERTAKING

Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
Antonio GRACIA / **BR&TE**	WP3 Leader	13/06/2022

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
Ian CROOK / **ISA Software**	Contributor	30/05/2022
Sandrine MOLTON / **ISA Software**	Contributor	30/05/2022
Natividad VALLE / **CRIDA**	Contributor	30/05/2022
María Florencia LEMA / **CRIDA**	Contributor	30/05/2022
Rubén RODRÍGUEZ / **CRIDA**	Project Coordinator Alternate	30/05/2022

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

Name/Beneficiary	Position/Title	Date
José Manuel Cordero / **CRIDA**	Project Coordinator	30/05/2022
Gennady Andrienko / **Fraunhofer**	Project Member	Silent approval
Hugo Salinas / **INDRA**	Project Member	Silent approval
Ian Crook / **ISA**	Project Member	30/05/2022
George Vouros / **UPRC**	Project Member	Silent approval

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
N/A		

Document History

Edition	Date	Status	Author	Justification
00.00.01	13/05/2022	Draft 1	A. Gracia	Principles for Transparency in AI/ML automation in ATM (Final)
00.00.02	24/05/2022	Draft 2	A. Gracia	Principles for Transparency in AI/ML automation in ATM (Final)
00.01.00	26/05/2022	Final	A. Gracia	Integration of Final Review Comments for Approval and SJU submission

00.02.00	30/05/2022	Final	A. Gracia	Integration of Final Review Comments for Approval and SJU submission
00.03.00	13/06/2022	Final	A. Gracia	Integration of SJU Review Comments for Approval

Copyright Statement

TAPAS

TOWARDS AN AUTOMATED AND EXPLAINABLE ATM SYSTEM

Abstract

This document presents a set of principles and recommendations for the use and application of transparent Artificial Intelligence (AI) and Machine Learning (ML) corresponding to the automation, at levels 2 and 3, of the Air Traffic Flow and Capacity Management (ATFCM) and Conflict Detection and Resolution (CD&R) operational use cases.

Table of Contents

Abstract 4

1 Executive Summary 9

2 Introduction 10

2.1 Purpose of the document 10

2.2 Intended Readership 10

2.3 Document Structure 11

2.4 Terminology and Acronyms 12

3 Deriving Principles for Transparency 14

3.1.1 Principles 14
3.1.2 Recommendations 17
3.1.3 Methodology 17
3.1.4 Recipients 19

4 Transparency Requirements Fulfilment 20

4.1 Assessment of the Degree of Fulfilment 21

4.2 ATFCM Transparency Requirements Fulfilment 22

4.2.1 Level 2 of Automation - Task Execution Support 22
4.2.1.1 Analysis of the imbalances detected - Identification of hotspots/optispots 22
4.2.1.2 Sector visualizer 22
4.2.1.3 Preparation of DCB measures to solve the hotspot 23
4.2.1.4 Selection of candidate flights 26
4.2.1.5 Analysis of the DCB measure impact – What-if 29
4.2.1.6 Hotspot Resolution monitoring 31
4.2.2 Level 3 of Automation – Conditional Automation 32
4.2.2.1 Declaration of hotspots / optispots 32
4.2.2.2 Decision on the DCB measure and flights impacted 33
4.2.2.3 Implementation of DCB measures 36
4.2.2.4 Other requirements 36
4.2.3 Transparency Requirements Refinement Proposal - ATFCM 38
4.2.4 Extracting Transparency Insights - ATFCM 40
4.2.4.1 Degree of transparency by ATFCM activity 40

4.3 CD&R Transparency Requirements Fulfilment 46

4.3.1 Level 2 of Automation - Task Execution Support 46
4.3.1.1 Assessment of the planned and desired trajectory profile 46
4.3.1.2 Identification of potential conflicts 48
4.3.1.3 Identification of conflict resolution strategies and clearances proposal 50
4.3.1.4 Conformance monitoring resolution 59
4.3.2 Level 3 of Automation – Conditional Automation 60
4.3.2.1 Clearances implementation 60
4.3.3 Transparency Requirements Refinement Proposal – CD&R 61
4.3.4 Extracting Transparency Insights – CD&R 63
4.3.4.1 Degree of transparency by CD&R activity 63

4.4 General fulfilment of transparency requirements 66
4.5 On the fulfilment of requirements and its impact on transparency and explainability 68
4.6 Learning to write transparency requirements in XAI 70
5 Expert Feedback Related to Transparency 72
5.1 Decisions on the General Supply of Transparency 72
5.1.1 Timing 73
5.1.2 Levels of Transparency 73
5.1.3 Combining Decisions to Unveil Further Insights 74
5.2 On the Impact of Transparency Quality on Human Efficiency 75
6 Extracting Insights from the Experiments 78
6.1 ATFCM experiments 78
6.1.1 Context 78
6.1.2 Validation Scenarios and Exercises 79
6.1.3 Validation Objectives and Results 81
6.1.4 Insights, Findings and Conclusions 82
6.1.4.1 Identifying principles for Transparency of AI-based solutions 82
6.1.4.2 Developing prototype XAI/VA methods for ATM use cases to address transparency at various levels of automation 84
6.2 CD&R experiments 89
6.2.1 Context 89
6.2.2 Validation Scenarios and Exercises 90
6.2.3 Validation Objectives and Results 93
6.2.4 Insights, Findings and Conclusions 94
6.2.4.1 Identifying principles for Transparency of AI-based solutions 94
6.2.4.2 Developing prototype XAI/VA methods for ATM use cases to address transparency at various levels of automation 97
6.3 General conclusions 103
6.3.1 Explainability/concept maturity 103
6.3.2 Transparency 104
6.3.3 Trust and confidence building 105
6.3.4 Technical feasibility 106
6.3.5 Efficiency (Key Performance Areas) 106
6.3.6 Safety (Key Performance Areas) 107
6.3.7 Human Performance (Key Performance Areas) 107
7 Principles and Recommendations for the Transparent Application of AI in ATM 108
7.1 Shaping Principles and Recommendations 109
7.1.1 From Insights to Principles and Recommendations 109
7.1.2 Format and Content 110
7.2 Principles 111
7.2.1 Transparency and Explainability 111
7.2.1.1 General Structure of Explanations 111
7.2.1.2 Level of Abstraction in Explanations 112
7.2.1.3 Off-line and On-line Transparency 113
7.2.1.4 Ongoing Actions Define Transparency Needs 114
7.2.1.5 Necessity of Transparency 115

7.2.1.6 Interfaces for Transparency 116
7.2.1.7 Features 118
7.2.1.8 Aggregating Information 119
7.2.1.9 Information Assimilation and Understanding 119
7.2.2 Human Factors - Trust 122
7.2.2.1 Trust Volatility 122
7.2.2.2 On Trust Building through Explanations 123
7.2.2.3 Trust Degradation by Unrealistic or Inaccurate Automations 124
7.2.2.4 'What-if' Mechanisms 125
7.2.3 Human Factors - Acceptance 126
7.2.3.1 Unbiasedness and Fairness 126
7.2.3.2 Closing the Gap Between Humans and XAI systems 127
7.2.4 Safety 129
7.2.4.1 Automation Levels and Situational Awareness 129
7.2.4.2 Intervention and Control Recovery in Automation 130
7.2.4.3 Safety at Partial and Full Automation 131
7.2.5 Certification 133
7.2.5.1 Tailoring Explanations for Certification 133

7.3 Recommendations 134
7.3.1.1 Seeking Accuracy in Transparency Requirements 134
7.3.1.2 Not Mixing Up Requirements Definition with Validation Aspects 134
7.3.1.3 Vocabulary Matters 135
7.3.1.4 Refining and Iterating on Transparency Requirements 135
7.3.1.5 Transparency Requirements Must be Ultimately Understood 136
7.3.1.6 Evaluation of the Level of Compliance 136
7.3.1.7 Automation Level and Transparency Requirements 137
7.3.1.8 Transparency, Explainability and Automation Recommendations 137

8 Conclusions 139

9 References 142

Appendix A - Schemas 143

Appendix B - Schemas 144

Appendix C – TAPAS Framework for the transparent use of AI/ML automation in ATM 145

List of Tables

Table 1. Proposal of refinement for the ATFCM transparency requirements. 39

Table 2. Proposal of refinement for the CD&R transparency requirements. 63

Table 2. Degree of fulfilment of the transparency requirements by the ATFCM prototype 66

Table 2. Degree of fulfilment of the transparency requirements by the CD&R prototype 67

List of Figures

Figure 1. Methodology used for the derivation of principles and recommendations for a transparent AI/ML automation in ATM..... 18

Figure 2. Transparency Requirements Fulfilment achieved by the ATFCM and CD&R prototypes..... 67

Figure 3. Levels of transparency proposed for the ATFCM and CD&R use cases 75

Figure 4. Illustration representing the flow of information in the derivation of principles and recommendations for the transparent application of AI in ATM..... 108

Figure 5. Format and content used in the definition of principles and recommendations. 110

Figure 6. Two major ways to tender transparency and explainability depending on the ongoing action. 113

Figure 7. Summary of the TAPAS ATFCM validation results per validation objective and criterion. Achieved sub-focuses are in green, partially achieved in orange, and non-achieved in red. 143

Figure 8. Summary of the TAPAS CD&R validation results per validation objective and criterion. Achieved sub-focuses are in green, partially achieved in orange, and non-achieved in red. 144

Figure 9. TAPAS framework for the transparent use of AI/ML automation in ATM. 145

1 Executive Summary

This document proposes a set of relevant principles and recommendations to be considered regarding the application of transparent and explainable Artificial Intelligence (AI) in the automation of Air Traffic Flow and Capacity Management (ATFCM) and Conflict Detection and Resolution (CD&R) use cases.

In particular, this document presents some innovative ideas and principles to be considered when automating, at different levels, ATFCM and CD&R use cases using technology based on Explainable AI (XAI) and Visual Analytics (VA). The idea is to be able to use the knowledge presented here and use it as a general guideline and foundations when carrying out the automation of these scenarios, in a more transparent and explainable way for different types of users and practitioners.

In general, these principles and recommendations contain revealing knowledge on how to implement a transparent and explainable automation process of the ATFCM activities and tasks, which are supposed to serve as reinforcement for conducting air traffic management tasks more fluently, efficiently, safely and robustly for the user. These principles and recommendations have been elaborated based on three important sources of knowledge gathered during the development of the TAPAS project: i) valuable feedback collected from different experts in *Air Traffic Management (ATM), human factors and AI* fields; ii) lessons learned through the development and implementation of the transparency requirements defined at the beginning of the project, which served as a guide and beacon for the technical implementation of the prototype; and iii) results from the validation activities of a set of expert operators when using and interacting with the prototypes developed for the ATFCM and CD&R use cases. From these three activities, multiple and diverse ideas, feedback, conclusions and insights were extracted which have served to elaborate and shape these principles and recommendations for transparency in AI/ML automation in ATM. In the document, these principles are organized into a set of general categories for the sake of clarity and to facilitate the reading.

The new knowledge presented in this document is intended to pave the way towards the uptake of AI/ML techniques as a step towards higher levels of Automation in accordance with the ATM Master Plan, as well as to serve as a general basis and a guideline in the application of transparent AI methods in ATM scenarios.

2 Introduction

2.1 Purpose of the document

The main objective of this document is to provide a set of principles and recommendations containing key knowledge on the application of transparent and explainable AI when automating activities across ATFCM and CD&R scenarios. This document considers automation levels 2 and 3, respectively partial and full automation, which are achieved by means of XAI and VA technology.

The principles and recommendations contained in the document are addressed to a wide target audience, such as *XAI developers*, *end users* and *operators*, *designers* to *certification agencies*.

To provide adequate background and allow the reader to understand why and in which situations principles have been used from a historical point of view, a review on concepts of 'principles' and their application to classic domains is provided. This is expected to help develop a basic intuition on why principles are used as a main transmitter of this new knowledge.

Also, the document presents the methodology that has been followed to collect key insights and findings over the project development, which have also served as major contributors to the definition of the principles and recommendations proposed here. The sources of information that are used as the main input to the elaboration of principles and recommendations are three, and these are properly presented, described, and explained to the reader. Thereafter, an extensive process of extraction of small pieces of information, or insights as we call them, is conducted. The idea is to identify, extract and elaborate the most promising ideas and findings to be later used as a main body of knowledge for the principles and recommendations. The document also poses relevant questions to be considered, carefully analysed, and hopefully answered in the near future.

The principles and recommendations are elaborated in the last section. They are classified according to different categories, depending on the main topic they belong to.

With this document, it is expected to provide the reader with a set of knowledge and tools that can be highly beneficial regarding how the automation of ATM scenarios using AI should be carried out in explainable and transparent manners, sometimes even daring to glimpse into the future. Ideally, this work should help lay the *foundations* to pave the way for the certification processes required for use and application of these technologies.

2.2 Intended Readership

This document is intended to be used by:

- SJU programme manager;
- TAPAS project members, in particular WP4 dealing with the XAI and Visual Analytics prototypes;

- SESAR2020 and international research community addressing automation in Air Traffic Management and Artificial Intelligence / Machine Learning.

2.3 Document Structure

This document is structured into the following sections:

- Section 1 is the Executive Summary and provides an overview of the general process followed to shape and define the principles and recommendations for the application of transparent and explainable AI in ATFCM and CD&R activities, the objectives of such principles and recommendations, as well as highlighting the categories in which they are organized.
- Section 2 presents an introduction that provides the purpose of the document, the intended readership, the document structure and the terminology and acronyms used throughout the document.
- Section 3 presents the general methodology followed to derive the transparency principles and recommendations presented at the end of the document. As part of this process, relevant background on historical concepts of 'principles' and their applications to classic domains is also provided, as well as a summary of the main recipients for which the principles and recommendations are meant to be.
- Section 4 presents the verification process followed regarding the level of compliance achieved by the transparency requirements after the implementation of the ATFCM and CD&R XAI/VA prototypes. In addition, relevant insights are provided during the process.
- Section 5 provides expert feedback and some decisions regarding the application of transparency and explainability aspects which were collected during the development of the TAPAS project, and therefore considered relevant in the elaboration of the principles and recommendations.
- Section 6 contains important insights and conclusions regarding transparency and explainability aspects. These are extracted from the validation activities corresponding to the ATFCM and CD&R experiments in which experts interacted with the developed XAI/VA prototypes.
- Section 7 presents the principles and recommendations for the transparent application of AI in ATM, organized in different categories, as well as some new ideas.
- Section 8 provides some conclusions and future lines.
- Section 9 provides the references used during the writing of this document.
- Appendixes A, B and C present some schemas used in the document.

2.4 Terminology and Acronyms

Term	Definition
3D	Three-Dimension
AI	Artificial Intelligence
AoR	Area of Responsibility
AoI	Area of Interest
ACC	Air Traffic Control Centre
ATC	Air Traffic Control
ATCo	Air Traffic Controller
ATFCM	Air Traffic Flow and Capacity Management
ATM	Air Traffic Management
CD&R	Conflict Detection and Resolution
CPA	Closest Point of Approach
CWP	Controller Working Position
DCB	Demand and Capacity Balancing
DL	Deep Learning
EC	European Commission
EC	Executive Controller
FMP	Flow Management Position
HEC	Hourly Entry Counts
HITL	Human-in-the Loop
KPI	Key Performance Indicator
KPA	Key Performance Areas
ML	Machine Learning
NM	Network Manager
NOP	Network Operations Plan

MDV	Multivariate Data Visualization
OCC	Occupancy Counts
PM	Projection Module
RTS	Real Time Simulations
SESAR	Single European Sky ATM Research
TAPAS	Towards an Automated and Explainable ATM System
TRL	Technology Readiness Level
VA	Visual Analytics
XAI	Explainable AI

3 Deriving Principles for Transparency

This section gives an overview and describes the main characteristic of the process followed to derive the transparency principles and recommendations in TAPAS.

Three sources of information have been considered to extract the insights that will shape the proposed principles and recommendations for transparency. These sources contain relevant knowledge collected over: i) the verification of the initial transparency requirements which allows understanding the level of technical compliance achieved by the prototype; ii) the results of the validation activities related to the performance of the ATFCM and CD&R prototypes, which collects valuable feedback from experts interacting with the actual implementations, and eventually extracts insights on the transparency and explainability achieved; iii) other valuable expert feedback provided throughout the development of the project. These three sources, in different degrees, serve as the main backbone from which to draw conclusions for the generation of new knowledge aimed at the real-word application of transparent and explainable AI in ATM. Therefore, and through this process, new knowledge regarding the use of transparent and explainable AI in ATM activities is inferred and extracted throughout this document by means of adequate '*principles*' and '*recommendations*'.

This section is organized as follows. A description of the main definitions used historically for the concept 'principle' and its applications to diverse classic domains is provided in Section 3.1.1. This is intended to allow the reader to outline and envision how principles might be presented, as well as to develop some initial intuition regarding the principles provided later in this document. The generated knowledge in this document is also presented in form of recommendations. Hence, Section 3.1.2 provides a description on the situations in which providing recommendations, instead of principles, might be more suitable. Section 3.1.3 3.1.3presents the main characteristics of the methodology followed to derive principles and recommendations. Finally, Section 3.1.4 describes the main recipients for such principles and recommendations.

3.1.1 Principles

According to 'Cambridge Dictionary' [12], the word '*principle*' means: 1) *a basic idea or rule that explains or controls how something happens or works*; and 2) *a moral rule or standard of good behavior or fair dealing*. These definitions are aligned with the ones provided by 'Oxford Languages' [13]:

1. *A fundamental truth or proposition that serves as the foundation for a system of belief or behaviour or for a chain of reasoning;*
2. *A general scientific theorem that has numerous special applications across a wide field;*
3. *A fundamental source or basis of something.*

The word 'principle', in its widest sense, refers to a base of ideals, foundations, rules and/or policies from which ideologies, theories, doctrines, religions and sciences are born. 'Principle' comes from the Latin 'Principium', which means origin or beginning. Despite being used to refer to the beginning of something, this word is mainly used in a moral and ethical philosophical sense. In addition to these

meanings, principles are also used to refer to foundations and/or laws about how a theory, ideology, doctrine, religion, justice, and science works.

In the scientific domain, which is probably the closest to the context of this research, *principles* are different to *laws* in several aspects. A law describes an event, but it does not explain why the event happens. Laws describe relationships, specific situations, and conditions. Principles describe why and how things happen. For example, Newton's law of gravitational attraction describes how objects are influenced by gravity. If you throw an apple in the air, it will follow a specific path while falling down. Newton's laws don't tell us why the apple falls or what causes it to fall, just that it does fall. In addition, scientific laws can be written as mathematical symbols and equations, for example:

- *Newton's Second Law*: F = ma;
- *Hooke's Law*: F = kx;
- *Gauss' Law*: $\nabla E=\rho$.

Principles, instead, have not the shape of rules that can be written down with mathematical symbols. They are guiding ideas that scientists use to make predictions and develop new laws. Principles are ideas based on scientific rules and laws that are generally accepted by scientists and they are fundamental truths that are the foundation for other studies. For instance:

i. Principle of Relativity: '*Physical laws take the same form in all systems of reference*';
ii. Principle of Special Relativity: '*The speed of light is the same for all observers*'.
iii. Pauli Exclusion Principle: '*No two particles with the same quantum numbers can be at the same position in space and time*'.

Digging further into the content of scientific principles, a well-known classic principle is the one called '*Archimedes' principle*'. Archimedes' principle is one of the most essential laws of physics and fluid mechanics. The principle states an object immersed in a fluid is buoyed up by a force equal to the weight of the fluid that it displaces. This principle, which is perhaps the most *fundamental* law in hydrostatics, explains many natural phenomena from both qualitative and quantitative points of view. For example, the principle of isostasy, which states that Earth's crust is in floating equilibrium with the denser mantle below [7] [8] , is simply based on Archimedes' principle. Or the *'Pascal's principle'*, also called Pascal's law [9], in fluid (gas or liquid) mechanics, statement that, in a fluid at rest in a closed container, a pressure change in one part is transmitted without loss to every portion of the fluid and to the walls of the container. The principle was first enunciated by the French scientist Blaise Pascal. '*Heisenberg's uncertainty principle*' is one of the most celebrated results of quantum mechanics and states, in essence, that there is inherent uncertainty in the act of measuring a variable of a particle, so one cannot know all things about a particle at the same time [11].

However, one of the most relevant examples in the definition of scientific principles is found in classical physics. One of the greatest exponents when defining principles in science was Sir Isaac Newton, in his work '*Philosophiæ naturalis principia mathematica*' [5] also known as *Principia*. First published in 1687, Isaac Newton provided a mathematical and comprehensive description of the laws of mechanics and gravitation and their application to planetary motion. The Principia is considered one of the most important works in the history of science [6], and a revealing example on how principles are defined and can be accurately used to model knowledge and produce predictive models.

Although less relevant to this research, the religious sense might be also interesting to mention. Here, *moral* principles are social norms that indicate what people should do or what they should avoid. They

also determine which actions should be promoted or recognized and which ones should be criticized or punished. These types of rules refer to general issues that can be applied in very different cases. They never refer to specific situations, therefore they can be interpreted and applied differently depending on the case. They come from the construction of human wisdom over time and are spread through time by oral tradition. Therefore, they are not compiled in any book or determined by a specific person. *Ethical* principles, on the other hand, reflect the 'appropriate' behaviour of people and the use of their specific knowledge in professional areas relevant to society (example: doctors). The moral principles together with the ethical principles make up what is called principles of the human being, and mostly have a religious meaning. As an example, the German philosopher, scientist, and writer Georg C. Lichtenberg defined four principles of morality [10]:

1) *Philosophical:* do good for its own sake, out of respect for the law.

2) *Religious*: do good because it is God's will, out of love of God.

3) *Human*: do good because it will promote your happiness, out of self-love.

4) *Political*: do good because it will promote the welfare of the society of which you are a part, out of love of society having regard to yourself.

All these examples of principles, in diverse domains, can be regarded as systematic attempts to model, infer and produce '*a body of knowledge*' that describes how humans, society and the world behave, as well as for defining useful rules and laws on how nature works. In an overall way, this inferred knowledge allows humans to make more informed decisions when facing and anticipating future situations. Similarly, the outcome of TAPAS is an attempt to provide a body of new knowledge to set foundations for the application of AI in ATM in a transparent and explainable manner, with the ultimate aim of building trust and reinforcing the human-machine paradigm.

The principles defined in some branches of human knowledge, as science, are born and built as a result derived from strong experimentation and validation processes. Likewise, all scientific research is based on a small set of assumptions. Although it is impossible to absolutely prove all these assumptions, they have been so widely proven and are so valid that they can be called scientific principles. Even though the TAPAS project has posed and conducted valuable experiments as well as proposed adequate validation processes of the interaction between human experts with the implemented prototypes, it is worth highlighting that it is not a main objective of TAPAS to establish principles that might be regarded as 'universal truths' in any case. The reasons are the exploratory research nature of the project, and the low TRL of the technology developed at this stage of the research. Thus, the principles derived in this research are not intended to be regarded as immutable truths, as they are highly susceptible to refinement and maturation processes as further and more sophisticated validation stages are implemented. In addition, XAI technology still requires a significant degree of maturation, improvement, testing and validation, and this is supposed to be addressed in the upcoming years.

However, it is certainly feasible to derive our own principles and recommendations from the results, insights, and feedback that we have collected over the project development. These, must be seen as a general way of encapsulating and presenting key knowledge that contains findings on aspects such as how the transparent automation of key activities are explained by the system and later interpreted by the human operator, how he/she interacts with and reacts to certain situations, elements of interest in ATM-related human decision making, or on the way that the information provided by the AI system must be organized, categorized and properly displayed/presented to the human.

These findings might help to pave the way and to establish the foundations on how the automation of some ATM scenarios, using transparent and explainable AI technology and VA tools, must be carried out.

3.1.2 Recommendations

In addition to establishing principles that contain knowledge on how explainable and transparent AI should be used in diverse ATM scenarios, TAPAS project also provides a series of more general sets of recommendations, practices, and indications. This can be especially useful in those cases where the knowledge and lessons learned are not yet firm, clear, or mature enough for directly becoming a principle, and therefore their application should remain optional. In other words, any insight derived from experimentation which might not yet have the entity for being considered a principle should probably lie on this category.

Therefore, providing a series of recommendations, indications or guidelines might suggest to different end users and practitioners aspects such as when, how, and even to what extent it is recommended to use of a type of transparency or explicability for the different casuistry found in the operational situations of the considered use cases.

3.1.3 Methodology

The methodology followed to derive principles and recommendations in the context of the TAPAS project is presented in Figure 1. As mentioned above, there are three main sources of information, from which the knowledge that will shape the principles and recommendations for the use of transparent and explainable AI in ATM is extracted. The overall process began with the definition of the transparency requirements.

The transparency requirements presented in [3] contained information on how, and to what extent, to impregnate with levels of transparency and provide adequate explicability to the underlying processes based on AI technology that are used to automate key activities in ATM. These requirements are the input, and serve as a technical guide, to the implementation of the prototypes for the ATFCM and CD&R use cases implemented by WP4. The prototypes make use of XAI and VA technology to try to comply with the previously set transparency requirements. Therefore, the main objective of the prototypes is to effectively achieve an automation of the ATM activities that will mean achieving automation levels 2 and 3, as defined in [2]. It is worth mentioning that the prototypes are not only limited to automating the corresponding activities to achieve levels 2 and 3, but it does so by providing explanations and insights about the logic used to generate certain outputs, providing information on key parameters that influence the inner mechanisms of the models, as well as making use of appropriate visualizations when necessary.

Once the implementation of the prototypes was done, validation activities of the experiments proposed for the different scenarios were designed, proposed and carried out during the month of June 2021 (for the ATFCM use case) and March 2022 (CD&R) [4]. To do so, several Human-in-the Loop (HITL) simulations were performed which involved operational experts in order to validate and extract conclusions and principles on the transparency and explainability requirements necessary for three

main scenarios at different levels of automation (level 1, 2 and 3). From these validation activities, it is also possible to check the level of *fulfilment* of the transparency requirements defined at the beginning. The degree of fulfilment might indicate to what extent the current XAI and VA technology is mature enough to capture the initial requirements and turn them into real functionalities provided by the prototypes. Also, it is interesting to understand and quantify our own limitations, regardless the current and available technology, when it comes to defining general requirements for transparency and set ways for further refinement and improving of these requirements. Such lessons shall also contribute to build a body of knowledge from which to extract ideas and insights for defining principles and recommendations. In addition, extremely valuable knowledge and feedback have been collected from highly recognized experts working in different domains such as ATM, AI, and human factors, which are undoubtedly supposed to contribute to a deep well of insights from which to outline principles and recommendations on the use of transparent and explainable AI.

The common factor across the document is the extraction of key *insights*. New knowledge is built by means of these insights. Insights are represented by bullet points with a phi symbol (Φ) and they can be seen as *small* or *basic* pieces of information containing relevant knowledge and findings related to transparency and explainability. These insights are collected, classified, and categorized according to their main topic. Such insights are combined into bigger and more elaborated pieces of information, and further developed according to their category. This way, when we need to shape the principles and recommendations for the different categories, some of these pieces of information will be already available to build upon them. This will significantly facilitate the process of shaping principles and recommendations.

As mentioned earlier, the main objective of the research carried out in TAPAS project is to shape and derive a set of principles, recommendations, and foundations in the form of new knowledge, which can serve as a general basis and a guideline in the application of transparent AI methods in the ATM domain.

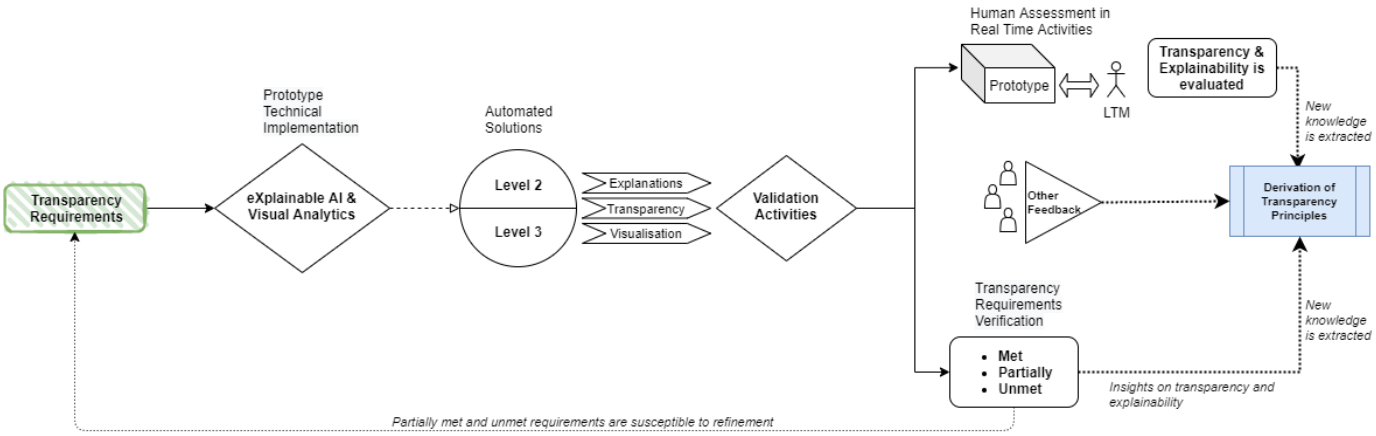

Figure 1. Methodology used for the derivation of principles and recommendations for a transparent AI/ML automation in ATM

3.1.4 Recipients

The principles and recommendations proposed in this document are devoted to a set of different target audiences. These, range *from XAI developers*, *end users* and *operators*, *designers* to *certification agencies*.

4 Transparency Requirements Fulfilment

This section is devoted to the verification of the level of compliance achieved by the transparency requirements defined in document '3.1 – Use Cases Transparency Requirements' [3], after implementing the prototype that automates the ATFCM and CD&R use cases and conducting the validation activities.

The process is based on a complete traceability that both validates the level of compliance that each solution has with respect to its own requirements and a description of the degree of transparency and explainability aspects achieved per activity of the ATFCM and CD&R use cases.

The core idea is that the lessons learned during this process shall allow us to foresee, outline and eventually define potential principles and recommendations regarding the use of transparent AI, whilst basing them on clear and well justified criteria and rationale.

TAPAS, as an exploratory research project, has an incremental nature which means that over the development of the project we have made incremental additions and improvements as it was progressing. This results in the fact that the processes followed for assessing the requirements fulfilment, for the ATFCM and CD&R use cases, are subtly different in the following ways.

For the ATFCM prototype, completed in *June 2021*, the initial assessment of compliance with the requirements is based on the process of going through them and, carrying out a subjective evaluation, to verify whether the technical solutions and functionalities provided by the prototype allow the requirements to be successfully met, partially met, or unmet. These decisions are duly supported by elaborated rationale and justified reasons. Several members of the WPs who participated in the implementation of the prototype and validation activities are involved in this evaluation process. Hence, leaders from WP02, WP04 and WP05 were asked to provide sincere feedback regarding the degree of fulfilment of the transparency requirements, from their point of view, as well as proper rationale supporting their opinions. Based on this feedback, the work conducted by the WP03 leader was to try to harmonize the provided opinions in order to reach a coherent degree of agreement. This process was iteratively conducted over several rounds until reaching a final decision on the fulfilment of the transparency requirements. Along with the final decision of the degree of fulfilment, other aspects are also provided. For instance, it is important to provide a justification attempting to accurately synthesize why a particular decision is reached. Some other times, the assessment of a given requirement led us to the conclusion that a refinement of the requirement might be needed. The reason for this could simply be that the requirement was not properly defined, it was ambiguous somehow, or even that the requirement is not relevant in terms of transparency/explainability and somehow superfluous. Therefore, some transparency requirements are reworded here for the sake of interpretability.

Sometimes, remarkable insights on the use of transparency and explainability are also extracted. These are derived from the expert feedback collected during the validation activities of the prototype. For some others, these are personal interpretations of this document's author. In any case, these insights are supposed to be a direct source of contribution to the principles and recommendations on the use of transparent AI proposed at the end of this document.

The CD&R prototype was implemented and completed a few months later, in *March 2022*. This gap in time allowed us to account for lessons learned during the first round of validation and to think of and carefully elaborate, more efficient and objective ways for analysing the fulfilment of the transparency requirements achieved by the prototype for CD&R use case. In particular, it was agreed that defining a set of 'compliance criteria', in advance, would likely facilitate the process of verifying the fulfilment of each of the transparency requirements for the CD&R use case. In this way, and in order to check more straightforwardly if a transparency requirement is met or not by the implemented prototype, we could simply verify if the prototype meets some pre-defined (and agreed) criteria. This would eventually benefit the objectivity of the process, as the final decision of fulfilment would not only rely on a human's assessment, but on well-defined criteria, values and parameters. Therefore, for the CD&R use case, a fourth column called 'Functional Compliance Criteria' is included, which allowed us to assess the requirements fulfilment in a clearer, simpler and more efficient manner.

Finally, it is worth highlighting that, in contrast to the ATFCM case in which the requirements fulfilment was assessed after the experiments, in the CD&R experiments the transparency requirement fulfilment was assessed in advance, and before the days of the final experiments.

4.1 Assessment of the Degree of Fulfilment

The evaluation of the fulfilment of the requirements is conducted as follows. The requirements are organized by level of automation and task. A description of the requirement is given, while complementing it with its corresponding decision, a justification and possible insights (if any) on transparency and explainability that are derived in the process.

The *final decision* is set by means of four degrees of fulfilment, depending on the level of compliance:

- **Unmet** requirement. A requirement that is unmet is represented in red. This is given when the functionality specified by the requirement is not fulfilled in any degree.
- **Partially met** requirement. A requirement can be partially met by two different degrees: **-** (orange) and **+** (light green).
    - Negative (-). Indicates a requirement that is partially met, but not in a sufficient way and therefore closer to be unmet. A few characteristics are implemented.
    - Positive (+). Indicates a requirement that is partially met, and very close to being fully met. An example is given when a requirement is almost fully met, but some minor functionalities have been not finally implemented, preventing it to be considered as fully met.
- **Met** requirement. A requirement that is fully met is represented in green. This is given when all the major functionalities requested by the requirement are fulfilled.

In addition, an adequate *justification* is provided along with the decision. This is intended to support the decision by means of comments and feedback, as well as providing feedback on the degree of agreement on the reached decision. Some other times this will describe why a requirement is not met, or why it is considered partially met by using proper rationale.

Finally, different *insights* can be sometimes extracted from the feedback given by experts during the validation exercises or by the personal interpretation of this document's author. These are supposed to eventually contribute in some degree to the recommendations and principles proposed in Section 7.

4.2 ATFCM Transparency Requirements Fulfilment

4.2.1 Level 2 of Automation - Task Execution Support

4.2.1.1 Analysis of the imbalances detected - Identification of hotspots/optispots

- **REQ-ATFCM-T001-BD.** The FMP shall be provided with the set of decision rules that led to the identification of a hotspot or optispot.

Decision	Justification	Insights
MET	Overall Agreement. Rules for the identification of hotspots are deterministic and very simple in the present form of the prototype.	

- **REQ-ATFCM-T002-BD.** Should a decision tree be built as part of the rule-based system, the FMP shall be provided the complete decision tree containing the rules regarding the identification of a hotspot or optispot.

Decision	Justification	Insights
PARTIALLY +	Full Agreement. The FMP was provided with explanations on the identified hotspot, but the cause of the hotspot was not always clear. Finally, no decision tree existed for this purpose.	

4.2.1.2 Sector visualizer

- **REQ-ATFCM-V003-BD**. In order to alleviate the congestion of sectors that are cluttered and overloaded with multiple individual aircraft trajectories, the FMP should be provided with effective visualization means in order to make the visualization clearer and more understandable.

Decision	Justification	Insights
MET	Overall Agreement	

- **REQ-ATFCM-V004-BD**. Trajectories should be clustered by a proper similarity measure and, depending on different density (number of trajectories) thresholds, a group of similar trajectories will be represented by one single, wider, and bigger trajectory flow.

Decision	Justification	Insights
UNMET	The requirement is considered unmet as it has not been included into the demonstrator as it was originally defined. However, a more appropriate technique is provided. Although tools for clustering of	

	trajectories based on a variety of similarity measures are available, we found these methods to be hardly applicable in the DCB scenario. Instead, grouping of trajectories that cross similar sequences of sectors is provided within the sector explorer component. In this tool, the grouped flight trajectories are presented in combination with time histograms displaying demand in involved sectors.	

4.2.1.3 Preparation of DCB measures to solve the hotspot

- **REQ-ATFCM-T005-BD**. The FMP shall be provided with the most effective DCB measure (or combination of DCB measures) to solve the hotspot.

Decision	Justification	Insights
MET	Full Agreement	

- **REQ-ATFCM-T006-BD**. The FMP shall be provided with the most relevant output parameters that influence the most in determining the proposed DCB measure (or combination of DCB measures).

Decision	Justification	Insights
MET	From the technical side it is met. XAI provides, for any single decision taken, all parameters that influence the decision	During the debriefing sessions the FMP experts stated that some **aggregated information, statistics about solutions impact would be of interest**, apart from the information already included in the tool

- **REQ-ATFCM-T007-BD**. Should any kind of importance be given to some input parameters used in the determination of the DCB measures, this information shall be explicitly provided to the FMP in form of explanations.

Decision	Justification	Insights
PARTIALLY +	The XAI prototype provides all parameters that affect a decision, but an explanation is not provided as required	

- **REQ-ATFCM-V008-OD**. Additionally to REQ-ATFCM-T005-BD, the FMP shall be provided with a preliminary visualization on the display (s) of alternative and backup solutions representing other ‘n’ most effective DCB measures to solve the hotspot.

Decision	Justification	Insights
PARTIALLY -	Internally, multiple types of alternative solutions (i.e., with different mixtures of measures) are considered, but only one solution is finally provided to the FMP via VA. Three types of solutions and one solution per type are given. For instance, one solution using regulations, one solution using level capping and one using level capping regulations. In Level 2 the aim of the requirement was to provide several backup and promising solutions for the FMP to decide based on his/her judgment and experience.	Further interaction of the users with the system would probably enforce the need for multiple solutions per type of solution.

- **REQ-ATFCM-V009-OD**. This preliminary visualization should include a motion-based representation of the specific action to be taken (e.g., level-capping or a delay), complemented with a textual display containing the following parameters:
  - A metric that measures the *effectiveness* of that DCB measure to solve the hotspot
  - On-load Areas
  - Off-load Areas
  - Flights involved and related flight information

Decision	Justification	Insights
MET	The most important functionality of this requirement is met by using visualisation means; except On-load Areas and Off-load, that are not supported	During the debriefing sessions the FMP experts stated that some **aggregated information, statistics about solutions impact would be of interest**, apart from the information already included in the tool

✓ **REQ-ATFCM-T010-BD**. The FMP shall be provided with explanations regarding the inner reasons of proposing some DCB measures over others.

Decision	Justification	Insights
MET	From the technical side it is met: the system provides all the parameters that affect agent's decisions, also providing the parameters for counter-decisions. The prototype describes the reasons for a specific measure decided. In addition, it provides the arguments that justify not to take that measure. The difference between them could be exploited so view the more important reasons for making a specific decision for a measure, but this is not presented.	✓ Information about explanations was presented but **hard to understand** by FMP experts ✓ The quality of explanations must be improved

✓ **REQ-ATFCM-T011-BD**. The FMP shall be provided with explanations containing detailed information on how the proposed DCB measure (or combination of DCB measures) impact (s) on other sectors.

Decision	Justification	Insights
MET	It is met	✓ Information about explanations was presented but **hard to understand** by FMP experts ✓ The quality of explanations must be improved

✓ **REQ-ATFCM-T012-BD**. The FMP shall be provided with a list of the most impacted sectors by the proposed DCB measure (or combination of DCB measures).

Decision	Justification	Insights
UNMET	Overall agreement	

✓ **REQ-ATFCM-T013-BD**. The FMP shall be provided with explanations and parameters in order to reinforce the trust of the FMP in the AI system in potential counterintuitive situations (to be identified during the validation phase) for the FMP.

Decision	Justification	Insights
MET	From the technical side, it is met	In general, the **requirements should not contain references to aspects that can be acquired during validation**, for example 'trust'. If the system foster trust, this must be tested later, during the validation. Also, during the validation it was highlighted that the FMP showed an improvement of trust in the system after several days. More time for getting familiarized with the system **would have been useful**

4.2.1.4 Selection of candidate flights

- **REQ-ATFCM-T014-BD**. The FMP shall be provided with explanations for her/him to understand the details of the inner process followed during the trade-off process.

Decision	Justification	Insights
MET	This requirement originally contained (probably wrongly) two aspects: technical and interpretability. The first part: 'providing explanations' is met. The second part: 'for the FMP to understand the details', is not met. However, the requirement is considered met from a technical point of view, even though it is not proved that the 'FMP understand the details of the inner process', as this must be tested in the validation activities, and it is a lesson learned for the future definition of transparency requirements.	▪ It is **not recommended mixing** technical and other aspects in a single requirement. The difficulty in the interpretations (or the irrelevance of the requirement) by humans is a thing that should not affect to the evaluation of the technical fulfilment. This insight must be part of the good practices or recommendations on an 'accurate definition of general transparency requirements'. ▪ This requirement was considered as **not relevant** to the FMP, who stated that they do not need such information. The FMP was provided with explanations on the different steps the algorithm uses to calculate the final solution. But since the algorithm was complex, it was difficult for them to understand and, in particular, they stated they did not need that information during the operation. In fact, only during training it will be useful to have it. ▪ Also, the term 'understandable' should be more explicit. Finding **more accurate ways** to better define what is understandable and what it is not is primordial and must be tackled.

- **REQ-ATFCM-T015-BD**. The FMP shall be provided with all the parameters for her/him to understand the details of the inner process followed during the trade-off process.

Decision	Justification	Insights
MET	Same than **REQ-ATFCM-T014-BD**	Same than **REQ-ATFCM-T014-BD**

- **REQ-ATFCM-T016-BD**. The FMP shall be provided with a list of the candidate flights or flows impacted by the DCB measure (or combination of proposed DCB measures), along with explanations informing about the key parameters involved in the decision for the selection.

Decision	Justification	Insights
MET	Full agreement. It was stated that 'The XAI does not finally work in that way, but we can conclude that the	

	requirement is met, since measures per flight are explained.'	

- **REQ-ATFCM-T017-BD**. The FMP shall be provided with several alternative solutions to the candidate flights or flows impacted by the DCB measure (or combination of proposed DCB measures), along with explanations informing about the key parameters involved in the decision for the selection.

Decision	Justification	Insights
PARTIALLY -	Internally, multiple types of alternative solutions (i.e., with different mixtures of measures) are considered, but only one solution is finally provided to the FMP. Three types of solutions and one solution per type are given. For instance, one solution using regulations, one solution using level capping and one using level capping regulations. In Level 2 the aim of the requirement was to provide several backup and promising solutions for the FMP to decide based on his/her judgment and experience. There were different interpretations of the requirements by the WP leaders. Specifically, in the understanding of what 'alternative or backup solutions' meant to them.	▪ The aforementioned different interpretations might indicate the **need for a possible rewording** or a more accurate definition of some of the requirements. ▪ Univocally defining general requirements (and accurately interpreting and technically implementing them by the recipient) that do not involve specific quantities, measures or values (for example as typically done in system engineering) is not trivial. Robust mechanisms to **diminish the ambiguity** in the understanding of these might be required in the specific case of requirements involving transparency or explainability aspects.

- **REQ-ATFCM-T018-BD**. The FMP shall be provided with the expected benefits obtained, in the sector and other possible sectors, by the different proposed solutions.

Decision	Justification	Insights
PARTIALLY +	Full agreement	During the debriefing sessions the FMP experts stated that some **aggregated information, statistics about solutions impact would be of interest**, apart from the information already included in the tool

- **REQ-ATFCM-T019-BD**. The FMP shall be provided with relevant values informing about the total number of affected flights for the different proposed solutions.

Decision	Justification	Insights

MET	Full agreement	

- **REQ-ATFCM-T020-BD**. The FMP shall be provided with relevant values informing about how long those flights will be affected by the DCB measure (or combination of proposed DCB measures) for the different solutions.

Decision	Justification	Insights
MET	Full agreement	

- **REQ-ATFCM-T021-BD**. The FMP shall be provided with relevant values informing about how many different airlines are impacted in the different solutions.

Decision	Justification	Insights
UNMET	Overall Agreement. The requirement is not met because the information is not explicitly shown by the prototype, which does not consider airlines. This information could be extracted from the tools but it was not shown directly.	

- **REQ-ATFCM-T022-BD**. Should there be any other type of valuable information that can help the FMP make accurate comparisons between the different solutions to make final choices, this information shall also be identified and provided by the system.
  - ✓ For example, *the FMP shall count on accurate and up-to-date information to avoid the possible situation of causing harm to some airlines by implementing DCB measures on a large number of aircraft from a specific airline, instead of first prioritizing fairer and more balanced measures*.

Decision	Justification	Insights
MET	It is considered met. The FMP was provided with a 'what-if' functionality that allow them to test the solutions proposed by the prototype, in Level 2. The system provides decision rules triggered for decision making, as well as arguments for taking counter-decisions. Tools for running the system simulation towards producing the final solution are provided, also allowing the visualization of all involved flights and joint decision making.	

4.2.1.5 Analysis of the DCB measure impact – What-if

- **REQ-ATFCM-T023-BD**. The FMP shall be provided with a PM that allows the implementation of hypothetical DCB measures (or combination of DCB measures) in order to project, beforehand, their consequences, as well as analysing and assessing the expected impact on the network and potential benefits obtained, in the sector and other possible sectors, by the DCB measure.

Decision	Justification	Insights
MET	It is met. What is finally presented per decision includes: (a) the parameters that affect the decision taken (b) the parameters that drive alternative decisions The impact of alternative decisions is provided by the XAI system. However, no interaction exists with operators for identifying the effects of hypothetical DCB measures.	

- **REQ-ATFCM-V024-BD**. The PM shall provide a suitable visualization depicting the expected effect of the implemented DCB measure so that the FMP can quickly and intuitively comprehend the possible consequences of his/her actions in the resolution of the active hotspot (s).

Decision	Justification	Insights
MET	From a strict technical point of view this must be considered met. The system provides the functionality even though the FMP 'can **NOT** quickly and intuitively comprehend the possible consequences of his/her actions in the resolution of the active hotspot (s)'. Mixing technical and interpretability aspects in a requirement is not correct, and it must be merely assessed from a technical side. This certainly indicates an **inaccuracy in the definition of the requirement** and needs rewording.	It should not be stated nor assumed (in the definition of the requirement itself) how the recipient of the requirement will eventually interact or feel respect to it. Drawing conclusions like this **must only come from the validation** results, not from the requirements definition.

- **REQ-ATFCM-T025-BD**. In the case of not solving the hotspot, the PM shall provide explanations informing about the underlying reasons of the failure in solving the spot.

Decision	Justification	Insights
UNMET	Overall agreement. No explanations informing about the underlying reasons for unresolved hotspots are provided. Information about a non-resolved hotspot	

	can provide reasons for failure, but no direct explanations are provided.	

- **REQ-ATFCM-T026-BD**. In the case of not solving the hotspot, the PM shall provide explanations suggesting alternative DCB measures which might help solve the hotspot.

Decision	Justification	Insights
UNMET	Overall agreement. In case of not solving the hotspot, no explanations containing and suggesting alternative DCB measures which might help solve the hotspot are provided. Potential solutions are provided, but in cases they cannot solve it, no explanations on failure are provided	

- **REQ-ATFCM-T027-BD**. The PM shall provide a list of the most important parameters involved in the resolution state of the hotspot.

Decision	Justification	Insights
MET	Full Agreement. The XAI prototype provides for any single decision taken all parameters that influence the decision. However, the decision of an agent does not consider a single hotspot, but all the hotspots to which it is involved during its flight	

- **REQ-ATFCM-T028-BD**. The PM shall allow setting up different parameters (e.g., hyperparameter tuning) to be used as input to the ML algorithms, and re-running simulations in order to evaluate the diversity of the possible solutions, as well as their potential impact on the network.

Decision	Justification	Insights
UNMET	Overall agreement. The parameters were fixed without offering the possibility to change them	

- **REQ-ATFCM-T029-BD**. In the case of implementing ATFCM regulations involving specific Regulation Period, Regulation Width and Regulation Rate, the PM shall provide the FMP with the following parameters so that he/she can understand the real magnitude of the expected impact and make pertinent decisions.
  - Occupancy Counts
  - Hourly Entry Counts
  - Total delay
  - Average delay per flight
  - Estimated additional flown distance

Decision	Justification	Insights
MET	It is considered met, but no estimated additional flown distance is provided.	The requirement must avoid explicitly mentioning aspects related to understanding. Technical aspects cannot be mixed with interpretability aspects, which are assessed later during the validation.

4.2.1.6 Hotspot Resolution monitoring

- **REQ-ATFCM-T030-BD**. The HRMM shall provide the FMP with up-to-date and relevant parameters in real-time to help the FMP univocally understand what the current resolution state of all the active hotspots is (e.g., number of flights over the declared capacity).

Decision	Justification	Insights
MET	This is met from a technical point of view, even though **it was not proved** that the FMP univocally understands what the current resolution state of all the active hotspots.	The requirement is ambiguously defined and requires rewording. It mistakenly mixes technical and interpretability aspects. As mentioned earlier, conclusions on interpretability: '...to help the FMP **univocally understand** what the current resolution state of all the active hotspots is', must be drawn from validation and so it **must be avoided** in the general definition of transparency requirements.

- **REQ-ATFCM-T031-BD**. Should the HRMM detects potential deviations, a list possible corrective actions shall be provided to the FMP for further assessment.

Decision	Justification	Insights
UNMET	Overall agreement. Deviations cannot be detected. However, the system can proceed the simulation from any state	

- **REQ-ATFCM-T032-BD**. Should the resolution of a hotspot be not resolved within D-1, the HRMM shall provide accurate and explanations stating the underlying reason (s) of why it has not been resolved.

Decision	Justification	Insights

UNMET	Overall agreement	

- **REQ-ATFCM-T033-BD**. Should the resolution of a hotspot be not resolved within D-1, the HRMM shall textually suggest initiate action(s) for preparation of alternative DCB measures (or combination of DCB measures) in order to solve the hotspot, taking into account the current situation of the network.

Decision	Justification	Insights
PARTIALLY -	This is partially met. Alternatives are provided, but not in the same way that the requirement originally states	

4.2.2 Level 3 of Automation – Conditional Automation

4.2.2.1 Declaration of hotspots / optispots

- **REQ-ATFCM-T034-BD**. The FMP shall be provided with the set of decision rules that led to the declaration, or not, of the hotspot to the NM.

Decision	Justification	Insights
MET	Full Agreement. XAI does not focus on individual hotspots, but on the global airspace state	

- **REQ-ATFCM-T035-BD**. Should a decision tree be built as part of the rule-based system, the FMP shall be provided with the complete decision tree containing the rules regarding the declaration, or not, of a hotspot to the NM.

Decision	Justification	Insights
MET	Overall agreement. No decision tree is finally built, but the decision process was able to be followed from the graphic display	

- **REQ-ATFCM-V036-BD**. In order to complement the information contained in the decision rules, the FMP should be provided with enhanced visualization means in order to better understand the analysed hotspot's root cause and complexity. The visualization shall represent key information in relation to complexity and cause factors, such as:
  - trajectories
  - impacts with other sectors
  - number of flights climbing/descending
  - number of potential trajectory crossings
  - number of flows interactions

- PRU Complexity Index
- Others

Decision	Justification	Insights
MET	Overall agreement. No information was provided related to flow interactions or complexity	

- **REQ-ATFCM-V037-BD**. This information shall be displayed into a map and located where required by using diverse VA tools, along with their numerical values (in form of text). This shall help the FMP understand in a more analytical objective and transparent way the decision of the AI system to declare, or not, the hotspot to the NM.

Decision	Justification	Insights
MET	Overall agreement. As a map representation may be overcrowded and often confusing due to 4D nature of data (dynamics over 3D space), appropriate representation for flights in a form of sequences of crossed sectors was designed. This representation in an interactive and dynamic form is provided within the sector explorer component	

- **REQ-ATFCM-T038-BD**. Should a set of additional parameters (e.g., those that have not been used in the building of the rule-based system for the identification of hotspots/optispots and might possibly have some relevance in the decision of declaring a hotspot to the NM. For example, parameters related to overall network load, or period of the year, etc.) have been used in the decision of the declaration, they will be provided to the FMP so that he/she can clearly understand the final decision.

Decision	Justification	Insights
UNMET	Full Agreement	The requirement must avoid explicitly mentioning aspects related to understanding. Technical aspects cannot be mixed with interpretability aspects, which are assessed later during the validation.

4.2.2.2 Decision on the DCB measure and flights impacted

- **REQ-ATFCM-T039-BD**. The FMP shall be provided with explanations containing the inner reason (s) about which DCB measure (or combination of DCB measures) has (have) been selected for implementation.

Decision	Justification	Insights
MET	Full Agreement	

- **REQ-ATFCM-T040-BD**. The FMP shall be provided with the set of the parameters that influenced the most in the final decision.

Decision	Justification	Insights
MET	Full Agreement	

- **REQ-ATFCM-T041-BD**. Should the AI system identified more than one DCB measure that are very similar in likelihood in terms of solving the hotspot (and so both are highly susceptible for being selected for implementation), the FMP shall be provided with explanations highlighting the differences that led it made the decision.

Decision	Justification	Insights
MET	It is considered met. Differences among sets of parameters considered for alternative decisions are provided. In addition, explanations for different types of solutions are provided and these can be further elaborated	

- **REQ-ATFCM-T042-BD**. The FMP shall be provided with a list of the flights impacted by the final selected DCB measure (or combination of DCB measures).

Decision	Justification	Insights
MET	Full Agreement	

- **REQ-ATFCM-T043-BD**. The FMP shall be provided with information regarding how long the flights are to be impacted by the selected DCB measure (or combination of DCB measures).

Decision	Justification	Insights
MET	Full Agreement	

- **REQ-ATFCM-T044-BD**. The FMP shall be provided with a list of the most impacted airline (s). ('Impacted' means those airlines whose aircraft are directly affected by the selected DCB measure, or combination of DCB measures).

Decision	Justification	Insights
UNMET	Full Agreement. A list on impacted airlines was not provided. Explanations were provided at a flight level, without differentiating the airline. The system does not consider information about airlines	

- **REQ-ATFCM-T045-BD**. The FMP shall be provided with the most relevant parameters that influence the most in determining the ordering of the airline (s) ranking.

Decision	Justification	Insights
UNMET	Full Agreement. The system does not consider information about airlines. Ranking is not available	

- **REQ-ATFCM-T046-BD**. The FMP shall be provided with a list of the most impacted sector (s) by the selected DCB measure (or combination of DCB measures).

Decision	Justification	Insights
MET	Overall agreement. The most impacted sectors by the selected DCB measures are provided by displaying them on a graph, rather than on a list	

- **REQ-ATFCM-T047-BD**. The FMP shall be provided with the most relevant parameters regarding the decision of which sectors are the most impacted by the selected DCB measure (or combination of DCB measures).

Decision	Justification	Insights
MET	Overall agreement	

- **REQ-ATFCM-T048-BD**. The FMP shall be provided with explanations to the question: '*How the selected DCB measure (or combination of DCB measures) can impact or create new and unexpected future possible hotspots*?

Decision	Justification	Insights
PARTIALLY -	Partially met. Even though the 'what-if' functionality for Level 2 allows to assess the impact of the solutions, **no explicit and clear explanations are provided** which allow getting insight on the creation of new possible hotspots.	

	The effects of joint decisions on DCB measures can be viewed, although this is quite complex given that the joint decision of all agents involved must be considered	

4.2.2.3 Implementation of DCB measures

- **REQ-ATFCM-T049-BD**. The FMP shall be provided with an explanation confirming that the DCB measure (or combination of DCB measures) has (have) been successfully implemented in the system, as well as providing the flights impacted register.

Decision	Justification	Insights
MET	Full Agreement	

- **REQ-ATFCM-T050-BD**. The FMP shall be provided with explanations informing about the reason (s) about why the implementation of the DCB measure (or combination of DCB measures) in the system has failed, as well as providing insight on further actions to take.

Decision	Justification	Insights
UNMET	Full Agreement. Unresolved hotspots are not justified. If a DCB measure fails, there **is no explanation on what fails** nor further insights	

4.2.2.4 Other requirements

- **REQ-ATFCM-T051-BD**. The FMP shall be provided with explanations regarding what is/are the ATFCM regulation(s) applied to any single flight.

Decision	Justification	Insights
MET	Full Agreement	

- **REQ-ATFCM-T052-BD**. The FMP shall be provided with explanations regarding what is the joint ATFCM regulation(s) being applied to a subset of flights (e.g., those crossing a specific sector within a period).

Decision	Justification	Insights
MET	Overall Agreement. The main functionality (explanations) is provided, regardless how difficult is to conduct actions as tracking the aggregation	

- **REQ-ATFCM-T053-BD**. The FMP shall be provided with explanations regarding what are the ATFCM regulations on flights satisfying specific criteria.

Decision	Justification	Insights
UNMET	No such possibility is provided	

- **REQ-ATFCM-T054-BD**. The FMP shall be provided with explanations regarding what are the flights that can be mostly affected by hotspots.

Decision	Justification	Insights
PARTIALLY +	It is partially met (+). No specific functionality providing explanations identifying the flights that are 'most affected' by hotspots is found. However, flights in a hotspot were able to be found, and explanations' parameters include the hotspots that justify a DCB measure for a flight	

- **REQ-ATFCM-T055-BD**. The FMP shall be provided with explanations regarding what are the ATFCM regulations applied on sectors in different periods.

Decision	Justification	Insights
MET	Full Agreement	

- **REQ-ATFCM-T056-BD**. The FMP shall be provided with explanations regarding how hotspots are linked in terms of flights crossing the corresponding sectors.

Decision	Justification	Insights
PARTIALLY +	It is partially met (+). Manual steps have to be done to get the explanation/information	

- **REQ-ATFCM-T057-BD**. The FMP shall be provided the inner reasoning considered in support of the assessment of the criticality of a hotspot to the overall situation.

Decision	Justification	Insights
PARTIALLY +	Overall Agreement. Explanations are not that specific, but the parameters for decision making include the hotspots and flights contributing to these.	

4.2.3 Transparency Requirements Refinement Proposal - ATFCM

From the insights derived in the previous section, the need for a process of *refinement* of the transparency requirements (originally defined in [3]) is outlined, and therefore such a process is proposed in this section.

This conclusion is mainly based on the fact that when defining general transparency requirements, it has been suggested that these shall contain aspects strictly referred to technical features to be implemented by the prototype. If these requirements are accurate enough, they are supposed to contribute in some degree to the final transparency of the solutions without the need for specifying or taking for granted how the final recipient should 'feel' or 'receive' the transparency, or even how the recipient should 'understand' these solutions. All these aspects and conclusions related to interpretability must be purely drawn from the validation activities, which are supposed to evaluate and test how the recipients interact with the solutions offered by our prototypes.

Mixing technical aspects with those involving interpretability in one single requirement is not a good idea. It is outlined that not mixing such aspects would largely facilitate the final evaluation on the fulfilment of a requirement, purely from a technical point of view, thus removing the need for finding complex ways of assessing subjective interpretations.

Therefore, here it is proposed a rewording of some of the original transparency requirements for the sake of technical rigorosity. A total of seven (7) transparency requirements have been identified for further refinement. These requirements are listed as follows, as well as the proposal of change.

The parts of the requirement that need some refinement are marked in bold, italic and underlined for a better readability.

Requirement	Original definition	New proposal
REQ-ATFCM-T013-BD	The FMP shall be provided with explanations and parameters in order to ***reinforce the trust*** of the FMP in the AI system in potential counterintuitive situations (to be identified during the validation phase) for the FMP.	*The FMP shall be provided with explanations and parameters for him/her to gain further insight in potential counterintuitive and/or non-nominal situations*
REQ-ATFCM-T014-BD	The FMP shall be provided with explanations for her/him ***to understand*** the details of the inner process followed during the trade-off process	*The FMP shall be provided with explanations informing about the details followed during the trade-off process*
REQ-ATFCM-T015-BD	The FMP shall be provided with all the parameters for her/him ***to understand*** the details of the inner process followed during the trade-off process.	*The FMP shall be provided with all the parameters involved in the trade-off process*

REQ-ATFCM-T024-BD	The PM shall provide a suitable visualization depicting the expected effect of the implemented DCB measure so that the FMP ***can quickly and intuitively comprehend*** the possible consequences of his/her actions in the resolution of the active hotspot (s).	*The PM shall provide a suitable visualization depicting the expected effect of the implemented DCB measure so that the FMP can see the possible consequences of his/her actions in the resolution of the active hotspot (s).*
REQ-ATFCM-T029-BD	In the case of implementing ATFCM regulations involving specific Regulation Period, Regulation Width and Regulation Rate, the PM shall provide the FMP with the following parameters *so that he/she can understand* the real magnitude of the expected impact and make pertinent decisions: occupancy counts, hourly entry counts, total delay, average delay per flight, estimated additional flown distance.	*In the case of implementing ATFCM regulations involving specific Regulation Period, Regulation Width and Regulation Rate, the PM shall provide the FMP with the following parameters informing about the magnitude of the expected impact and make pertinent decisions: occupancy counts, hourly entry counts, total delay, average delay per flight, estimated additional flown distance.*
REQ-ATFCM-T030-BD	The HRMM shall provide the FMP with up-to-date and relevant parameters in real-time to help the FMP ***univocally understand*** what the current resolution state of all the active hotspots is (e.g., number of flights over the declared capacity).	*The HRMM shall provide the FMP with up-to-date and relevant parameters in real-time, informing on the current resolution state of all the active hotspots is (e.g., number of flights over the declared capacity).*
REQ-ATFCM-T038-BD	Should a set of additional parameters (e.g., those that have not been used in the building of the rule-based system for the identification of hotspots/optispots and might possibly have some relevance in the decision of declaring a hotspot to the NM. For example, parameters related to overall network load, or period of the year, etc.) have been used in the decision of the declaration, they will be provided to the FMP so that he/she can ***clearly understand*** the final decision	*Should a set of additional parameters (e.g., those that have not been used in the building of the rule-based system for the identification of hotspots/optispots and might possibly have some relevance in the decision of declaring a hotspot to the NM. For example, parameters related to overall network load, or period of the year, etc.) have been used in the decision of the declaration, they will be provided to the FMP.*

Table 1. Proposal of refinement for the ATFCM transparency requirements.

These new requirements will be included in the document containing the original transparency requirements (see [3]), thus replacing the old ones.

4.2.4 Extracting Transparency Insights - ATFCM

This section is responsible for identifying and extracting knowledge that may be useful for the derivation of transparency principles and recommendations, based on compliance with the requirements defined at the beginning of the TAPAS project.

The way to highlight some insights will be through the use of bullets (Φ), for the sake of traceability of them throughout the document.

4.2.4.1 Degree of transparency by ATFCM activity

In relation to the fulfillment of requirements based on the breakdown of activities belonging to the ATFCM use case, the following synthesized knowledge is extracted in terms of transparency and explainability aspects provided by the prototype.

The aim here is to provide, in plain language and vocabulary, what are the transparency and explainability capabilities and functionalities achieved and offered by the implemented prototype.

[LEVEL 2] Analysis of the imbalances detected - Identification of hotspots/optispots

In general, the explanations provided by the prototype allow an adequate understanding of the rules used to identify the hotspots occurred. These rules are simple and deterministic, and it was determined that the use of a decision tree was not necessary.

[LEVEL 2] Sector visualizer

The VA module provides effective visual representations for the FMP to gain insight on the actual traffic of the operating sector, thus attempting to abstract her/him to some degree from the actual complexity.

[LEVEL 2] Preparation of DCB measures to solve the hotspot

The prototype manages to provide the various parameters involved and that have the most impact on the calculation of the most effective DCB measure to solve the identified hotspot. In addition, the prototype is able to provide all parameters that affect a decision, but explanations are not provided.

The FMP is provided with parameters and explanations regarding the reasons on why the prototype proposes specific DCB measures, and how such DCB measure impacts on other sectors, as well in potential counterintuitive and/or non-nominal situations.

At a visual level, internally, multiple types of alternative DCB solutions (i.e., with different mixtures of measures) were considered, but eventually only one solution is provided to the FMP via preliminary visualizations on the display. No backup or alternative measures are provided to the FMP which would let him/her decide based on his/her judgment. Visualizations include most of the important parameters.

Regarding the information on the most impacted sectors by the proposed DCB measure, this is not provided by the prototype. It was indicated that the reason for not meeting such functionality was the

lack of time, as additional processing steps are needed to exploit and explore explanation data to identify these sectors.

Main insights:

- Φ Information about explanations was presented but hard to be understood by FMP experts.
- Φ The quality of some explanations must be improved, in terms of clarity and simplicity.
- Φ It is outlined that when explanations are to be requested in transparency requirements, such explanations must be accompanied by some mechanism ensuring that the requirement not only offers explanations, but that they are *understandable* and *actionable* by the recipient. Transparency requirements must include not only the functionality of offering the explanation, but also *ensuring that this meets a minimum degree of interpretability*. For instance, the explanation can be complemented with some metric capable of quantifying the degree of satisfaction with the explanation by the recipient. In this way, and until a minimum level of interpretability/actionability is not achieved by the offered explanation, the transparency requirement should not be considered as fulfilled and this means the technical implementation still require some degree of refinement. Otherwise, the explanations might be *irrelevant* or *meaningless*.
- Φ Transparency requirements must not contain references to aspects that must be acquired during training stages, for example 'trust'. Trust must be tested before system deployment.
- Φ It was highlighted that FMP experts showed an improvement of trust in the system after several days. More time for getting familiarized with the system would have been useful.
- Φ During the debriefing sessions the FMP experts stated that some aggregated information, statistics about solutions impact would be of interest, apart from the information already included in the tool.
- Φ Further sessions and longer interactions of the users with the prototype would probably enforce the need for offering multiple options per type of solution. Currently, the XAI provides several types of solutions, but only one specific option per type, for example one solution with delay regulations, one with Flight Level Capping and one with a mixture of Level Capping and delays.

[LEVEL 2] Preparation of DCB measures to solve the hotspot - *Selection of candidate flights*

The prototype is able to provide the FMP with relevant explanations and parameters about the inner details followed during the trade-off process to select the candidate flights impacted by the DCB measures. This way, the XAI selects the candidate flights to regulate, and decides how much to be regulated and it explains the reason. However, FMPs stated that they do not really need such information. They were provided with explanations on the different steps the algorithm uses to calculate the final solution. But since the algorithm was complex, it was difficult for them to understand it and they stated that they did not need that information during the operation. However, *during training* this information was identified as being useful.

In addition, the prototype accurately provides essential information containing values such as the total number of affected flights for the different proposed solutions; and how long those flights will be affected.

Even though FMPs are not explicitly provided with a list of the candidate flights impacted by the DCB measure, this information was available through other means – e.g., by consulting the list of flights impacted by the measure in the FMP client tool. Also, FMPs were not provided explanations informing about the key parameters involved in the decision for the selection, but the VA prototype does explain the measures per flight.

Regarding the information on how many different airlines are impacted in the different solutions, this is not provided by the prototype. It was indicated that the reason for not meeting such functionality was the *lack of time*, however that functionality is easy to implement.

There also seems to be some discrepancies with respect to different interpretations of requirements by the parts involved in the implementation. For example, in relation to providing the FMP with some alternative or backup solutions. Generally, the developers of the prototype stated that even if one solution is finally provided, multiple options are considered internally before providing the final solution. In Level 2, the ideal case would have probably been to provide backup and promising solutions for the FMP to make decisions based on his/her judgment from the potential proposed alternatives. However, at the same time it was highlighted that the need for multiple solutions per type of solution is something that would need to be tested with the interaction of the users with the prototype.

Main insights:

- Recommendation regarding the accurate definition of general transparency requirements: *'Technical/functional and other (validation/testing related) aspects must not be mixed into one single transparency requirement'*. In other words, transparency requirement must only contain technical aspects. For example, it should not be stated nor assumed in the definition of the requirement how the recipient of the requirement will eventually interact or feel respect to it. The reason is because drawing conclusions on how useful or transparent the functionality provided by a requirement will be received by a human, must solely come from the validation results.

- Recommendation regarding the use of terms such as 'trustable', 'interpretable', 'understandable', 'comprehensive', 'quickly stimulate'. These, if used, must be explicitly defined by means of *adequate descriptions* and the use of *quantifiable and measurable criteria*. Such terms are highly susceptible to subjective interpretations and their definitions must be univocal.

[LEVEL 2] Preparation of DCB measures to solve the hotspot - *Analysis of the DCB measure impact – What-if*

The FMPs were provided with a *'what-if'* functionality that could allow them to test the solutions proposed by the prototype (level 2). The functionality provides decision rules triggered for decision making, as well as arguments for taking counter-decisions, if needed. Tools for running the system

simulation towards producing the final solution are provided, also allowing the visualization of all involved flights and joint decision making. Per decision it is included:

- The parameters that affect the decision taken.
- The parameters that drive alternative decisions.

The impact of the alternative solution types is provided by the prototype too. However, no interaction exists with operators for identifying the effects of hypothetical DCB measures.

In relation to ATFCM regulations, the prototype offers parameters informing about the real magnitude of the expected impact and to make pertinent decisions, for example: occupancy counts, hourly entry counts, total delay, average delay per flight.

Also, representative visualizations of the expected effect of the implemented DCB measure in the resolution of the active hotspot (s) are provided. Solution explorer explicitly displays changes from one situation (e.g., baseline) to another one (e.g., intermediate or final solution), with an appropriate representation of relevant indicators (e.g., accumulated delays, number of hotspots) in their dynamics. This, however, was shown not to be straightforwardly understandable by FMPs.

Regarding the unmet transparency related functionalities in this task, the prototype does not offer explanations informing about the underlying reasons of unresolved hotspot. It was stated the *high technical complexity* as main reason, since explaining the reasons of failure for a specific hotspot may involve testing different measures, not for that hotspot only, but for others that are impacted by the measures. Also, no explanations suggesting alternative DCB measures which might help solve an unresolved hotspot are provided. Similarly, it was also indicated the *high technical complexity* as main reason, explaining that the prototype provides 3 types of solutions. This means that no mixture of measures available can resolve a hotspot. Alternative DCB measures means expanding the system with alternative options (e.g., DAC). Finally, no functionality for setting up different parameters to be used as input to the ML algorithms, and re-running simulations in order to evaluate the diversity of the possible solutions, as well as their potential impact on the network is offered. The stated reason for such unfulfillment was that this functionality entails a high degree of interaction with prototype.

[LEVEL 2] Hotspot Resolution Monitoring

The prototype offers, in real-time, up-to-date and important parameters informing about the current resolution state of all the active hotspots. In particular, the demand per sector and time instance, as well as the flights involved are presented to the FMP.

Regarding the resolution state of the hotspot, the prototype provides all the parameters that influence the decision, for any single decision taken. However, the decision of an agent does not consider a single hotspot, but all the hotspots to which it is involved during its flight.

Should the resolution of a hotspot be not resolved within D-1, the prototype internally considers different alternative DCB measures in order to solve the hotspot considering the current situation of the network. Finally, only one action is offered for the FMP to execute it. The ideal case, as requested in the requirements, would have been to provide several alternatives to allow the FMP to decide whether to select a sub-nominal solution instead of the one being proposed.

Regarding the unmet transparency related functionalities in this task, a list of corrective actions for further assessment is not provided. The highlighted reason was that it did not fit in the paradigm followed. Also, no explanations are provided that describe the underlying reason(s) why a hotspot is not resolved within D-1. This involved *high complexity*, since explaining the reasons of failure for a specific hotspot may involve testing different measures, not only for that hotspot, but also for others that are impacted by the measures.

[LEVEL 3] Declaration of hotspots / optispots

It was required to provide the set of decision rules that led to the automatic declaration, or not, of the hotspot to the NM. It was stated that the prototype works in a way such as it does not focus on individual hotspots, but on the global airspace state. Likewise, in case that a decision tree was to be built as part of the rule-based system, the FMP shall be provided with the complete decision tree containing the rules regarding the declaration, or not, of a hotspot to the NM. Finally, no tree was built but a functionality is offered to follow the considered decision process using the graphic display.

Appropriate visualization means were provided by the VA tool regarding the analyzed hotspot's root cause. The visualization offers information such as the features that justify a decision, factors, trajectories and impacts to other sectors. Additionally, appropriate representation for flights in a form of sequences of crossed sectors was designed. This representation in an interactive and dynamic form is provided within the sector explorer component.

Regarding the unmet transparency related functionalities in this task, no additional parameters that have been used in the decision of the declaration are provided to the FMP. No specific reason for the unfulfillment of this functionality is declared by the authors.

[LEVEL 3] Decision on the DCB measure and flights impacted

The prototype is able to provide explanations containing the reason(s) why each DCB measure is selected for implementation, as well as highlighting key parameters that influenced that decision. Furthermore, differences among sets of parameters considered for alternative decisions are also provided, and adequate explanations for the different types of solutions are provided, which can be further elaborated.

Regarding the impact of the applied DCB measure, the FMP is provided with a comprehensive list of the flights impacted by the measure, as well as with information regarding how long the flights will be impacted. Besides, the most impacted sectors are displayed on a graph, and the most relevant parameters for calculating those sectors. However, no information in relation to the most impacted airlines is given, as initially required. The declared reason is *lack of time*, as this involves further exploration and possible clustering on the explanation data.

Regarding obtaining insights on the possible creation of new hotspots by implementing the selected DCB measure, no explicit and clear explanations are provided. Nonetheless, the 'what-if' functionality for Level 2 allows to assess the impact of the solution at some degree.

[LEVEL 3] Implementation of DCB measures

FMPs are properly provided with explanations confirming that the DCB measure has been successfully implemented in the system, as well as the corresponding flights impacted register.

However, the prototype fails to offer explanations which contain details on the reason (s) about why an implementation of the DCB measure in the system has failed, as well as providing possible actions to take. The indicated reason is *high technical complexity*.

[LEVEL 3] Other requirements

In relation to ATFCM regulations, the prototype offers pertinent explanations on what is/are the regulation (s) applied to any single flight, as well as what are the regulations applied on sectors in different periods. In this line, explanations containing information on what is the joint regulation (s) being applied to a subset of flights are given, even though tracking the aggregation is shown to be difficult. However, no explanations are presented related to what are the regulations on flights satisfying specific criteria. The main reason was the *lack of time*, and additional filtering criteria should be added in the VA modules.

Regarding flights that can be mostly affected by hotspots, no specific explanations identifying the flights that are 'most affected' by hotspots are given. However, flights in a hotspot can be found, and explanations' parameters include the hotspots that justify a DCB measure for a flight.

Even though not by means of clear explanations, exploratory actions involving computational steps can be done to get information to find out how hotspots are linked in terms of flights crossing the corresponding sectors.

Finally, and referring to the assessment of the criticality of a hotspot to the overall situation, key parameters for decision making that include the hotspots and flights contributing to these are given by the prototype.

4.3 CD&R Transparency Requirements Fulfilment

4.3.1 Level 2 of Automation - Task Execution Support

4.3.1.1 Assessment of the planned and desired trajectory profile

➢ **REQ-CDR-T058-OD.** The EC shall be provided with all the parameters that had impact during the evaluation step of the planned against the current trajectory profile.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide the parameters that show the deviation of radar track from the planned trajectory during the evaluation step? The following (3) parameters are expected: 1. The 2D NM deviation (in NM) between planned vs actual trajectory in time t 2. The actual vs planned Cruise Flight Level 3. The actual vs planned speed ➢ YES: **FULLY MET** For this criteria to be true, all 3 parameters are provided by the system. ➢ NO: In the case that the criteria is not fully met: Does the XAI provide only a subset of them? or does the XAI only provide it when a conflict is detected (or for any deviation)? ▪ YES: **PARTIALLY+**: this is the case if 2 of the expected 3 parameters are provided **PARTIALLY-**: this is the case if only 1 of the expected 3 parameters is provided ▪ NO: **UNMET** (if 0 parameters are provided)	The XAI/VA tool provides alerts whenever the actual trajectory deviates from the planned one through a pop-up window. The window contains information on (including the exact time when this deviation takes place): - Actual vs planned flight level - Actual vs planned speed - Actual vs planned 2D route (course deviation)	

➢ **REQ-CDR-T059-OD.** The EC shall be provided with simple explanations, highlighting and comparing the numerical differences in magnitude produced for a same type of parameter (e.g., cruise requested flight level vs. actual flight level, cruise requested speed vs. actual speed) between the planned and the current trajectory profile.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide the values of the parameters above? 1. YES: **FULLY MET** Explanations for all 3 parameters are provided. 2. NO: If not #1, Does the XAI provide only a subset of them? or does the XAI only provide it when a conflict is detected (or for any deviation)? ▪ YES: **PARTIALLY+**: if explanation for 2 of the 3 parameters is provided **PARTIALLY-**: if explanation for only 1 parameter is provided ▪ NO: **UNMET** (no explanation of the parameters is available)	Same as REQ-CDR-T058-OD	The term '*simple*' must be avoided in the definition of transparency requirements if not complemented by adequate description. *This requirement needs refinement*.

➢ **REQ-CDR-V060-OD.** The EC shall be provided with appropriate visualization means in order to visually quantify the resulting gap between the planned and the current trajectory profile.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the VA provide the visualization? 1. YES: In this case, does the provided visualization allow to accurately quantify the requested functionality? (+/- 250 ft, +/- 0.5NM) ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY MET** 2. NO: **UNMET**	The XAI/VA did not provide any 2D/3D map for the conformance monitoring functionality. But it does provide a table with all the parameters above, allowing the ATCO to quantify the gap between planned and actual trajectory. Additionally, this req. is considered to be fully met since the map visualization is also provided to the EC through the ATC platform and radar screen, so there is no need to duplicate	Similar to REQ-CDR-T059-OD, the terms '*appropriate*' and '*visually quantify*' must be fully described by proper metrics and/or vocabulary, or avoided. *This requirement needs refinement*.

		all these functionalities in the VA tool.	

➢ **REQ-CDR-T061-OD.** The EC shall be provided with explanations informing about the possible reason (e.g., restrictions applied by upstream sectors, etc.) of the gap between the planned and the current trajectory profile.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide the reasons of why the flight deviated from the planned trajectory? 1. YES: In this case, does the provided explanation provide an accurate reason that is able to describe the gap? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY MET** 2. NO: **UNMET**	Note that this can only be provided when the deviation is due to an ATC resolution action. Due to this reason, the VA/XAI tool provides which resolution action proposed by the algorithm is the cause behind this deviation (it shows the flightID and the particular action and value).	

➢ **REQ-CDR-T062-OD.** The EC shall be provided with all the relevant parameters that might have impacted/led to the current gap between the planned and the current trajectory profile.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide the requested parameters? 1. YES: In this case, are the provided parameters relevant respect to the resulting gap? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY MET** 2. NO: **UNMET**	Same comment as REQ-CDR-T061-OD	

4.3.1.2 Identification of potential conflicts

➢ **REQ-CDR-T063-OD.** The EC shall be provided with an explanation describing what is the flight state (climb, cruise, descent) that has been considered in the conflict identification process (i.e., the flight state expected when the separation minima is breached), as well as possible corresponding parameters of importance.

Decision	Functional Compliance Criteria	Justification	Insights

MET	Does the XAI provide the requested explanation? 1. YES: In this case, does the provided explanation adequately describe the flight state? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY MET** 2. NO: **UNMET**	The XAI/VA tool provides the attitude of both flights included in the conflict through an arrow indicating if it is climbing, ascending or in the cruise phase.	

➢ **REQ-CDR-T064-OD.** The EC shall be provided with an explanation describing what is the most critical type of violation and severity corresponding to the identified conflict (horizontal or vertical separation), as well as possible identified corresponding parameters (e.g., how much the horizontal separation, or the vertical separation will be violated).

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide the following three parameters about severity of conflicts? 1. MOC (Measure of Compliance) 2. ROC (Rate of Closure) 3. An explanation that clearly describes the type of violation and the severity of the conflict YES: **FULLY MET** (all 3 parameters are provided) NO: 2 parameters: **PARTIALLY MET +** 1 parameter: **PARTIALLY MET -** 0 parameter: **UNMET**	The XAI/VA tool provides the separation minima achieved (horizontal and vertical), including a percentage of compliance with the separation minima MOC and severity score according to this MOC and ROC. Use of colour code (red coloured bars) to highlight the severity. An explanation that clearly describes the type of violation and the severity is exactly the presentation of these parameters, no further explanation is needed.	

➢ **REQ-CDR-T065-OD.** The EC shall be provided with information stating if the aircraft speed has influenced or caused the identified conflict and to what extent.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide any of the requested information? 1. YES: In this case, does the provided information clearly state if the aircraft speed has caused the conflict and to what extent? ▪ YES: **FULLY MET**	The XAI/VA tool provides the rate of closure indicating how close the flights are moving to the collision point. And this is a parameter that affects the final score for severity.	This requirement needs *refinement* since all conflicts are influenced or caused by the speed of the aircraft involved.

	▪ NO: **PARTIALLY MET** 2. NO: **UNMET**		

➢ **REQ-CDR-T066-OD.** Should the EC be unaware of in which specific sector a conflict has been identified, he/she shall be provided with this information.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Is 'the sector at which the conflict is detected' provided? 1. YES: **FULLY MET** 2. NO: **UNMET**	The name of the sector where the conflict is detected is included in the VA visualisation.	

4.3.1.3 Identification of conflict resolution strategies and clearances proposal

➢ **REQ-CDR-T067-OD.** The EC shall be provided with metrics that accurately quantify the uncertainty around the flight trajectories.

Decision	Functional Compliance Criteria	Justification	Insights
UNMET	Is any tolerance error margin around the flight trajectory provided, describing under which margin the XAI takes action? 1. YES: **FULLY MET** 2. NO: **UNMET**	This was out of the scope for the XAI/VA prototype.	

➢ **REQ-CDR-T068-OD.** The EC shall be provided with parameters related to efficiency (flight and time efficiency) along with an explanation that accurately quantifies and describes to what extent these parameters influence the final decision of the most effective conflict resolution strategies.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide the increase of flown distance and increase of time flown for the aircraft involved in the conflict and other aggregated metrics for all the sector? 1. YES: In this case, does the XAI provide an explanation containing its view of the problem? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY MET** 2. NO: **UNMET**	The XAI/VA tool provides the added miles flown, added seconds flown and conflicts foreseen. It also includes a characterisation of the state of the flight after the solution is implemented.	The term *'accurately'* must be described in terms of accuracy metrics, or avoided. *This requirement needs refinement*.

➢ **REQ-CDR-T069-OD.** The EC shall be provided with an explanation that accurately quantifies and describes to what extent the aircraft's ground speed influences the final decision of the most effective conflict resolution strategies.

Decision	Functional Compliance Criteria	Justification	Insights
UNMET	Does the XAI explain why it proposes HDG, FL or DCT instead of speed change due to restrictions or less impact of those solutions? 1. YES: **FULLY MET** 2. NO: **UNMET**	There is no explanation on why the XAI proposes one solution over other, only through the presentation of a rank showing the likelihood of that solution to solve the problem/conflict according to its AI algorithm.	The term *'accurately'* must be described in terms of accuracy metrics, or avoided. *This requirement needs refinement*.

➢ **REQ-CDR-T070-OD.** Should weather data be available, the EC shall be provided with the parameter wind speed along with an explanation that accurately quantifies and describes to what extent this parameter influences the final decision of the most effective conflict resolution strategies.

Decision	Functional Compliance Criteria	Justification	Insights
UNMET	Should weather data be available, does the XAI provide the requested parameter and the explanation? 1. YES: In this case, does the provided explanation accurately quantify and describe what it is required? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY MET** 2. NO: **UNMET**	Weather is out of the scope for this XAI/VA prototype and validation. Requirement unmet.	The term *'accurately'* must be described in terms of accuracy metrics, or avoided. *This requirement needs refinement*.

➢ **REQ-CDR-T071-OD.** The EC shall be provided with adequate metrics informing about the expected accuracy or likelihood of the proposed strategies to resolve the conflict.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide a ranking of the solutions proposed according to effective resolution of conflict (does solve the conflict) and expected impact? 1. YES: **FULLY MET** 2. NO: **UNMET**	There is a rank showing the likelihood of that solution to solve the problem/conflict according to its AI algorithm.	The term *'adequate'* must be described, or avoided. *This requirement needs refinement*.

➢ **REQ-CDR-T072-OD.** The EC shall be provided with the flights impacted by the proposed conflict resolution strategies.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide the requested parameter? 1. YES: **FULLY MET** 2. NO: In this case, is it possible to derive the requested parameter in a different way or by conducting other steps? ▪ YES: **PARTIALLY MET** ▪ NO: **UNMET**	If a flight is impacted with the proposed resolution strategy and creates a new conflict, this is shown in the "Conflict Foreseen" column.	

➢ **REQ-CDR-T073-OD.** The EC shall be provided with a quantification of the impact caused by the proposed conflict resolution strategies.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide the following parameters required for impact assessment? This is planned using 4 parameters: 1. Flight efficiency 2. Time efficiency 3. Subsequent conflicts 4. Foreseen deviations ➢ YES: **FULLY MET** (if all 4 parameters are provided) If all 4 are not provided, is only a subset of them provided? ▪ YES: **PARTIALLY+** (3 parameters) ▪ YES: **PARTIALLY-** (1 OR 2 parameters) ➢ NO: **UNMET** (0 parameters)	XAI/VA tool provides the impact on flight efficiency (added miles), time efficiency (added seconds, duration of the action), subsequent conflicts (foreseen conflicts) and foreseen deviations (state of the aircraft after the solution is applied).	The term '*quantification*' must be described in such a way that clearly states what are the required parameters for conducting the quantification process. If this description is not provided, the term can be vague or meaningless. *This requirement needs refinement*.

➢ **REQ-CDR-T074-OD.** The EC shall be provided with adequate values informing about how long those flights will be affected by the proposed conflict resolution strategies (e.g., when the flights will resume their planned route).

Decision	Functional Compliance Criteria	Justification	Insights

PARTIALLY +	Does the XAI provide information about how long the resolution action will take to complete and how long it will take to resume its planned trajectory? 1. Both parameters are provided: **FULLY MET** 2. One parameter is provided: **PARTIALLY** 3. None of them are provided: **UNMET**	The duration of the resolution action is provided. However, not how and how long it will take the aircraft to resume its FPL. During the validation exercises at level 2 it was shown that it is necessary to inform the ATCO in a more straight forward way (not only through the indication of duration of the resolution action) when the aircraft needs to resume its FPL (e.g.: using an alarm, check list or by including additional clearances to support FPL recovery).	

➢ **REQ-CDR-T075-OD.** Should the resolution of a critical conflict be prioritized over another, the EC shall be provided with explanations clearly stating the reason (s) for such priority (e.g., time to conflict, conflict severity), as well as highlighting the possible consequences in case that it is not successfully resolved.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide a classification or prioritisation of conflict resolution? 1. YES: **FULLY MET** 2. NO: **UNMET**	Prioritisations are done using the severity field, together with MOC and ROT fields. Using red colour bars to indicate which one of the conflicts is more critical.	The term *'clearly'* must be avoided in the definition of transparency requirements if not complemented by adequate description. *This requirement needs refinement*.

➢ **REQ-CDR-T076-OD.** The EC shall be provided with all the additional information that has been used (if so) to calculate the most effective conflict resolution strategies.

Decision	Functional Compliance Criteria	Justification	Insights
MET	If additional information is used in the calculation of the most effective strategies, does the XAI provide it? (AGENTS' States and CONFLICTS tables) 1. YES: **FULLY MET** 2. NO: **UNMET**	Agent states, conflict information is provided, as well as state after the resolution. Simplified information since it is a safety critical situation where too much information is not useful and time consuming.	

➢ **REQ-CDR-T077-OD.** The EC shall be provided with explanations describing the underlying reasons of proposing one conflict resolution over another. These shall cover aspects like the likelihood of triggering secondary conflicts and the expected benefit in terms of flight efficiency, time efficiency, ATCo workload, etc.

Decision	Functional Compliance Criteria	Justification	Insights
PARTIALLY +	Does the XAI provide the requested reason? 1. YES: If this is the case, does the provided explanation clearly describe the underlying reasons of proposing measures over others? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY** 2. NO: **UNMET**	A rank is provided according to the likelihood of the resolution to solve the conflict according to the XAI algorithm.	

➢ **REQ-CDR-T078-OD.** The EC shall be provided with explanations containing insights on to what extent the proposed conflict resolution strategies manage to reconcile with original planned trajectory.

Decision	Functional Compliance Criteria	Justification	Insights
UNMET	Does the XAI provide the waypoint to return after the resolution action? 1. YES: **FULLY MET** 2. NO: **UNMET**	Only the duration of the resolution action is provided.	

➢ **REQ-CDR-T079-OD.** Should the proposed conflict resolution strategies might generate additional conflicts (in-sector or downstream), the EC shall be provided with a list containing all the aircraft to be affected.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Should that case occur, does the XAI provide a list containing all the aircraft to be affected? 1. YES: **FULLY MET** 2. NO: **UNMET**	Resolution actions indicate if they create another conflict (foreseen conflicts) and by clicking on that field more information about the new conflict is provided, for instance flights affected, using a similar view to the first one.	

➢ **REQ-CDR-T080-OD.** Should the proposed conflict resolution strategies might generate additional conflicts (in-sector or downstream), the EC shall be provided with explanations describing in detail the potential impact on other aircraft and in the overall traffic situation.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Should that case occur, does the XAI provide the requested feature (explanation)? 1. YES: If this is the case, does the provided explanation clearly describe the potential impact on the overall traffic situation? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY** 2. NO: **UNMET**	Same comment as REQ-CDR-T079-OD	*Terms as 'in detail'* must be avoided, if not complemented by adequate descriptions of what 'in detail' means. They are at risk of being vague if not defined. *This requirement needs refinement*.

➢ **REQ-CDR-T081-OD.** Should the proposed conflict resolution strategies might generate additional conflicts (in-sector or downstream), the EC shall be provided with explanations including parameters about the conditions of the conflicts occurrence (e.g., location, horizontal and vertical separation at CPA, time to CPA).

Decision	Functional Compliance Criteria	Justification	Insights
MET	Should that case occur, does the XAI provide the requested feature (explanation)? 1. YES: If this is the case, does the provided explanation include all the requested parameter about the conditions of the conflict's occurrence? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY** 2. NO: **UNMET**	Same comment as REQ-CDR-T079-OD	

➢ **REQ-CDR-T082-OD.** Should the proposed conflict resolution strategies might generate additional conflicts (in-sector or downstream), the EC shall be provided with their likelihood of occurrence.

Decision	Functional Compliance Criteria	Justification	Insights
UNMET	Should that case occur, does the XAI provide the likelihood of occurrence? 1. YES: **FULLY MET**	This is unmet.	

	2. NO: In this case, is it possible to derive the requested value in a different way or by conducting other steps? ▪ YES: **PARTIALLY** ▪ NO: **UNMET**		

➢ **REQ-CDR-T083-OD.** The EC shall be provided with information describing what is the planned flight plan of the different aircraft involved in the conflict, after the potential conflict time and location.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide the change in FPL? 1. YES: **FULLY MET** 2. NO: **UNMET**	The FPL only changes when updating the cleared FL or WP when giving a direct. These updates/changes are directly shown in the resolution action and need to be updated in the CWP, in not a non-conformance alert will be produced.	

➢ **REQ-CDR-T084-OD.** Should the proposed conflict resolution strategies might generate additional conflicts (in-sector or downstream), the EC shall be provided with explanations describing alternative solutions and actions in order to mitigate or prevent the occurrence of such conflicts.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Should that case occur, does the XAI provide the requested explanation? 1. YES: In this case, does the provided explanation clearly describe alternative solutions and actions in order to mitigate or prevent the occurrence of such conflicts? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY** 2. NO: **UNMET**	The user can consult alternative and different solutions (depending on the situation and complexity), which imply different subsequent conflicts and impact.	

➢ **REQ-CDR-T085-OD.** The EC shall be provided with explanations describing why a conflict resolution strategy has failed, as well as the expected location, horizontal and vertical separation at CPA, and time to CPA.

Decision	Functional Compliance Criteria	Justification	Insights

PARTIALLY +	Does the XAI provide the requested feature (explanation)? 1. YES: In this case, does the provided explanation describe why the conflict resolution has failed, the expected location, horizontal and vertical separation at CPA and time to CPA? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY** 2. NO: **UNMET**	The prototype won't know why it failed, but it will provide the time and separation at CPA even when the collision has happened (loss as the type of conflict).	

➢ **REQ-CDR-T086-OD.** The EC shall be provided with explanations detailing the inner mechanism for generating the most appropriate clearance. The explanation shall give rationale about the automation process in terms of how the proposed resolution strategies and the agreed exit sector conditions influence the decision for the proposed ATC clearance.

Decision	Functional Compliance Criteria	Justification	Insights
UNMET	Does the XAI provide the requested feature (*detailed explanation of the mechanism for generating the clearance*)? 1. YES: In this case, does the provided explanation offer rationale about the automation process related to how the proposed resolution strategies and the agreed exit sector conditions influence the decision for the proposed ATC clearance? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY** 2. NO: **UNMET**	Not provided. At the same time, it was concluded that it is not necessary in this operational phase, but in the training phase. Since in this safety critical use case, solutions are self-explanatory.	

➢ **REQ-CDR-T087-OD.** The EC shall be provided with explanations about how the selected clearance complies with the agreed restrictions and/or coordination conditions.

Decision	Functional Compliance Criteria	Justification	Insights
UNMET	Does the XAI provide the requested feature (explanation *how the clearance complies with restriction/coordination conditions*)? 1. YES:	Not provided. At the same time, it was concluded that it is not necessary in this operational phase, but in the training phase.	

	In this case, does the provided explanation clearly describe how the selected clearance complies with the agreed exit point and flight level? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY** 2. NO: **UNMET**		

➢ **REQ-CDR-V088-OD.** The EC shall be provided with an appropriate visualisation of unexpected deviations.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide the requested feature? (visualization *of unexpected deviations*): 1. YES: Does the provided visualization allow to clearly see unexpected deviations on the 4D trajectory compared to the planned? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY** 2. NO: **UNMET**	This visualisation was available through the ATC platform and radar screen, apart from the parameters of deviation that are shown in the VA tool.	See REQ-CDR-T060-OD. *This requirement needs refinement.*

➢ **REQ-CDR-V089-OD.** The EC shall be provided with an appropriate visualisation for conformance monitoring related aspects.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide visualisation for conformance monitoring in the evaluation step? YES: **FULLY MET** In this case, does it provide it only for resolution actions? ▪ YES: **PARTIALLY** ▪ NONE OF THEM: **UNMET**	This visualisation was available through the ATC platform and radar screen, apart from the parameters of deviation that are shown in the VA tool.	See REQ-CDR-T060-OD. *This requirement needs refinement.*

➢ **REQ-CDR-V090-OD.** The EC shall be provided with an appropriate visualisation of all conflicts at the time of their appearance, focusing on those that are going to be resolved.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide the requested feature? (visualization *of all conflicts as they appear*): 1. YES: In this case, does the provided visualization allow to clearly see all conflicts at the time of their appearance, focusing on those that are going to be resolved? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY** 2. NO: **UNMET**	Information of the conflict is available for the EC, including time of appearance, 4D representation (through the VA tool and radar screen).	See REQ-CDR-T060-OD. *This requirement needs refinement.*

➢ **REQ-CDR-V091-OD.** The EC shall be provided, on demand, with an appropriate 2D visualisation of the expected conflict. This visualization must represent, intuitively, the following parameters:

- Φ Horizontal separation at CPA
- Φ Vertical separation at CPA
- Φ Time to CPA
- Φ A depiction of the trajectories of the different aircraft involved in the conflict, from the time of request up to the expected CPA
- Φ Flight levels and speeds for the different aircraft.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Does the XAI provide the requested feature? (visualization *of the conflict in 2D with the associated parameters*): 1. YES: In this case, does the provided visualization show all the requested parameters? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY** 2. NO: **UNMET**	The XAI/VA tool provides horizontal separation at CPA, vertical separation at CPA, time at CPA, depiction of the trajectories, including a 4D map, 2D + time and a vertical depiction of the conflict with flight levels of aircraft. Speed and flight levels as well as 2D map was also available through the ATC platform and radar screen.	See REQ-CDR-T060-OD. *This requirement needs refinement.*

4.3.1.4 Conformance monitoring resolution

➢ **REQ-CDR-T092-OD.** Should the identified conflict, for which a resolution has been already proposed, be caused by a non-conformance of the ATC clearance, this information shall be provided to the EC.

Decision	Functional Compliance Criteria	Justification	Insights

MET	Should the identified conflict be caused by a non-conformance of the ATC clearance, does the XAI provide inform about it in the evaluation step? YES: **FULLY MET** NO: In this case, does it provide it only for resolution actions? ▪ YES: **PARTIALLY** ▪ NONE OF THEM: **UNMET**	The XAI/VA tool indicates if the conflict comes from another resolution action through the "Due to" column, where information on the previous resolution actions is specified.	

4.3.2 Level 3 of Automation – Conditional Automation

4.3.2.1 Clearances implementation

➢ **REQ-CDR-T093-OD.** Should a flight be deviated from a given ATC clearance and any conflict (s) is (are) triggered, the EC shall be provided with explanations informing about the potential impact that such conflict (s) might cause in-sector or downstream.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Should that case occur, does the XAI provide the requested feature (explanation *about possible in-sector/downstream impact*)? 1. YES: In this case, does the provided explanation clearly inform about the potential impact in in-sector or downstream? ▪ YES: **FULLY MET** ▪ NO: **PARTIALLY** 2. NO: **UNMET**	If a flight deviates from its trajectory and creates a new conflict, this new conflict is detected, and new resolution actions are proposed.	

➢ **REQ-CDR-T094-OD.** Should a flight be deviated from a given ATC clearance and any conflict (s) is (are) triggered, the EC shall be provided with possible actions to take.

Decision	Functional Compliance Criteria	Justification	Insights
MET	Should a case occur where a deviation from an ATC clearance results in conflict, does the XAI offer possible actions to take? 1. YES: **FULLY MET**	If flight deviates from its trajectory and creates a new conflict, this new conflict is detected and new resolution actions are proposed.	

	2. NO: **UNMET**		

4.3.3 Transparency Requirements Refinement Proposal – CD&R

In this section, some refinements for the original CD&R transparency requirements are proposed. A total of fourteen (14) transparency requirements have been identified for further refinement. These requirements are listed as follows, as well as the proposal of change. The parts of the requirement that need some refinement are marked in bold, italic and underlined for a better readability

Requirement	Original definition	New proposal
REQ-CDR-T059-OD	The EC shall be provided with ***simple*** explanations, highlighting and comparing the numerical differences in magnitude produced for a same type of parameter (e.g., cruise requested flight level vs. actual flight level, cruise requested speed vs. actual speed) between the planned and the current trajectory profile.	The EC shall be provided with explanations, highlighting and comparing the numerical differences in magnitude produced for a same type of parameter (e.g., cruise requested flight level vs. actual flight level, cruise requested speed vs. actual speed) between the planned and the current trajectory profile.
REQ-CDR-V060-OD	The EC shall be provided with ***appropriate visualization*** means in order to ***visually quantify*** the resulting gap between the planned and the current trajectory profile.	The EC shall be provided with visualization means in order to quantify the resulting gap (+/- 250 ft, +/- 0.5NM) between the planned and the current trajectory profile.
REQ-CDR-T065-OD	The EC shall be provided with information stating if the aircraft speed has influenced or caused the identified conflict and to what extent	The EC shall be provided with information informing about the rate of closure to the collision point.
REQ-CDR-T068-OD	The EC shall be provided with parameters related to efficiency (flight and time efficiency) along with an explanation that ***accurately*** quantifies and describes to what extent these parameters influence the final decision of the most effective conflict resolution strategies.	The EC shall be provided with parameters related to efficiency (flight and time efficiency) along with an explanation that describes to what extent these parameters influence the final decision of the most effective conflict resolution strategies.
REQ-CDR-T069-OD	The EC shall be provided with an explanation that ***accurately*** quantifies and describes to what extent the aircraft's ground speed influences the final decision of the most effective conflict resolution strategies.	The EC shall be provided with an explanation that describes to what extent the aircraft's ground speed influences the final decision of the most effective conflict resolution strategies.
REQ-CDR-T070-OD	Should weather data be available, the EC shall be provided with the parameter wind speed along with an explanation that ***accurately*** quantifies and describes to what extent this parameter influences	Should weather data be available, the EC shall be provided with the parameter wind speed along with an explanation that describes to what extent this parameter influences the

	the final decision of the most effective conflict resolution strategies.	final decision of the most effective conflict resolution strategies.
REQ-CDR-T071-OD	The EC shall be provided with ***adequate*** metrics informing about the expected accuracy or likelihood of the proposed strategies to resolve the conflict.	The EC shall be provided with a ranking of the proposed solutions according to effective resolution of conflicts and expected impact.
REQ-CDR-T073-OD	The EC shall be provided with a ***quantification*** of the impact caused by the proposed conflict resolution strategies.	The EC shall be provided with the following parameters to assess the impact caused by the proposed conflict resolution strategies: • *Flight efficiency* • *Time efficiency* • *Subsequent conflicts* • *Foreseen deviations*
REQ-CDR-T075-OD	Should the resolution of a critical conflict be prioritized over another, the EC shall be provided with explanations ***clearly stating*** the reason (s) for such priority (e.g., time to conflict, conflict severity), as well as highlighting the possible consequences in case that it is not successfully resolved.	Should the resolution of a critical conflict be prioritized over another, the EC shall be provided with an explanation describing the reason (s) for such priority (e.g., time to conflict, conflict severity), as well as highlighting the possible consequences in case that it is not successfully resolved.
REQ-CDR-T080-OD	Should the proposed conflict resolution strategies might generate additional conflicts (in-sector or downstream), the EC shall be provided with explanations describing ***in detail*** the potential impact on other aircraft and in the overall traffic situation.	Should the proposed conflict resolution strategies might generate additional conflicts (in-sector or downstream), the EC shall be provided with an explanation describing the foreseen conflicts and flights affected.
REQ-CDR-V088-OD	The EC shall be provided with an ***appropriate visualisation*** of unexpected deviations.	The EC shall be provided with a visualisation of the unexpected deviations on the real 4D trajectory compared to the originally planned.
REQ-CDR-V089-OD	The EC shall be provided with an ***appropriate visualisation*** for conformance monitoring related aspects.	The EC shall be provided with a visualisation for conformance monitoring in the evaluation step.
REQ-CDR-V090-OD	The EC shall be provided with an ***appropriate visualisation*** of all conflicts at the time of their appearance, focusing on those that are going to be resolved.	The EC shall be provided with a visualization that allows to see all conflicts at the time of their appearance, focusing on those that are going to be resolved.
REQ-CDR-V091-OD	The EC shall be provided, on demand, with an ***appropriate 3D visualisation*** of the expected conflict. This visualization must represent, ***intuitively***, the following parameters: • *Horizontal separation at CPA* • *Vertical separation at CPA* • *Time to CPA* • *A depiction of the trajectories of the different aircraft involved in the conflict,*	The EC shall be provided, on demand, with a 2D visualisation of the expected conflict. Such visualization must represent the following parameters: • *Horizontal separation at CPA* • *Vertical separation at CPA* • *Time to CPA* • *A depiction of the trajectories of the different aircraft involved in the*

	from the time of request up to the expected CPA • *Flight levels and speeds for the different aircraft.*	*conflict, from the time of request up to the expected CPA* • *Flight levels and speeds for the different aircraft.*

Table 2. Proposal of refinement for the CD&R transparency requirements.

4.3.4 Extracting Transparency Insights – CD&R

4.3.4.1 Degree of transparency by CD&R activity

In relation to the fulfilment of requirements based on the breakdown of activities belonging to the CD&R use case, the following synthesized knowledge is extracted in terms of transparency and explainability aspects provided by the prototype.

[LEVEL 2] Assessment of the planned and desired trajectory profile

The prototype is able to provide alerts whenever the actual trajectory deviates from the planned one through a pop-up window. Thus, the transparency provided by the prototype comes in form of key parameters that had impact during the evaluation step of the actual against the planned trajectory such as: actual vs planned flight level, actual vs planned speed, actual vs planned 2D route (course deviation), and the exact time when the deviation takes place.

Regarding visualisation, the prototype is able to provide a visualisation that facilitates the quantification (+/- 250ft, +/- 0.5NM) of the resulting gap between the planned and actual trajectory. Although the prototype does not provide any 2D/3D map for the conformance monitoring functionality, it does also provide a table with all the required parameters, thus promoting further transparency of the solution. This was proved to facilitate the ATCO the quantification of the resulting gap between planned and actual trajectory. Furthermore, a map visualization is also provided to the EC through the ATC platform and radar screen, so there is no need to duplicate all these functionalities in the VA tool.

In those cases when the deviation is due to an ATC resolution action, the prototype is also able to provide which specific resolution action proposed by the algorithm is the cause behind this deviation between the planned and the current trajectory profile. This is done through the display of key parameters as the flightID and the particular action and value.

Insights:

- Φ It is outlined that using some terms in the definition of transparency requirements, such as ‘simple’, must be accurately defined in terms of what simplicity means in that context. This could represent: low number of parameters, low complexity, low number of steps, etc. Therefore, terms such as *'simple'* must be subject to objective evaluation, by these criteria, in order to assess the fulfilment of the requirement. In addition, other terms such as ‘*appropriate*’ and ‘*visually quantify’* must be fully described in the transparency requirements by proper metrics and/or vocabulary, or avoided.

[LEVEL 2] Identification of potential conflicts

Regarding the identification of potential conflicts, the implemented prototype is able to give accurate insights about the flight states that have been considered in the conflict identification process. This is achieved by providing the EC with the attitude of both flights included in the conflict through an arrow indicating if it is climbing, descending, or in the cruise phase.

Additionally, the EC is provided with key parameters about the type of violation and the assessed severity of the conflict. In particular, the tool provides the separation minima achieved between the two involved flights (horizontal and vertical), including a percentage of compliance with the separation minima MOC and severity score according to this MOC and ROC. Likewise, colour codes (red coloured bars of differing length and colour intensity) are used to highlight the conflict severity. Additionally, the time when the conflict starts and the time at the CPA are also available through the VA display.

Regarding the possible influence of the aircraft speed in the conflict, the prototype contributes to transparency in terms providing the EC with the Rate of Closure (ROC) indicating how close the flights are moving to the final collision point. Furthermore, since the XAI component detects conflicts inside the sector of study and in the downstream sector (in the surrounding area, close to the sector of focus), the exact sector where the conflict is detected is also provided by the VA visualisation.

[LEVEL 2] Identification of conflict resolution strategies and clearances proposal

Attending to how the aircraft's ground speed might influence the final decision of the most effective conflict resolution strategies, it was initially required some degree of explainability to understand this possible fact. Although there is *no direct explanation* on why the prototype proposes one solution over another (for instance, why it proposes a HDG, FL or DCT instead of speed change), the prototype manages to offer clear insights on this regards through the presentation of a rank according the likelihood of the offered solution to effectively solve the problem and its expected impact.

Besides this, if a flight is somehow expected to be impacted by the proposed resolution strategies, and consequently a new conflict is going to be caused (in-sector or downstream), the EC will be accordingly notified about that. Specifically, and for the sake of transparency in such impact assessment, the prototype can provide relevant parameters on *flight efficiency* (added miles), *time efficiency* (added seconds, duration of the action), *subsequent conflicts* (foreseen conflicts) and *foreseen deviations* (state of the aircraft after the solution is applied). Also, even though no explicit explanations describing alternative solutions and actions that mitigate or prevent the occurrence of such conflicts are given, the EC can consult *alternative* and *different* solutions. These depend on the situation and complexity, and imply different subsequent conflicts and impact, which also adds significant value to the decision-making process. Note that the users indicated that, due to the short time frame in which decisions need to be made, this level of detail was unlikely to be required/used in an operational context but it may be useful when training users on the tools or during certification/verification activities once the prototypes are more mature.

In the case that weather data was available, it was original required that the EC shall be provided with the parameter wind speed along with an explanation that described to what extent this parameter influences the final decision of the most effective conflict resolution strategies. However, this turned out to be irrelevant for these experiments, which were conducted using SACTA traffic simulator in a standard atmosphere with no variations on wind nor other weather conditions, therefore *out of the scope* of the prototype.

In relation to key information such as how long the offered resolution action will take to complete, as well as how long it will take to resume its planned trajectory, only the duration of the resolution action is provided by the prototype, but not how and how long it will take the aircraft to resume its FPL. Likewise, regarding *explainability*, the prototype is not able to provide any explanation regarding to what extent the proposed conflict resolution strategies manage to reconcile with original planned trajectory. That is, the waypoint to return after the resolution action is not provided, only the duration of the resolution action.

The prototype also deals with and provides transparency on *prioritisation* and *classification* of conflicts. These are handled by using the severity field, together with MOC and ROC fields. In particular, red colour bars are used to indicate which one of the conflicts is more critical, which gives prioritisation insights to the EC.

Also, the prototype offers *simplified information* that was used in the calculation of the most effective conflict resolution strategies. For example, agent states and conflict information are provided to the EC, as well as state after the resolution. Simplified information is given by the prototype since it is a safety critical situation were too much information is not useful and time consuming to the EC.

About the potential *changes in flight plans*, the prototype does not directly provide changes in FPL to the EC. Specifically, the FPL only changes when updating the cleared FL or WP when giving a direct. These updates/changes are directly shown in the resolution action and need to be updated in the CWP, if not a non-conformance alert is produced and offered.

A *missing* and *relevant functionality* would have been the inclusion of explanations describing *why* a conflict resolution strategy has *failed* and time to CPA. The implemented prototype does not know the reason why a conflict resolution strategy failed. Instead, it provides the time and separation at CPA even when the collision has happened (loss as the type of conflict).

Regarding *clearances*, it was originally requested to provide the EC with explanations regarding two aspects. The first one was to describe the inner mechanism used to generate the most appropriate clearance. Such explanation should give rationale about the automation process of how the proposed resolution strategies and the agreed exit sector conditions influence the decision for the proposed ATC clearance. A second explanation was requested to the prototype to give insights on how the selected clearance complies with the agreed exit point and flight level. The prototype does not provide none of these explanations, and it was stated that it was not necessary during the operational phase.

In relation to *visualisations*, the prototype offers:

- The actual 4D trajectory compared to the planned one (that is available through the ATC platform) to better understand unexpected deviations.
- A visualisation for conformance monitoring related aspects in the evaluation step.
- All the conflicts at the time of their appearance, focusing on those that are going to be resolved. Specifically, information of the conflict is available, including a 4D representation (through the VA tool and radar screen), horizontal separation, vertical separation, time at the start of the conflict, time at the end of the CPA, a vertical depiction of the conflict with flight levels of aircraft, speed, and flight levels as well as a 2D map.

Main insights:

- The term '*accurately*' must be described in terms of accuracy metrics, scores, or avoided during the definition of transparency requirements.

- The term '*adequate*' must be fully described to give it a meaning, or avoided during the definition of transparency requirements.

- The term '*quantification*' must be described in such a way that clearly states what are the required parameters for conducting the quantification process. If a univocal description is not provided during the definition of transparency requirements, the term can be vague or meaningless.

- Terms as 'in *detail*' must be avoided during the definition of transparency requirements, if not complemented by adequate descriptions of what 'in detail' means. They are at risk of being vague otherwise.

[LEVEL 2] Conformance monitoring resolution

In this step, should the identified conflict be caused by a non-conformance of the ATC clearance, the prototype is able to indicate to the EC if the conflict comes from another resolution action.

The prototype indicates whenever a non-conformance event arises according to the proposed solutions implemented. However, the resolution of this non-conformance event is responsibility of the controller.

[LEVEL 3] Clearances implementation

For the clearances implementation, should a flight be deviated from a given ATC clearance and any conflict (s) is (are) triggered, this new conflict can be detected by the prototype and therefore new resolution actions are proposed to the EC and applied automatically (through a '*Ghost controller*') in full automation level 3.

4.4 General fulfilment of transparency requirements

The degree of fulfilment of the transparency requirements reached by the prototype for the **ATFCM** use case is as follows:

Degree of fulfilment	MET	PARTIALLY +	PARTIALLY -	UNMET
Number of requirements	34	6	4	13

Table 3. Degree of fulfilment of the transparency requirements by the ATFCM prototype

From the 57 transparency requirements originally defined, 34 requirements have been fully met, 6 partially met +, 4 partially met -, and 13 are unmet. From these results, and considering that partially met + requirements are close to being considered as fully met, it can be concluded that the degree of technical fulfilment of the transparency requirements for the ATFCM use case is **high**.

Respect to the **CD&R** use case, the degree of fulfilment of the transparency requirements implemented by the prototype is:

Degree of fulfilment	MET	PARTIALLY +	PARTIALLY -	UNMET
Number of requirements	27	3	0	7

Table 4. Degree of fulfilment of the transparency requirements by the CD&R prototype

From the 37 transparency requirements originally defined, 27 requirements have been fully met, 3 partially met + and 7 unmet. In other words, 30 out of 37 requirements are almost fully fulfilled, results that confirm a ***remarkably high*** degree of technical fulfilment of the CD&R prototype respect to the original transparency requirements.

If the fulfilment of requirements is analysed from an overall point of view, that is by considering the CD&R and ATFCM use cases together, the aforementioned insights scale in a clearer manner.

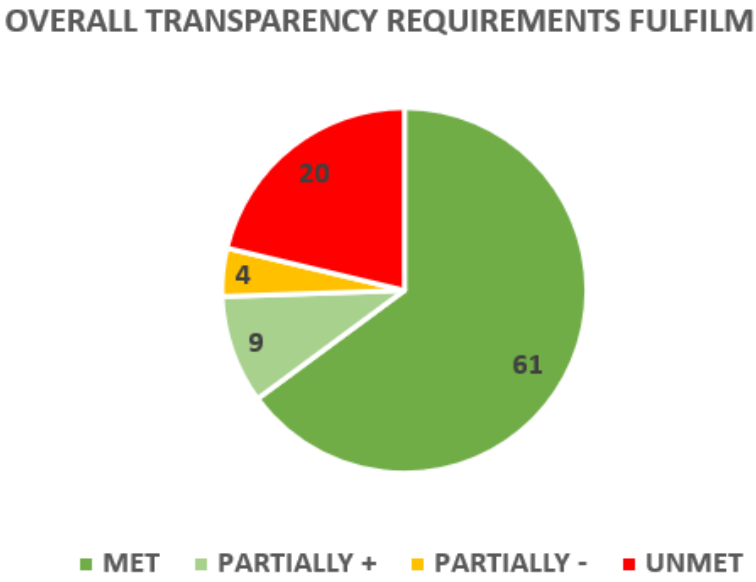

Figure 2. Transparency Requirements Fulfilment achieved by the ATFCM and CD&R prototypes

70 out of 94 requirements are very close to being fully met, which represents roughly **¾ of fulfilment** respect to the total number of original transparency requirements. This certainly represents a *high degree of fulfilment in terms of the transparency and explainability* requested to be achieved by the prototypes. These results must be highlighted due to the fact that TAPAS is an exploratory research project, the expected TRL at the project completion is very low and the current level of maturity of XAI technology is increasingly, but steadily, making progress.

However, according to Figure 2, almost ¼ of the transparency requirements are *unmet* or *vaguely* implemented. The exact reasons are provided above, but in the next section (Section 4.5) we aim at providing further rationale regarding the fulfilment of requirements and the impact on the transparency achieved, and vice-versa.

4.5 On the fulfilment of requirements and its impact on transparency and explainability

Previously, in Sections 4.2.4 and 4.3.4, the most important reasons describing why some requirements have not been technically implemented by the prototypes are given. However, this information seems insufficient to analyze some other relevant questions.

In particular, an aspect of interest is the traceability between the fulfilment of transparency requirements and the real impact on transparency and explainability achieved, and vice versa. For example, some of the questions we are interested in analyzing are:

- *Is it possible to identify from 'undesired or incomplete results' after the validation activities the potential absence of key requirements that if existed would have prevented such 'failures/lack of transparency functionalities' from occurring?*
- *Or those requirements that, although being met or partially met, should have been formulated in a more accurate way, present fundamental flaws or are simply irrelevant in terms of transparency?*
- *Similarly, to what extent the unmet requirements turned out to be actually important for some transparency or explainability aspects?*
- *Is it feasible to outline some of these missing functionalities or 'gaps' based on the results and the human feedback collected after validation?*

Regarding the previous questions and the impact of the requirements coverage in the transparency and explainability, for the ATFCM use case it was found that:

- In general, all relevant information for the users was included in the prototypes. No major issues were detected with the partially accomplishment or no coverage of the requirements.
    - However, during the validation activities, it was shown that aggregated information on the impact of the solutions was also considered to be useful for the users. This adds another dimension to the already formulated requirement REQ-ATFCM-T024-BD on the impact of the solutions. This impact is also included in the requirements REQ-ATFCM-T012-BD, REQ-ATFCM-T018-BD, REQ-ATFCM-T021-BD, REQ-ATFCM-T044-BD and REQ-ATFCM-T045-BD that due to technical and time constrains were not finally covered in the prototypes.
    - Additionally, the requirement REQ-ATFCM-T045-BD also points out an important topic highlighted by the operational experts during the exercise execution, that is, in order to *avoid bias* and *guarantee all airlines are considered in the same way and no prioritisation* is being made. Nonetheless, this information was considered not relevant to be presented to the user during the operational phase, but in the *training* or *certification phase*.
    - Some unmet requirements were not considered to have a major impact in the exercises. In other words, they can be considered as *irrelevant*.

- This is case of REQ-ATFCM-V004-BD as trajectories were not shown clustered in flows, this could be desirable but *not necessary*.

- REQ-ATFCM-T028-BD was also not covered, but the operational experts participating in the trials *did not show the need* to tune the algorithm. What was important to them is that the solutions work efficiently, and they did not mind the exact parameters the algorithm considered in this operational phase (during training and certification would be needed such information). As in the current operation methods where the FMPs use CASA algorithm to compute regulations and, they do not know what parameters or what conditions it considers when computing the delays.

- Explanations were provided by the prototype, but they were *too detailed*.

  - For example, *there was no need* for a decision tree for the identification of a hotspot since this identification was self-explanatory to the user (REQ-ATFCM-T002-BD). Similarly, *no detailed explanations on the decisions proposed were needed* in the operational phase, but in training or certification phase (REQ-ATFCM-T007-BD, REQ-ATFCM-T038-BD).

  - *When the solutions failed* the prototype did not provide explanations on the underlying reasons why they failed to solve the problem nor further actions were proposed in those cases (REQ-ATFCM-T025-BD, REQ-ATFCM-T026-BD, REQ-ATFCM-T032-BD, REQ-ATFCM-T050-BD). This would be desirable to have for the user, but time limitations prevented that the developed tools and algorithms met these requirements.

- REQ-ATFCM-V008-OD and REQ-ATFCM-T017-BD were not fully covered, as the prototype provided alternative solutions but not in the desired way by the user. Especially in automation level 2, where multiple solutions would be found useful by the operational experts. This was one of the lessons learnt, together with the above comments, which was tackled in the following CD&R use case.

As for the CD&R exercise, regarding the *partially met* and *unmet requirements* and their impact, it was found:

- Generally, *the prototype provided the desirable information* on the detection and resolution actions of the conflicts.

  - Nevertheless, it was stated that the 'resume to FPL trajectory' action *should have been provided in a clearer and more straight forward* way by the prototype (REQ-CDR-T074-OD, REQ-CDR-T078-OD). This partial coverage of the requirements had no major impact on the exercises, but *it contributed to increase the workload* of the ATCO participants.

  - REQ-CDR-T069-OD was not covered, and **it was shown to be useful** to know why the algorithm chooses one solution over another, that is, especially during the training and certification phase because during the operational phase the user has limited time and mainly wants the solution to work.

- Other *unmet requirements* were considered out of the scope for the prototype as they did not add value in this operational phase, but in the training phase. This is the case of REQ-CDR-T067-OD, REQ-CDR-T082-OD, REQ-CDR-T086-OD and REQ-CDR-T087-OD (uncertainty would be desirable but not necessary since the amount of information to be provided should be simplified and also it was not included is the computation of uncertainty is not an easy matter to considered, in particular with restricted time development).

4.6 Learning to write transparency requirements in XAI

The overall process of writing transparency requirements allowed us to extract a series of valuable lessons. These are indeed worth highlighting for consideration in future writings of transparency requirements in XAI.

Defining effective requirements that foster transparency and explainability in XAI prototypes has been proved not to be something trivial. It requires careful thinking, accurate planning and design, and extense domain knowledge on which the XAI system is going to be built. Making it harder, as of 2022, very few research works (or even none) have provided relevant clues, methodologies or general ways on how to do it adequately and effectively. We made it from scratch, with very little and even lack of reference or guide.

This section lists some of the main insights extracted and lessons learnt regarding the definition of transparency requirements in XAI for the TAPAS project. These are supposed to serve as a general guideline when writing transparency requirements in different types of XAI projects. Sometimes these insights will be directly applicable to other projects, some others not. This will depend on the domain of application. In general, though, we made big efforts when providing these lessons as *generic* as possible in order to make them *reusable*, regardless the field in which XAI technology is applied.

Later, Section 7.3 translates and further develops these insights into the form of *recommendations* that should *be considered by future XAI developers* as a basis for processes involving the definition of requirements on transparency and explainability.

- Φ Transparency requirements must contain clear and specific information and instructions about what to do. They must avoid ambiguity and must focus on describing how to provide transparency about a given process.

- Φ Sometimes, clear explanations will give rationale about processes more effectively and can justify the reason for a system's decision.

- Φ Requirements must not contain information about aspects that shall/cannot be demonstrated during prototype validation. Drawing conclusions on how transparent the functionality provided by a requirement will be perceived by a human must come from the validation results.

- Φ Certain vocabulary must be avoided in the transparency requirements. For example: 'trustable', *'quickly stimulate', 'clearly', etc.* If any of these terms is used, it must be complemented by *descriptions* and adequately supported by *quantifiable criteria.*

- Φ Experimentation suggests that if explanations are requested in any transparency requirement, these must be accompanied by adequate mechanism to ensure that not only the explanation is offered, but also understood by the recipient. Otherwise, explanations can be irrelevant.

- Φ The evaluation of the fulfilment of the transparency requirements achieved by the implemented prototypes must be conducted as objectively as possible and based on well-defined criteria.

- Φ Refining, iterating, and updating transparency requirements is sometimes necessary and a good practice. It benefits XAI prototype *maintenance* and *enhancement*.

- Φ There is no clear evidence to conclude that the need to achieve high automation level necessarily impacts the way transparency requirements are defined. More experimentation is needed.

5 Expert Feedback Related to Transparency

This section presents valuable expert feedback that has been provided as a result of interactions in progress meetings and workshops with well-recognized experts in diverse domains such as ATM, AI, Human Factors, etc. Such feedback has been collected through the development of the project and is presented here.

This feedback yields important insights on how elements of transparency and explainability should be provided to humans interacting with AI-based solutions in ATM. In particular Section 5.1 presents some of the decisions agreed on the general provision of transparency. For example, some proposals regarding the way of providing transparency to the end-user depending on the ongoing operational action and its time horizon are considered. Also, revealing ideas on the different human necessities of transparency depending on the operational context and other factors are provided. Finally, Section 5.2 poses some questions that, if addressed, could provide clues about the most important aspects to consider when analyzing the potential impact of a varying quality in the transparency provided to humans.

From these exposed ideas and proposals, a set of insights are also derived which are later considered and further elaborated in the definition of the principles and recommendations.

5.1 Decisions on the General Supply of Transparency

The nature, quantity and timing of the transparency provided by the prototype developed in TAPAS rely on several aspects, presented next. There is not yet a universal and widely agreed definition for the term 'transparency' in AI, thus we consider that this term should be merely understood as a simplification concept of all those elements that foster and contribute to the understanding of the solutions given by the prototype to the end user. For instance, these elements sometimes come in the shape of adequate explanations (e.g., *visual, textual, examples*) giving rationale or understanding to some decisions behind automation processes, some others they are specific *parameters/indicators/values* that had importance in the outcome of an optimization process, as well as appropriate *visualizations* intentionally devised to visually stimulate and facilitate the human's comprehension of a task involving multiple parameters or multidimensional values.

In order to adequately provide transparency to the different situations and tasks given across the ATFCM and CD&R use cases, relevant feedback provided by different ATCOs experts and members of our Advisory Board has been considered in order to define how transparency must be provided by means of adequate transparency requirements. The feedback presented here is related to aspects such as how many different levels of transparency must be provided, how the criticality of the action might impact the required transparency, as well as the time-horizon in which the action takes place.

In general, the most important elements that we have identified and considered in order to provide different degrees of transparency to the requirements are those related to:

- Time-horizon: ranging from Pre-tactical (focused on D-1, but could be further extended up to D-7) to Tactical D0.

- Criticality: safety critical, non-safety critical.
- Complexity: there are tasks that are more demanding in terms of cognitive load than others.
- Accuracy: the accuracy of the solution provided by the XAI prototype might vary for the different tasks, and this aspect must be considered to provide more or less degree of transparency to the end user.
- Uncertainty: the higher the uncertainty associated to a solution, the higher transparency might be needed.

Insights:

Φ Aspects such as *time-horizon, criticality, complexity, accuracy and uncertainty* must be considered when designing and tailoring general transparency requirements in ATM.

5.1.1 Timing

Regarding the timing, two major ways to tender transparency to the end-user are proposed:

- **Off-line transparency.** This involves providing any means of transparency *-pre* or *-post* the operational action. For example:
  - *-Pre* mainly refers to the training process of ATCOs in which they are instructed and taught on how to operate with specific tools. Also, this applies to system development and testing, verification, and certification stages.
  - *-Post* refer to the fact of providing transparency once a specific event has occurred and so the end-user requests some transparency in order to possibly understand the outcome of a process. For example, in case of formal post-ops investigations after an operational incident.
- **On-line transparency**. This involves providing any means of transparency during the operational action itself. This can be seen as real-time transparency supply.

Insights:

Φ Taking into account the right timing when applying transparency to the end user is important. '*Off-line transparency*' allows providing transparency -pre or -post the operational action; whilst '*On-line transparency*' focuses on offering transparency during real operations.

5.1.2 Levels of Transparency

Different levels of transparency might be needed and so provided, depending on the ongoing action.

- **No transparency**. Should the solution given by the system be explicative enough (self-explanatory) or it is known to be reliable by past experience, then *no additional information is needed* nor provided.
- **Transparency on demand**. The system may provide additional information to explain the solution reached, if requested by the human operator. For example, human operators should be able to ask for rationale at any time, or to request explanations on what happened when certain solutions go out of the acceptable frame. The decision to ask for further justification about the proposals/decisions of the system should remain with human operators, depending

on the operational context. This level also encompasses the possibility of '*switching off*' or '*disabling*' the provided transparency, if the operator is not trusting it. For example, it might be that if transparency is provided all the time, excessive workload will add up to the cognitive task without a possible benefit for the operator.

- **Transparency by default**. The system acknowledges the unusual condition of a solution reached and provides, by default, transparency on it.

Note: the TAPAS project focuses on 'On Demand' and 'By Default'. These modes were regarded as being the most interesting and innovative from a research point of view, therefore we focused on exploring and experimenting with these ways of providing transparency for the ATFCM and CD&R operational use cases.

Insights:

- Φ Transparency can be offered in *multiple ways*, or *levels*, depending on aspects such as the operator's prior knowledge/experience of a process, the need (or lack of it) to obtain relevant insights on specific system's solutions/decisions, or even the operator's current cognitive workload.

5.1.3 Combining Decisions to Unveil Further Insights

By combining the aforementioned concepts, more elaborated insights can be obtained.

- Regarding the timing:
  - Pre-tactical FMPs at D-1 can manage very complex situations at the network level and therefore operators' need to understand the solution or alternative of solutions proposed. They have more time to analyze in detail the **on-line** explanations and other transparency elements. There is *no time pressure* and it is non-safety critical.
  - Tactical controllers at D0 manage conflict resolution with a more limited number of parameters. The tasks that they conduct are safety critical and need to be done potentially under varying levels of pressure: therefore, there *is little or no time for on-line explanation*. They need to understand the logic of the resolution to take action. This might suggest the need for supplying **off-line** transparency, supported perhaps by post-operational analysis techniques when required.

Insights:

- Φ From these takeaways it can be outlined that the transparency provided by the requirements over non-safety critical use cases can apply under the **on-line** schema, while safety critical use cases should take place under an **off-line** schema of transparency.

- Regarding the levels of transparency:
  - ✓ In *non-safety critical* scenarios, transparency should be generally provided **by default**. This might include the option to disable or enable (on demand) the transparency, if requested by the human operator.
  - ✓ In *safety critical* scenarios, transparency should be generally provided **by default** (in -pre operations), and **on demand** during and -post operations.

Insights:

- Φ To comply with this, general transparency requirements might contain an explicit reference to an identifier indicating the level of transparency in which it applies (No transparency: **NT**; On demand: **OD**; By default: **BD**). For example:
    - REQ-ATFCM-T001-***BD***. The FMP shall be provided with the most effective DCB measure (or combination of DCB measures) to solve the hotspot.

Figure 3 summarizes these decisions in a more intuitive way.

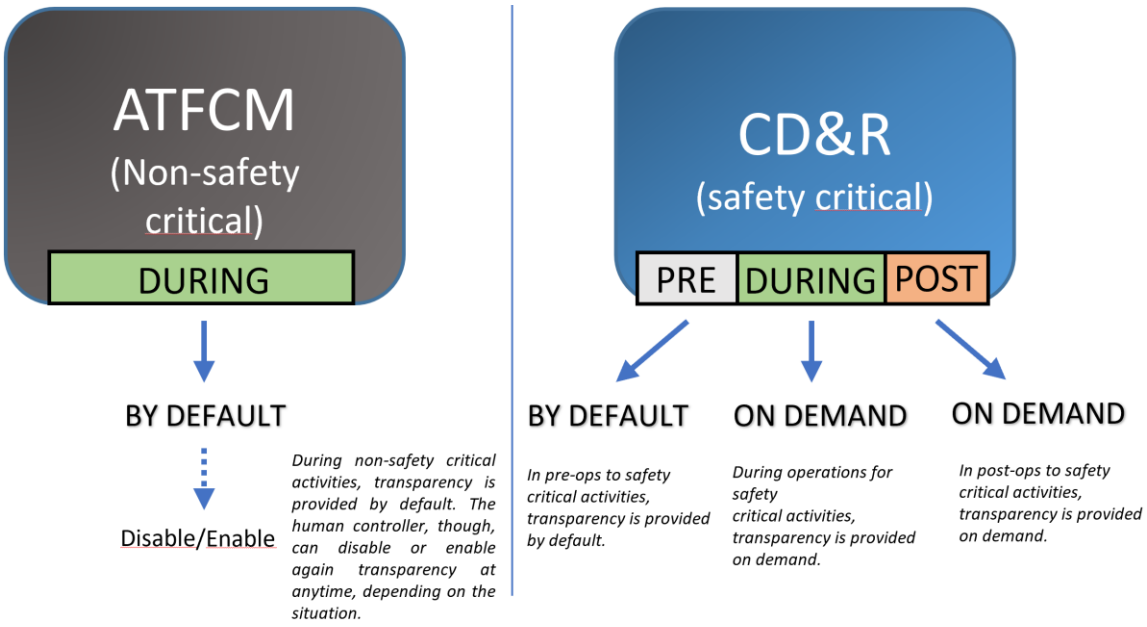

Figure 3. Levels of transparency proposed for the ATFCM and CD&R use cases

5.2 On the Impact of Transparency Quality on Human Efficiency

In addition to the mere fact of providing transparency and explainability in the solutions, it is extremely important to pay attention to the *quality* of such explanations.

However, a current a bottleneck is the fact that there are many unknowns about what the term *quality* really means in transparent and explainable AI. To make it worse, quality can be interpreted in multiple ways depending on the domain of application. Section **¡Error! No se encuentra el origen de la r eferencia.** attempts to provide some new ideas on how to start *defining* and *measuring* the quality of transparency and explainability in the solutions provided by an AI system in ATM.

Regardless what the term quality means to us, it is surely expected that such quality in the provided transparency will have a direct impact on operational efficiency, from a human factors point of view. In this sense, there are several questions and approaches to consider, that could provide interesting

insights into the most important aspects when analyzing the potential impact of a varying quality in the transparency provided to humans.

- **Variable cognitive workload**. It could happen that the level in the quality of transparency and explanations provided to the users is not of sufficient (in terms of timing and accuracy), which could result in two situations:

  1. The users consider that the information provided is not useful or relevant, and therefore discards such information. This could lead to a significant ***loss of trust*** in the AI system, which could mean that the system is simply not used, or is used without considering the full range of capabilities that it is able to offer.

  2. The users have to discriminate, manually, which parts of the information provided are relevant and which are not. This could result in a significant ***increase in their cognitive workload*** with respect to the current operation, since they would be creating 'a new task' to solve. This, in turn, could cause other tasks to be neglected or put on the back burner with a consequent impact.

- **Human error.** TAPAS is about providing a decision support tool at different levels of automation. This means that the quality of the explanations/transparency provided to the user will have a direct impact on the decision to be made (when the actions are not fully automated: Level 2), as well as on the supervision capability (when the system not only proposes, but also decides: Level 3). This impact will possibly have an impact on the final probability that the human will make a mistake when interpreting the information provided by the AI system. An interesting case to analyze would be, for Level 3 of automation, when the human only supervises the actions initiated by the machine. If the explanations are not of sufficient quality, the human could interpret the system's decision as incorrect and take manual control. This transition from automatic to manual, and vice versa, is certainly *susceptible to human error*. Some immediate questions would be:

  - *Should we label it as human error to switch from automatic to manual as a result of poor-quality explanations?*
  - *What consequences on the system could this action have? Can we quantify them?*
  - *In addition to the potential errors associated with the transition, is it correct to regain control?*
  - *Is this an error whose consequences are not being properly assessed because the information presented is not accurate or of high quality?*

  It is important to note that, although these new systems will (like the current ones) need to have a specific mode of operation and procedures that should *mitigate human error* in these cases, these factors will need to be analyzed in detail and in depth during future validation activities.

- **Situational awareness**. The quality of explanations is expected to have a direct impact on the human's situational awareness. By the mere fact of providing explanations, the situational awareness associated with the decision-making process in an automated environment should be improved. A priori, two interesting case studies are identified:

  1. The quality of transparency/explanations is not good enough and *the user is aware* of such limitations. In this case, the impact is likely to be limited to a reduction in situational awareness of the process (in some cases, trust may have been previously gained and these situations may

even be accepted, given that the safety case is positive). Therefore, in this situation, there would be an impact on the human factor, but the efficiency of the decisions, if the system works correctly, would not be affected to a great extent.

2. The quality of transparency/explanations is not good enough *and the user is not aware* of it. Here a deeper and more dangerous degradation of situational awareness could be generated. This would certainly lead to an increase in the probability of human error and impact on safety.

6 Extracting Insights from the Experiments

This section contains important conclusions, findings and insights derived from the results of the validation activities corresponding to the ATFCM and CD&R experiments. These results are mostly focused on transparency and explainability aspects regarding the interaction of experts with the XAI prototype developed for the ATFCM and CD&R use cases.

TAPAS experiments were designed in two selected ATM/ATC domains – the ATFCM domain and Air Traffic Control (ATC) domain. These experiments were executed in two distinct sets of validation experiments in June 2021 (ATFCM) and March 2022 (CD&R). The results corresponding to both experiments were collected and presented in the deliverable called 'D5.2 TAPAS Validation report' [4]. In the experiments, different Human-in-the-Loop Simulations (HITL) were performed in ATFCM and CD&R scenarios that involved the participation of operational experts and their interaction with the implemented XAI/VA prototypes. The overall aim of these tests was to experimentally assess the human-machine interaction in order to validate and extract conclusions that can be useful to derive principles on the transparency and explainability at three different levels of automation (level 1, 2 and 3).

This chapter is organized in the following way. There are three main separate sections containing a brief description of the experiments for the ATFCM (6.1) and CD&R (6.2) scenarios, and some *general conclusions* and *results* extracted from both scenarios (6.3). For each of the different use cases, a first section (6.1.1 and 6.2.1) introduces the *context* in which the experiments were carried out, highlighting aspects the human operators were required to focus during the execution of the experiments. The second section (6.1.2 and 6.2.2) presents the proposed validation scenarios as well as the different *exercises*, at different automation levels, to be conducted along the three days of experiments. The third section (6.1.3 and 6.2.3) presents the different *validation objectives* with diverse focuses and success criteria that were used to evaluate the results of the experiments.

The approach followed is similar to previous sections. The main idea is to identify and extract relevant insights from the experiments, which are expected to later contribute to the main content of the presented principles and recommendations.

6.1 ATFCM experiments

6.1.1 Context

The TAPAS research aims to demonstrate how AI based Decision Support Tools (DST) can be used to assist operators in the execution of Air Traffic Flow and Capacity Management (ATFCM) and Air Traffic Control (ATC) tasks in a manner that can be understood by a human operator. By ‘understood’ it is meant that the human can ultimately unveil some of the 'why's' and 'how's' related to the functioning of the automation processes behind the provided solutions. This way, human understanding about how and why a phenomenon works in a particular way is increased when complemented with explanations, and the learning processes are largely enhanced by examples using visual representations. This fact is key for the human to build trust during his/her collaboration with the machine and the TAPAS research focuses on this objective.

To execute the validation experiments, XAI and VA prototypes were integrated in the INNOVE platform - a human in the loop simulation and gaming platform that simulates the majority of the NM B2B services used in the Network Planning process. Also, a prototype FMP client working position was developed to allow operators to perform Demand-Capacity Management activities in real-time.

A set of validation experiments were performed that included prototype DST working at varying levels of automation, ranging from the provision of advisory information to the automated execution of actions identified by the DST. These tools were accompanied by a dedicated set of transparency tools, which provided interactive VA and explanatory information, designed to help the human operator to understand the proposed decisions.

During the execution of the experiments, human operators were required to:

- Monitor and manage air traffic demand against the available capacity for all airspace sectors located in the Madrid ACC
- Identify overload periods (*Hotspots*) where demand exceeds the defined capacity thresholds
- Investigate the characteristics of the traffic that contributes to the overloads
- Determine suitable ATFCM solutions that can be applied to mitigate the identified problems
- Apply and test those solutions and assess the results of the mitigation actions

In the case where the XAI components were used to automatically identify Hotspots in the region and optionally propose or automatically implement solutions, the operators we required to:

- Review and validate the Hotspots that were identified by the XAI
- At automation level 2:
    - Select one or more of the proposed solutions and implement them
- At automation level 3:
    - Review the solutions that had been automatically applied by the platform
- Review the explanations provided and the associated Visual Analytics which help to understand the choices and solutions being proposed
- Report on their level of understanding of the solutions and their confidence that those solutions have been based on reliable reasoning

6.1.2 Validation Scenarios and Exercises

Three different scenarios, spread over three days of experiments, were considered based on the levels of automation tackled by the project, called automation level 1, 2 and 3. For each scenario, several runs were executed using different traffic dates and studied sectors.

This resulted in a total of eight individual exercises that were performed using the platform. The first two exercises were executed with a heavy focus on *training* purposes, to ensure that the FMP experts could acquire sufficient knowledge and familiarity with the DSTs and the platform that was used during the different sessions. Interim sessions were aimed at enhancing the FMP familiarity with the platform and the associated DST running at different levels of automation.

The final day of the experiment included execution of the platform at each of the levels of automation to allow the FMP users to operate in scenarios with increasing levels of automation, and since time

permitted, included a second scenario running at full automation to allow the FMP to delve deeper into the VA support component and its capabilities. During the exercise, facilitators provided interactive training and advice as well as performing observations on how the FMP were interacting with the various tools. A summary of the experiments is presented next:

- Day 1.
    - **Exercise 1, Automation Level 1**: Initial exercise for 2 FMP users working manually using the FMP client application. The exercise was primarily aimed at the provision of training and platform familiarisation.
    - **Exercise 2, Automation Level 2**: Initial exercise for 2 FMP users working with the FMP client application and supported by XAI/VA prototypes to automatically detect Hotspots, provide suggestions for potential solutions and offer explanations supported by the VA application on a co-located display. The exercise was primarily aimed at the provision of training and familiarisation with the XAI and VA components.
- Day 2.
    - **Exercise 3, Automation Level 1**: Interim exercise for a single FMP user working manually using the FMP client application. The exercised aimed at improving familiarity with the platform to help gain more confidence in its use to identify solve DCB problem scenarios using Regulation and/or Flight Level Capping measures as well as the gathering of results and FMP user feedback.
    - **Exercise 4, Automation Level 3:** Initial exercise for a single FMP user working with the FMP client application and full automation from the XAI prototype to automatically detect Hotspots and implement solutions without human intervention. The VA application was available on a co-located display and could be interrogated as needed by the FMP user. The exercise was primarily aimed at the provision of training and platform / VA-component familiarisation in full automation mode.
- Day 3.
    - **Exercise 5, Automation Level 1**: Additional exercise for a single FMP user working manually using the FMP client application. The exercised provided an additional opportunity for users to work manually with the platform to continue to gain more confidence in its use to identify and solve DCB problem scenarios using Regulation and/or Flight Level Capping measures as well as the gathering of results and FMP user feedback.
    - **Exercise 6, Automation Level 2:** Exercise for a single FMP user working with the FMP client application and partial automation from the XAI prototype to automatically detect Hotspots and provide proposals for potential solutions. The VA application was available on a co-located display and could be interrogated as needed by the FMP user. The exercise was a repeat of the initial exercise, performed previously on day 1, but with a higher level of automation, with the aim of acquiring more feedback from the FMP user.

- **Exercise 7, Automation Level 3:** Additional exercise for a single FMP user working with the FMP client application and full automation from the XAI prototype to automatically detect Hotspots and implement solutions without human intervention. The VA application was available on a co-located display and could be interrogated as needed by the FMP user. The exercise was aimed at enhancing FMP user familiarity with the fully automated scenario as well as in the interactive use of the VA-component to help understand the solutions that had been implemented. FMP user feedback was gathered including information on workload, situational awareness, levels of trust and confidence in the solutions.

- **Exercise 8, Automation Level 3:** Additional scenario running with full automation from the XAI component and the FMP user interacting with the VA component to investigate and to help understand how and why solutions had been implemented by the DST. The exercise was aimed at further enhancing FMP user familiarity with the fully automated scenario as well as in the interactive use of the VA-component to help understand the solutions that had been implemented. FMP user feedback was gathered including information on workload, situational awareness, levels of trust and confidence in the solutions.

For further information and details about the considered scenarios, see *D5.2 TAPAS Validation report*, Section 3.2.4.1 [4].

6.1.3 Validation Objectives and Results

In order to accurately evaluate the results of the conducted validation activities, different validation objectives with diverse focus and success criteria were used. These were specifically designed to test different and important aspects, such as: i) *the identification of potential principles for transparency when AI-based solutions are being used to support the decision-making process*; and ii) *to develop and deploy XAI automation tools with varying levels of automation being emulated to help identify and solve demand-capacity issues and propose suitable actions using either Regulations, Flight Level Cap measures or a combination of both*.

In general, the aim of the experiments was to test and evaluate to what extent these objectives were fulfilled. In order to reach a verdict on the fulfilment of the different criteria corresponding to each of the objectives, three methods were proposed and applied:

- '**Over the shoulder**' observation and discussion during the execution of each validation exercise
- **Debriefing sessions** held immediately after each exercise
- Completion of **scoring questionnaires** for each exercise.

These methods were applied to recognized FMP experts and their answers and feedback were properly collected. Afterwards, the authors of the experiments and validation activities went through a complete and detailed analysis, interpretation and discussion of the results obtained along the execution of the eight exercises, by highlighting differences in the FMP's answers for automation levels

2 and 3, and well as extracting valuable insights and findings, to end up drawing pertinent conclusions on explainability, transparency, user trust/confidence, technical feasibility, and performance assessment.

In addition, the authors also indicated deviations respect to the initial TAPAS D5.1 Validation Plan, as well as some limitations of the validation results in terms of quality and significance. For further information and a more detailed breakdown about the evaluation criteria used in the validation, see *D5.2 TAPAS Validation report*, Section 4.1 [4].

Appendix A (Figure 7) presents a table that summarizes the aforementioned validation objectives, with considered the sub-focus, the corresponding success criteria and the final results that were achieved during the experiments.

6.1.4 Insights, Findings and Conclusions

This section presents a review of the main results of the validation, from the point of view of the different validation objectives and sub-focuses defined in [4] and the presented conclusions. This way, the main objective here is to further spot relevant information contained in the validation results, to condense and squeeze it in order to eventually present it in an intuitive and simple way to the reader. This facilitates the identification and extraction of relevant insights and findings, which are expected to later contribute to the content and form of the principles and recommendations.

It is worth mentioning that these insights are, sometimes, a combination of the information explicitly contained in [4] with further subjective elaborations, merging of concepts and some rewordings of this document's author.

6.1.4.1 Identifying principles for Transparency of AI-based solutions

- **Focus**:

  1. *Determining of how much additional information is needed at automation levels 2 & 3 to ensure that the human operator is able to make informed decisions to help solve ATM problems*.

- The considered success **criterion** was:

  ✓ VA and explanatory support information is clear and understandable and the tools are able to provide the required information at the right time.

- **Result**: Partially OK.

- **Extracted insights**:

  Φ The use of displays showing information via VA tools suggest *direct*, *easy* and *timely* access to different levels of information. This certainly helps human operators maintain situational awareness and contributes to decision making in order to resolve ATM problems.

- Φ Progressive training and familiarization with the XAI tool to be used leads to an *improvement in the* situational awareness of the human operator.

- **Focus**:

  2. *Identifying when support information is required, what level of detail is needed and how should it be provided in a timely manner*.

- The considered success **criterion** was:

  ✓ Key data that can be easily understood by the human has been identified that supports transparency needs and is provided in the required time frame and at an appropriate frequency.

- **Result**: Partially OK.

- **Extracted insights:**

  Φ The effort to scan VA displays by FMPs is generally considered *low*.

  Φ *Simplicity* and *clarity* are key factors when providing explanations.

  Φ An increasing interaction of the human with the offered tools *does improve* the understanding of the explanations and the information provided. Similarly, the effort to gather and interpret information is high initially, but this improves *as users became more familiar* with the XAI / VA tools.

- **Focus**:

  3. Evaluating areas where the levels of transparency may need to be improved.

- The considered success **criterion** was:

  ✓ Information that is unavailable but could help during the use of the proposed XAI has been identified and catalogued for future analysis.

- **Result:** OK.

- **Extracted insights:**

  Φ In some cases, the access to information is *obscure* and must be *improved*.

  Φ The provision or visualization of *aggregated information* stands out as a future promising mechanism to describe and highlight the overall effect and impact of the proposed measures on the overall system when assessing the effectiveness of the XAI-based solutions.

- **Focus**:

  4. *Proposing suitable methods by which the level of understanding and trust in the AI automation can be measured.*

- The considered success **criterion** was:

  ✓ Questionnaires, 'over-the-shoulder' observation and debriefing analysis metrics have been identified to support the necessary measures.

- **Result:** OK.

- **Extracted insights:**

  Φ It is observed that, in general, *the levels of trust progressively improve* as human users became more familiar and practise with the provided tools.

6.1.4.2 Developing prototype XAI/VA methods for ATM use cases to address transparency at various levels of automation

- **Focus:**

  1. *Producing customised VA views to support transparency and explanatory information to the human operator at different levels of automation*.

- The considered success **criterion** was:

  ✓ VA display tools are able to consume data provided by the XAI component to support interactive drill down views for the human operator.

- **Result:** OK.

- **Extracted insights:**

  Φ The use of dedicated displays to present information through multiple customized views using VA tools can *support* and *stimulate* the reception of transparency and explicability at different levels of automation. Having *multiple views* on a display is proved to facilitate the consultation and interactive acquisition of highly detailed information and explanations, at any time, as well as helping human operators to understand the issues that had been identified and the reasons behind the proposed solutions. For example, these multiples views can contain:

    - General views showing information about the overall state of the sectors in terms of traffic demand, highlighting the detected hotspots by the XAI algorithm.
    - Specific views to analyse a certain sector, through which the user can see the flights crossing that sector in a 2D representation and analyse the declared hotspot and flights involved during that time.

- Another series of views exclusively to present to the user the explanations giving by the XAI algorithm on the different solutions.

Φ Some solutions are *difficult to understand* due to the way in which XAI operates by proposing solutions to solve all in one go, rather than a traditional paradigm which used to solve issues one by one.

- **Focus**:

2. *Assessing how the VA methods can help enhance operator understanding and trust in AI-based automation*.

- The considered success **criterion** was:

✓ Elements provided in the VA provide clear visual evidence related to the actions being performed by the XAI tools.

- **Result:** Partially OK.
- **Extracted insights:**

Φ The solutions and reasons proposed are slightly difficult for the user to understand as they are *too complex* and *disruptive* with the current operating method the FMP follows. This has impact on level 2 runs, where the FMP must decide whether or not to implement the proposed solutions. This is because the way in which XAI operates is by proposing solutions to solve all in one go, rather than a traditional paradigm which used to solve issues one by one.

- **Focus:**

3. *Evaluating the effectiveness of the transparency solutions being deployed*.

- The considered success **criterion** was:

✓ Human operators are able to use the visualisation to interrogate the on-going scenario and solutions being considered.

- **Result:** OK.
- **Extracted insights:**

Φ The combination of the XAI prototype, VA display and FMP Client interface stand up as *relevant and efficient providers* of solutions involving transparency and explainability aspects. This original way of deploying solutions allows the users to interrogate the scenario and investigate the proposed solutions in an intuitive manner, depending on the ongoing action. Also, it is highlighted that the explanations provided by the XAI and displayed in the VA prototype should be eventually integrated into the FMP local tool. Two main ways of evaluating the effectiveness of the deployed transparency solutions, depending on the level of automation, are provided:

- **Level 2**. *What-if simulation*. This functionality represents an accurate means for the human operator to understand, evaluate and foresee the *overall impact on the system if the proposed measures were applied*. In addition to this functionality, the VA display also provided the FMP with views through which to consult the different solutions and the reasons behind those.
- **Level 3.** This level provides the user the option to interrogate the tool to further analyse and extensively explore the reasons about why some solutions have been automatically implemented.

- **Focus**:

4. *Determining the different needs for transparency at different automation levels*.

- The considered success **criterion** was:

✓ Human operators classify the information being provided and confirm that it is sufficient to explain the decisions being made.

Result: Partially OK.

Extracted insights:

Φ At levels 2 and 3 of automation, human operators agree that the provided transparency information is *useful*, *understandable,* and *easy* to access. At the same time, they experienced issues to classify such information and confirmed that *further details on transparency and explainability are required* to completely justify the decisions being made. More specifically, the supply of more data, *aggregated in diverse ways*, could largely enhance the interpretation of the information being presented. This can include statistics about the number of impacted flights by the solutions being proposed, most impacted airlines, minutes of delay saved, etc.

- **Focus**:

5. *Evaluating the level of understanding and situational awareness of the human as the automation proposes / implements solutions*.

- The considered success **criterion** was:

✓ Human operators are able to describe what the automation is doing and why solutions have been proposed.

- **Result:** OK.
- **Extracted insights:**

Regarding the human capacity to understand why an automation has proposed solutions, or even when it implements them automatically, it is outlined that:

- Φ *Operators are generally capable of maintaining a controlled level of SA*, as well as of describing what the automation consists of, or why certain solutions have been proposed. This remains valid for levels 2 and 3 of automation. However, and at the same time, there are also *difficulties*. One of them rises when tackling the problem as a whole, rather than by individual problems. Due to this, operators are *sometimes surprised by unexpected events*, which they do not expect according to their experience. They may also *fail to detect* some problems displayed on the screen, and they even *forget things* as they are used to focusing on problems individually, one after another. In addition, operators do not pronounce on their ability to keep things under control. This is especially evident for level 3, where automation has the capacity to implement solutions and the operator supervise.

- **Focus**:

  6. *Verifying that the human can successfully take over and recover control of the situation if the automation fails for any reason*.

- The considered success **criterion** was:

  ✓ The human was able to either take over and complete the current task when automation failed.

- **Result**: Experiment not conducted.

- **Focus**:

  7. Ensure that the human is able to identify and resolve any remaining issues at the end of the XAI process, if present.

- The considered success **criterion** was:

  ✓ *The human operator was able to identify and to complete any remaining issues that were not successfully solved at the end of the process*.

- **Result**: Experiment not conducted.

- **Focus**:

  8. *Demonstrating how transparency can promote operational and social acceptance of 'black-box' AI solutions*.

- The considered success **criterion** was:

  ✓ The operator confirms that the solutions provided by the XAI were fit for purpose.

- **Result:** Partially OK.

- **Extracted insights:**

  - Φ Human operators indicate that, in general, the solutions provided by the automation which contain certain levels of transparency and explainability *do not seem to be sufficient* in terms of *usefulness*, *reliability*, *accuracy* and *comprehensibility*.

  - Φ At the same time, an improvement in the understanding of the tools and the provided solutions is detected as *more time* is spent interacting with them, thus achieving a greater familiarization.

  - Φ FMPs do not need *long* and *intricate* explanations, sometimes *they do not even need explanations during the operational phase*. A higher level of explanations should be left for the training and tool familiarization stage. FMPs do need that the solutions are *highly effective* and, most importantly, to be able to understand the impact they will have on the overall system. This is especially necessary at level 2 automation, which allows the operator to make the decision to initiate the action. In this sense, the what-if functionality is recognized as very useful. However, there is a need to provide the FMP with more aggregated information regarding the impact of solutions, which allows the FMP to more effectively and intuitively assess the appropriateness of solutions.

  - Φ There is general agreement on that *greater acceptance and confidence* in the solutions could be achieved if the solutions that were implemented by the automation were reasonable across all airspace users. Furthermore, it is suggested that the new way of solving all problems at one go could be accepted as a new working paradigm if it is demonstrated that the algorithm does not contain bias.

  - Φ At level 2, there exist inconsistencies between the XAI paradigm (solve all issues at one go) and the selection and implementation of specific solutions, which is detrimental for trust as well. A suggested improvement is to offer all the solutions provided at once to the user.

  - Φ As highlighted in the requirements verification, a factor that would increase confidence in decision making is to *offer more than one solution* to the FMP in order for him/her to compare solutions based on their potential impact and thus be able to choose the most suitable one.

- **Focus:**

  9. *Assessing shortfalls and areas where transparency can be improved in future solutions*.

- The considered success **criterion** was:

  - ✓ Operational experts identify areas where information was insufficient to support understanding.

- **Result:** Partially OK.

- **Extracted insights:**

Operational experts have identified areas where further information related to transparency is required to support and enhance understanding. Particularly, in level 3, where automation proposes

solutions and implements some of them with human supervision, the operational experts were able to review the Hotspots that the XAI identified with relative ease, together with the solutions that were included in the FMP tool and their explanations in the VA display. However:

- Φ *Some pieces of information are difficult to understand*, some explanation features are no auto explanatory and more focused on how the XAI algorithm worked.

- **Focus:**

  10. *Identifying opportunities for additional training*.

- The considered success **criterion** was:

  - ✓ Additional training or processes to enhance the ability for the XAI/VA to assist the human in understanding the process at different automation levels has been identified by the team.

- **Result:** OK.

- **Extracted insights:**

Following the completion of the various exercises, the analysis team performed a general review of the design, execution and results of the validation scenarios and results. Key points were presented to the TAPAS project partners to provide feedback on Explainability, Tooling and Lessons Learned.

6.2 CD&R experiments

6.2.1 Context

For the CD&R validation experiments, the XAI and VA prototypes were connected to the ENAIRE/CRIDA SACTA real-time ATC simulator platform, from where they consumed the necessary data to detect conflicts, elaborate solutions and present all this information through visual aids to the air traffic controllers.

SACTA is the system that manages ATC in all en-route, approach, and terminal centres in Spain. A replica of this ATC platform, installed at the CRIDA premises, was used as a human in the loop simulation platform that both manages how aircraft operate in one or more ATC sectors and which provides a highly realistic ATC Controller Working Position (CWP). The CWP provides all the same features as the actual ATC system used to manage the Spanish airspace. Flight trajectories are calculated using the SACTA flight models and control of flights in the region is performed by support staff operating via a pseudo pilot interface. In addition, to emulate fully automated execution of conflict resolution clearances proposed by the AI tool, a 'ghost controller' is included in the platform who can implement the changes being recommended by the DST without the need for the ATC user to intervene.

Using the CWP and other SACTA tools, users can perform all the functions that are carried out by both the Planner and Executive/Radar Controller for the selected airspace in real-time, including any

communication that is required with pseudo aircraft pilots that are also connected to the simulator platform.

During the execution of the CD&R experimental scenarios, operators were required to:

- Monitor and manage air traffic as it flies through one of two upper airspace sectors (Toledo and/or Domingo upper) located in the Madrid ACC
- Acquire and hand-off flights as they arrive into / depart from the managed sector
- Observe the radar picture and flight plans as traffic progresses through the region
- Monitor aircraft-aircraft separation in the controlled sector and the bordering region
- Identify potential separation issues (with the help of the conflict alerting tool provided by the AI component) and prevent collisions between aircraft in flight
- Identify of potential solutions to predicted separation issues, with or without the help of the AI automation tool
- Provide of instructions/clearances to traffic to avoid separation losses
- Monitor the traffic compliance to proposed flight plans and any separation management instructions provided (manually, using proposed solutions and/or automatically)
- Provide of instructions/clearances to allow traffic to recover its original plan following separation management actions towards, or at, the sector exit / transfer point
- Select and execute efficient solutions to traffic separation issues, e.g. through the use of direct to solutions or manoeuvres that are as efficient as possible
- Timely deliver conflict avoidance instructions to ensure the safe and efficient conduct of flight operations in the region
- Select and implement the most suitable solution being recommended by the AI DST (automation level 2)
- Monitor and understand the solutions that have been automatically implemented by the AI DST and recovery of control whenever the automatic XAI system fails partially or completely to resolve conflicts (automation level 3)

6.2.2 Validation Scenarios and Exercises

Three different scenarios, spread over three days of experiments, were considered based on the three levels of automation. For each scenario, several runs were executed using different traffic dates densities, and studied sectors.

The validation took place over a period of three days, with exercises on Day 1 focused on the provision of *training* to the users as well as to ensure they became familiar with the various XAI and VA solutions provided as part of the validation process. Days 2 and 3 were used to run the full validation scenarios and to assess the effect on *human performance* and *understanding* when using the XAI and VA tools at different levels of automation. One of the two selected sectors from the Madrid ACC – Toledo Upper (TLU) and Domingo Upper (DGU) was used as the measured sector in each scenario.

This resulted in a total of *ten* individual exercises that were performed using the platform. This way, the planning included 3 exercises executed *at Level 1 automation* (1 scenario with 2 ATCo managing the sector and 2 scenarios with a single ATCO, resulting in 4 sets of questionnaire responses); 5 exercises at *Level 2 automation* (2 scenarios with 2 ATCO experts managing the airspace and 3 scenarios with a single ATCo, resulting in 7 questionnaire responses); and 2 exercises executed at *Level*

3 (including unexpected degradation of the automation) both with only a single ATCo monitoring the scenario resulting in 2 questionnaire responses. A summary of the experiments is presented next:

- Day 1.
    - **Exercise 1, TS1.1 – TLU, Automation Level 1, Low density traffic**: The TS1.1 scenario provided a baseline reference scenario against which other results could be compared. TS1.1, executed at the start of the first day was used as a training and familiarisation exercise. 2 ATCo worked with the system providing Radar/Executive controller functions for the TLU sector between FL345 and UNL. A low-density traffic sample was used to allow the users to familiarise themselves with the working position and available tools provided by the SACTA simulator platform. Conflict alerts were provided by the XAI DST component. TS1 traffic samples presented low complexity, low traffic demand and a low number of conflicts (2 conflicts per 15 minutes). Alerts were provided via the connected VA support display which was co-located on a small screen placed next to the main Radar view. This allowed the users to familiarise with the information being provided in relation to any conflict that was detected by the tools. All the resolution decisions and resulting actions were made and implemented by the human in this scenario without help from the AI tools.
    - **Exercise 2, TS2.1 – DGU, Automation Level 2, Medium to Low density traffic**: The TS2.1 scenario was executed for the DGU sector with partial automation support being provided by the XAI tools with the 2 ATCo providing support. Traffic in the TS2 level samples presented medium to low complexity, medium traffic demand (with average OCC of more than 5 flights and less than 10 flights in windows of 5 minutes). A low number of conflicts (3 conflicts per 15 minutes) was included in the traffic sample. Users received conflict alerts and a set of proposed solutions that had been determined by the XAI tool. Using the VA support, the users could review the actions proposed and decide to apply one or more of them to try to solve the conflict. Actions were based on clearances for one or other of the flights involved and were prioritised to offer users an idea of which may be the most effective clearance(s) to attempt. Following the issue of the clearance(s) the users were responsible for monitoring traffic to ensure it complied with the instructions given. The monitoring process was also supported by the XAI/VA components which provided visual alerts if traffic was identified to be ‘off-track’ or non-compliant with the instruction.
- Day 2.
    - **Exercise 3, TS4 – DGU, Automation Level 1, Medium density traffic**: The TS4-DGU scenario was the first of the measured validation scenarios to be executed for the CD&R study and focused on Radar control at Level 1 automation in the DGU sector. The airspace was managed by 1 ATCo who provided support for Tactical/Radar Control functions via the SACTA CWP. TS4 level traffic presents very high complexity, medium traffic demand (average OCC of more than 5 flights and less than 10 flights in windows of 5 minutes) and a medium number of conflicts (6 conflicts per 15 minutes). The XAI component was used to identify conflicts and to provide alerts which were shown in the associated VA display component. No potential resolution actions were provided, and the user was requested to solve the conflicts that the XAI identified with no additional assistance from the automation components.

- **Exercise 4, TS1.2 – TLU, Automation Level 2, Low density traffic**: TS1.2-TLU was executed with 1 ATCo providing the Tactical/Radar control function on the TLU radar position. TS1 traffic samples presented low complexity, low traffic demand (average OCC of less than 5 flights in windows of 5 minutes) and a low number of conflicts (2 conflicts per 15 minutes). Partial automation support was provided to the user. The automation performed conflict alerting and provided a prioritised list of potential actions for flights that could be used to solve those issues. The ATCo users were able to review the conflicts that were reported using both features available in the CWP and information provided by the co-located VA display tool. At level 2, the user was asked to select and apply solutions from the proposed list but the implementation of any clearances chosen had to be performed by the Radar controller using the facilities available in the SACTA CWP HMI and using voice-based instructions to the pseudo pilot. Conformance monitoring (both against the proposed flight plan and any clearance provided by the Radar controller) was also active in the scenario.

- **Exercise 5, TS2.1 – DGU, Automation Level 3, Medium density traffic**: TS2.1-DGU provided users with a more complex, higher density traffic scenario running with Level 3, full automation. 1 ATCo was responsible for monitoring the DGU sector and intervening if needed. In this scenario, the user was requested to allow the XAI components to perform all the conflict detection, alerting, and to initiate the preferred resolution actions without help from the Radar controller. The human monitored the scenario, looking at why conflicts had been identified, and assessed how and why the proposed solutions were chosen as well as their suitability in solving those issues. The user had to maintain a good situational awareness as well as to monitor the conformance of traffic to the automated resolution clearances – with the assistance of the conformance monitoring functionality in the XAI components. In addition, the user was requested to intervene and recover the situation when a sudden and unexpected failure of the automation occurred. This included identifying conflicts which were not captured by the automation, or which were captured but not solved. The scenario used TS2 level traffic samples which present low to medium complexity and medium traffic demand (average OCC of more than 5 flights and less than 10 flights in windows of 5 minutes) and a medium number of conflicts (3 conflicts per 15 mins). For these traffic samples morning hours were selected (from 8:00 to 8:30) from the 4th of July 2019.

- **Exercise 6, TS3 – DGU, Automation Level 2, Medium density traffic**: The TS3 – DGU scenario performed on Day 2 provided users with a more complex, higher density traffic scenario running at level 2, partial automation. The Radar/Executive controller function was provided by 2 ATCo during the simulation exercise. Users were requested to allow the XAI components to perform all the conflict detection and alerting, as well as to provide a set of resolution actions that could be used to solve conflicts using one or more of the recommended clearances. The human role was then both monitor traffic in the scenario using the CWP, looking at why conflicts had been identified, and assessing the proposed solutions to determine which were suitable to help solve those issues. Users were not required to use the highest priority solution(s) and were free to choose other alternatives being suggested, or to implement their own solution(s). Tasks also included identifying conflicts which were not captured by the automation, or which were captured but no suitable solution was offered. To support this task the

users had to maintain a good situational awareness as well as to monitor the conformance of traffic any resolution clearances that they had selected – with the assistance of the conformance monitoring functionality in the XAI components. Traffic for the TL3 scenario presents medium complexity, medium traffic demand (average OCC of more than 5 flights and less than 10 flights in windows of 5 minutes) and a medium number of conflicts (4 conflicts per 15 minutes). For these traffic samples morning hours were selected (from 8:00 to 8:30) from the 25th of June 2019.

- Day 3.
  - **Exercise 7, TS4 – DGU, Automation Level 1, Medium density traffic**: The TS4-DGU scenario performed on Day 3 of the validation was a repeat of the same scenario from Day 2 with a different Radar controller operator performing the exercise. 1 ATCo was responsible for providing Radar/Executive controller functions during the execution of the scenario with limited automation support at Level 1.
  - **Exercise 8, TS1.2 – TLU, Automation Level 2, Low density traffic**: The TS1.2-TLU scenario performed on Day 3 of the validation was a repeat of the same scenario from Day 2 with a different Radar controller operator performing the exercise. 1 ATCo provided support for the Radar/Executive controller function with the assistance of the XAI-based automation and VA information support tool. Suggestions offered by the automation could be selected by the ATCo then appropriate clearances delivered to traffic with automation running at Level 2.
  - **Exercise 9, TS3 – DGU, Automation Level 3, Medium density traffic**: The TS3-DGU scenario performed on Day 3 of the validation was a repeat of the same scenario from Day 2 with a different Radar controller operator performing the exercise. 1 ATCo was responsible for monitoring conflicts identified by the automation and the solutions that were being automatically applied to try to solve them. The user was also requested to indicate when conflicts were not correctly identified and/or if solutions were inappropriate or did not solve the problems.
  - **Exercise 10, TS2.1 – TLU, Automation Level 2, Medium to low density traffic**: The TS2.1-TLU scenario performed on Day 3 of the validation was a repeat of the same scenario from Day 2 with a different Radar controller operator performing the exercise.

For further information and details about the considered scenarios, see *TAPAS D5.2 TAPAS Validation report*, Section 3.2.4.2 [4].

6.2.3 Validation Objectives and Results

Similar to what was done for the ATFCM experiments, in order to evaluate the results of the conducted validation activities, different validation objectives with diverse focus and success criteria were defined for the CD&R experiments as well.

The validation objectives themselves are roughly the same in the CD&R respect to the ATFCM experiments, with minor additions that are detailed in the next section. This way, such methods were applied to recognized ATCOs and their answers and feedback were properly collected. Later on, the authors of the validation activities went through a complete and detailed analysis, interpretation and discussion of the results obtained along the execution of the ten exercises, by highlighting differences in the ATCOs answers for the different automation levels.

Appendix B (Figure 8) presents a table that summarizes the aforementioned validation objectives, with considered the sub-focus, the corresponding success criteria and the final results that were achieved during the experiments.

6.2.4 Insights, Findings and Conclusions

Following the same procedure done for the ATFCM experiments in previous sections, here we further spot relevant information contained in the validation results for the CD&R experiments.

6.2.4.1 Identifying principles for Transparency of AI-based solutions

- **Focus**:

  1. *To determine how much additional information is needed at automation levels 2 & 3 to ensure that the human operators can make informed decisions to help solve conflicts identified by the system at various levels of automation*.

- The considered success **criterion** was:

  ✓ VA and explanatory support information that is clear and understandable is provided in a short timeframe and the tools provide the required information to allow the user to rapidly understand the situation being managed and context of the proposed solution.

- **Result**: OK.

- **Extracted insights**:

  Φ Information was readily available and easy to access and that the information was *clear* and able to be understood in a *timely manner.*

  Φ The information provided a clear vision of the conflict resolution actions being proposed, they were able to *easily identify all the conflicts* and were not disturbed or overloaded by too much information.

  Φ Users indicated that they had a *good global overview* and were ahead of the traffic and fully capable of planning and organising the work that they needed to do.

  Φ ATCOs, when instructed to use the available information provided in the co-located VA display, *quickly assimilate the information* that is provided for a conflict and use the detailed information to review the solutions being proposed.

- Φ At levels 2 and 3, ATCOs prefer to use existing ATC/traffic monitoring features instead of the provided graphical display of the conflict trajectories.
- Φ Other information related to conflict alerts and proposed actions, when provided via the co-located display, is *useful* and allowed the users to *quickly understand* the conflict and traffic involved, as well as the solutions that were proposed.
- Φ ATCOs seem prone to accept solutions that solve the conflicts (those proposed at Level 2, and also those automatically implemented at Level 3), even though the solutions differ from those that ATCOs would have applied.
- Φ ATCOs tend to question the system's solutions and ask for reasons when conflicts are *partially solved, or unresolved*.
- Φ Due to the short times for conflicts to be identified, solved and instructions given to traffic, offering more information to ATCOs than was already provided will not necessarily change the understanding that they could usually acquire in a very short time due to their own experience and expertise in the domain.
- Φ ATCOs agree that the *level of information provided is sufficient* for their needs in the CD&R use case.
- Φ ATCOs work more efficiently and use more transparency-related information when this is presented via VA displays that are properly integrated a SACTA CWP HMI, instead of an adjacent co-located display. The latter might lead the ACTs to a loss of focus and awareness about the evolving traffic conditions in the sector, especially when traffic loads are high and complex.

- **Focus**:
  2. *To identify when support information is required, what level of detail is needed and how should it be provided in a timely manner*.
- The considered success **criterion** was:
  - ✓ Key data that can be easily understood by the human has been identified that supports transparency needs and is provided in the required time frame and at an *appropriate frequency*. Additional information providing more detailed information that can help explain more complex situations and the decisions that were made is available for consultation by the user in an 'on-demand' mode if required.
- **Result**: OK.
- **Extracted insights:**
  - Φ The *level of detail* of the provided transparency-related information was *useful* and presented in a way that allowed ATCOs to quickly comprehend the situation and understand any actions that were being proposed or implemented.

- Φ When presenting transparency-related information, low refresh data rates can result in *disorientation of the ATCOs*, which in turn might impact decision-making regarding response and delivery of a suitable clearance(s) to solve the identified conflict(s).
- Φ Given the short timeframe available between the identification of conflict situations and the need to identify then deliver clearances to resolve those conflicts, the need for additional 'drill down' information was *low*.
- Φ ATCOs feel that the transparency information provided is enough to understand the situation and the proposed actions. Such information allowed ATCOs to maintain a *good global overview of the situation* and it was provided with sufficient time to allow them to understand the situation, review the options being proposed and select a solution.
- Φ ATCOs consider that low effort is needed to establish a *good understanding* about what the automation was trying to do.

- **Focus**:

  3. *To evaluate areas where the levels of transparency may need to be improved.*

- The considered success **criterion** was:

  ✓ Information that is unavailable but could help during the use of the proposed XAI has been identified and catalogued for future analysis.

- **Result:** OK.

- **Extracted insights:**

  - Φ ATCs are satisfied with the information made available by the XAI/VA prototype.
  - Φ *Information* provided *is easy to use* and complemented the expertise of ATCOs as well as existing SACTA tool suite.
  - Φ No additional information relating to conflicts and the resolution clearances being proposed was required, even if on some occasions the specific actions were not necessarily the action that they would take without the help of the automation tool.
  - Φ It is outlined that, specially at level 3 in which the XAI prototype automatically implements an action, it is useful to keep ATCOs informed in *real-time* about the *status of automation*. This transparency would yield light on if the machine is currently processing an action or if the view has not change yet due to refresh issues.

- **Focus**:

  4. *To propose suitable methods by which the level of understanding and trust in the AI automation can be measured.*

- The considered success **criterion** was:
  - ✓ Questionnaires, 'over-the-shoulder' observation and debriefing analysis metrics have been identified to support the necessary measures.
- **Result:** OK.
- **Extracted insights:**
  - Φ Over the shoulder observation and the use of debriefing and questionnaires at the end of each exercise proved to be very useful, and in general how the scores tended to improve as users became more familiar with the platform.

6.2.4.2 Developing prototype XAI/VA methods for ATM use cases to address transparency at various levels of automation

- **Focus:**
  1. *To produce customised VA views to support transparency and explanatory information to the human operator at different levels of automation*.
- The considered success **criterion** was:
  - ✓ VA display tools are able to consume data provided by the XAI component to support interactive views for the human operator in a timely and concise manner.
- **Result:** OK.
- **Extracted insights:**
  - Φ At level 2, the *usefulness* of the *co-located VA support display is very limited* due to the short times between conflict identification and clearance implementation. ATCOs do not have time to focus on different displays while solving conflicts.
  - Φ At level 2, ATCOs still tended to use the existing functions and features within the SACTA CWP over any new features that were provided by the VA support display.
  - Φ At level 2, a very interesting fact suggested by ATCOs is that the *level of required explanations may depend on the time horizon* on which the proposed conflict resolution operates. That is, if the proposed solutions only solve conflicts in a *short-term horizon*, without looking at the effects at medium or long term, *no explanations are actually needed*. ATCOs tend to know almost all the possible solutions to a specific situation, they seem to control all the possible outcomes by their experience.
  - Φ Another situation in which ATCOs would acknowledge the provision of explanations is in case that the XAI would give a list of actions to 'not execute'. ATCOs manifest that, in this case, offering explanations to *why specific actions should not be executed* would be truly useful.

Φ *ATCOs do not seem to need the additional graphical view of conflict situations that accompanied conflict details in the VA display*. The CWP HMI, combined with the ATCOs' expertise and experience are sufficient to understand all the conflict situations rapidly without the need to consult the additional VA display.

- **Focus**:

  2. *To assess how the VA methods can help enhance operator understanding and trust in AI-based automation*.

- The considered success **criterion** was:

  ✓ Elements provided in the VA provide clear visual evidence related to the actions being performed by the XAI tools.

- **Result:** Partially OK.

- **Extracted insights:**

  Φ ATCOs widely agree on the need to *further improve the XAI solutions' accuracy and variety* to solve the proposed conflicts. The solutions sometimes led to further issues which the ATCOs would have avoid based on their own experience. Because of this, ATCs' *confidence in the XAI solutions is lower* than hoped.

  Φ ATCOs generally seem reluctant to accept proposals to solve conflicts *that look very different to those* that they would have chosen by their experience. It is suggested that if the proposed XAI solutions were more similar to those by the ATCOs a *greater acceptance could be achieved*.

- **Focus:**

  3. *To evaluate the effectiveness of the transparency solutions being deployed*.

- The considered success **criterion** was:

  ✓ Human operators are able to use the visualisation to interrogate the on-going scenario and solutions being considered.

- **Result:** OK.

- **Extracted insights:**

  Φ ATCOs seem to be able to *easily access the information* provided in the VA support to understand solutions being proposed and using features in the CWP they could measure the impact/applicability easily. ATCs seem able *to maintain very high levels of* situational awareness and their level of understanding is high.

- **Focus**:

  4. *To determine the different needs for transparency at different automation levels*.

- The considered success **criterion** was:

  ✓ Human operators classify the information being provided and confirm that it is sufficient to explain the decisions being made. Optional detailed views are able to support more complex situations and can provide additional detailed understanding in an *acceptable timeframe*.

- **Result:** OK.

- **Extracted insights:**

  Φ In the CD&R scenario at level 2, ATCOs feel that the *information* provided *was sufficient* for their purpose and *no additional transparency-related information is necessary*.

  Φ At level 2, whenever the provided solutions are reasonable to the ATCOs, then no additional explanations are required. ATCs are capable of *understanding* conflicts rapidly due to their own experience. ATCOs are able to *evaluate* and *understand* the potential impact of the proposed solutions with relative ease.

  Φ At level 3, some concerns were expressed regarding conflicts not being captured or resolutions being ineffective and/or not solving the conflict. However, this situation does not affect the transparency needs at different automation levels which are considered to be low by ATCOs.

- **Focus**:

  5. *To evaluate the level of understanding and situational awareness of the human as the automation proposes / implements solutions*.

- The considered success **criterion** was:

  ✓ Human operators are able to describe what the automation is doing and why solutions have been proposed.

- **Result:** Partially OK.

- **Extracted insights:**

  Φ At level 2, there is an evident *discrepancy* related to ATCOs fully understanding *why* the conflict resolution actions have been proposed and their *consequences*. There is *no clear evidence* to conclude anything that is significant in this regard.

  Φ More work for improving the technical solutions' accuracy, more thinking of what transparency information has to be given, as well as further training for ATCs to understand the tools is needed.

- Φ At Level 3 the situation gets worst as the system takes control to implement actions. ATCOs *seem to struggle to understand the conflict resolution actions* that have been automatically implemented by the system, as well as their consequences in the traffic.

- Φ A *wrong selection of the automation tools used might impact* these results, as some of them were considered to be unrealistic or inefficient by ATCOs.

- Φ ATCOs generally agreed that even if the resolutions were not ones that they would have chosen themselves, *the information provided by the system was sufficient* to allow them to easily understand the consequence of those actions and to *maintain a good overall situational awareness*.

- **Focus**:

  6. *To verify that the human can successfully take over and recover control of the situation if the automation fails for any reason*.

- The considered success **criterion** was:

  - ✓ The human was able to either take over and complete the current task when automation failed.

- **Result**: Partially OK.

- **Extracted insights:**

  - Φ In a *monitoring role* at level 3, when the automation does not capture all the conflicts or cannot provide a solution, ATCOs seem fully aware of the situation and able to *provide suitable actions that resolve the issue* when requested. Their situational awareness does not degrade.

  - Φ At *monitoring* roles at level 3, ATCOs manifest that the confidence in the tool *increases* if they can stay for hours by simply *looking at the display* and doing *nothing*, just by observing how the automation is able to accurately solve the conflicts without enforcing them to take over.

  - Φ ATCOs agree on if control recovery automation is to be deployed in the real world, there is a risk that over a sustained period, *ATCOs may become 'de-skilled',* with a *significant negative impact to SA,* which could lead to issues when trying to recover following unexpected failures of the automation.

- **Focus**:

  7. *To ensure that the human is able to identify and resolve any remaining issues at the end of the XAI process, if present*.

- The considered success **criterion** was:

- ✓ The human operator was able to identify and to complete any remaining issues that were not successfully solved at the end of the process.

- **Result**: Partially OK.

- **Extracted insights:**

  - Φ At levels 2 and 3, the *XAI automation is not able to detect all the conflicts*, and some of them are *not solved in the most efficient manner* or using 'open loop' manoeuvres.

  - Φ *The XAI is not yet able to provide additional instructions to resume their original plan*. The conformance monitoring feature of the XAI help ATCOs to capture the need to re-establish traffic on its former route, however, the way of processing that information is not ideal. More work to enhance and complete the prototype is needed on this regard.

  - Φ Still, *ATCOs are generally able to detect those unaddressed conflicts*, to *provide* the *appropriate instructions* to the pseudo pilot to solve those conflicts, and even *include additional clearances* to recover the flight plan following a resolution action.

  - Φ ATCOs prefer to receive a straightforward notification of the exact moment when instructions could be given to allow flights to resume their original plan. This can be achieved through the creation and presentation of a set of actions needed via check list.

- **Focus**:

  8. *To demonstrate how transparency can promote operational and social acceptance of 'black-box' AI solutions*.

- The considered success **criterion** was:

  - ✓ The operator confirms that the solutions provided by the XAI were fit for purpose.

- **Result:** Partially OK.

- **Extracted insights:**

  - Φ ATCOs *are generally able to understand and use the solutions* being proposed using the provided transparency-related information, but on some occasions those solutions are not considered suitable to solve the problem.

  - Φ ATCOs indicate that the provided clearances resulted in additional and sometimes more critical downstream conflicts. Similarly, the lack of additional instructions to help traffic to resume its original planned route also led to *a reduction in the confidence* established by the users in the solutions being proposed/implemented by the tool.

  - Φ It is outlined that transparency and explainability for a solution, if this does not accurately solve a problem in first place, cannot be enough to promote operational and social acceptance of an XAI-based automation. Trust building toward XAI tools is significantly degraded when a solution is inefficient, unreliable or incomplete.

- **Focus:**

  9. *To assess shortfalls and areas where transparency can be improved in future solutions*.

- The considered success **criterion** was:

  ✓ Operational experts identify areas where information was insufficient to support understanding.

- **Result:** OK.

- **Extracted insights:**

  Φ In the CD&R scenarios, ATCOs indicate that all the *information* they require *is easily obtainable from the XAI*. Also, the additional information about the conflicts and proposed solutions provided by the tool *is enough to support understanding* of the solutions and actions proposed, the *foreseen consequences* of such solutions, and the *maintaining of good SA*. No further transparency-related information is needed.

  Φ The way the automation conducted the conformance monitoring task was not initially easy to understand. However, once the ATCOs became familiar with the way that this information was provided, they accepted that *no further information is required*.

  Φ In level 3 scenarios, *ATCOs refuse to accept the way the XAI automatically solves conflicts as a natural way of problem resolution*. ATCOs seem to accept solutions *if these are similar* to those they would employ without the help of automation. That is, simply by using their experience and intuition.

- **Focus:**

  10. *To identify opportunities for additional training*.

- The considered success **criterion** was:

  ✓ Additional training or processes to enhance the ability for the XAI/VA to assist the human in understanding the process at different automation levels has been identified by the team.

- **Result:** OK.

- **Extracted insights:**

  Φ At level 2, ATCOs agreed that the *platform and tools are easy to understand*, *quick* to *learn* and they do not need the help of technical support personnel to use.

  Φ At Level 3, *responses are less positive*. ATCOs tend to rationalise the solutions being applied by the automation tools, which was sometimes confusing, and the lack of additional actions that helped flights to return to their original plan. However, *in general they indicated that they were*

happy with the information provided and how the tools can be used. They express that if these issues were rectified, then *little or no additional training would be necessary*.

6.3 General conclusions

In addition to the previous insights, the document [4] also provides general relevant conclusions regarding explainability, transparency, user trust/confidence, technical feasibility, and performance assessment for the ATFCM and CD&R validation experiments. Some of these conclusions are highlighted here, sometimes reworded, reformulated, and further organized for the sake of an optimal understanding on how principles and recommendations are derived at the end of the document, as well as to facilitating the tracking of the content that are part of such principles.

6.3.1 Explainability/concept maturity

- **Human users prefer trusting the system, rather than obtaining explanations.** Through the constant use of the system, especially during the *training phase*, the human actor is able to *develop trust* in the system through how it performed and the solutions it was providing. For example, a booster for *stimulating the human understanding and building of trust* is to see the impact of the solution implemented/proposed by the system before making decisions. This seems to be more valuable than the current explanations provided by the support tools by the users.

- **The time horizon defines the different levels of required explainability/transparency.**

  - **During the operation.** Humans *do not need to see all information or explanations* related to the proposed solutions by the XAI algorithm (it also requires a time that they generally do not have). Humans do not need to see the *intermediate solutions* the XAI algorithm considers and analyses before arriving to the final solution.

  - **During training.** In this phase humans need more explanations on the algorithm and solutions provided by the AI. Once the proposed approach is understood, *humans do no longer need this information*, as they already are instructed.

- **Explanations must contain aggregated information.** When providing explanations, the user wants to see information *in different fashions depending on the criticality* of the ATM scenario.

  - ATFCM scenarios. Users here wants to see information in an *aggregated* way, such as statistics on the impact of implementing the proposed solution, which could also include a breakdown by region, airline operator etc. They also indicated that they would like to be able to go from *a more general level to a particular one,* for instance from hotspot related information to flight related information and vice versa.

  - CD&R scenarios. This is a more safety critical one, and the operator would prefer to see the most relevant information only in a *straightforward manner by default*, as they have a *limited amount of time* to assess and solve the conflicts.

- **In real-time applications, in-depth or intermediate explanations are not always the primary way to go**. Depending on the criticality of the ATM scenario and the available time for action, having explanations to a process or reached solution may degrade human decision making. Although explanations can be beneficial sometimes, in scenarios as CD&R there might not be enough time for users to be able to consult all of them. In general, it is the accuracy of the presented solutions what will lead the user to accept and gain confidence in the tools so that she/he can understand why those solutions are suitable without the need for detailed explanations. Indeed, on many occasions the users repeated that seeing solutions which work, even if they are different to those that the user may have chosen themselves is sufficient to develop trust in the system and to accept how it is performing without the need for more explanations.

- **Complexity of the solutions limits human capacity to understand the real-time explanations and systems supervision.** In cases where the solution is too complex the human will have neither the time nor the ability to understand the solutions. Likewise, in such cases the human capability to conduct an effective oversight of an AI/XAI system might be diminished at some extent.

- **Ranking the best proposed solutions is something valuable**. When multiple solutions are possible for the same issue, a clear ranking of the solutions from best to worst is very valuable to the user. This is particularly important in time-constrained, safety critical situations typically seen in CD&R scenarios.

- **When actions are being automatically performed by the XAI, their current status should be clearly communicated**. For example, whether those actions are pending, in progress or completed.

- **If multiple actions have the same ranking it should be clarified whether both actions must be performed or not**. Sometimes two actions are too many (e.g. *to communicate and monitor*) and other times two actions with less impact to each recipient can be of added value in terms of capacity / efficiency / equity etc.

- **For the safety critical use case of the CD&R, automation level 3 does not seem feasible to implement**. As an example, performing a monitoring task alone may result in ATCO loss of expertise in the controlling tasks and whenever the XAI fails (even though it will supposedly work well most of the times) the ATCOs will not have the capability to recover control in complex situations in a safe manner.

6.3.2 Transparency

- **In ATFCM scenarios, the traceability of explanations is key for transparency.** The user needs, not only to see the final explanation of the solutions but have a clear traceability of the elements related to each measure/solution. They prefer to see aggregated information, but they appreciate the possibility of following the thread of certain solution down to the level of the flights to which it is related. This gives a clear transparency to the solutions or explanations provided, *making it easier for the user to build trust in the system*.

- **In CD&R scenarios, the provision of the solutions by the XAI system, and the corresponding transparency depends on several aspects.**

- Regarding the *moment* and *manner* of providing the solutions, as these are time critical, ATCOs prefer that the solutions are given *rapidly* and in *priority* order if more than one solution is available. In cases where multiple actions are proposed to solve the same conflict, users need clear indication to distinguish these from other cases where a choice of independent solutions is provided.

- About *transparency* needs for those solutions, ATCOs agree on that in CD&R scenarios the importance lies on providing *solutions that work well and are accurate*, rather than focusing on explanations. This is sufficient for ATCOs to accept those proposals. ATCOs consider that little or no additional explanatory information is needed since the combination of information already provided (usually linked to conflict characteristics) combined with a prioritisation of choices is sufficient to allow them to rapidly understand the proposals being made/implemented and the consequence of those actions.

- The solutions provided by the XAI system are *incomplete*, as the conflict resolution process must also *include clearances* to allow the traffic to *recover its original flight plan*. Nevertheless, the information provided via the VA display was considered to be sufficient to promote good levels of transparency.

6.3.3 Trust and confidence building

Φ **In ATFCM and CD&R scenarios, trust and confidence must be gained mostly during the training phase.** The human needs to see that the solutions proposed/implemented by the XAI system are efficient in terms of the impact that those solutions imposed on the traffic (number of delays, hotspots solved, etc.).

Φ **Automation factors that degrade trust in CD&R scenarios.** Different aspects, such as a combination of *unrealistic* (from the human way of thinking perspective) *solutions*, or solutions that may have led to *more complex issues* further downstream, along with the *lack of additional actions* that are also considered to be an integral and necessary part of the conflict resolution process may contribute to *a reduction in trust and confidence in the automation*.

Φ **Confidence and trust can be volatile.** Developing trust and confidence in a XAI system does take a long time and also relies on that system providing reliable solutions that the user accepts as being a valid response to a problem. In the event where something subsequently fails badly, even after trust has been achieved, that confidence in the system *can be lost very rapidly* and rebuilding it can be hard. This is especially critical in the CD&R domain, and therefore heavy focus on the reliability and suitability of solutions being proposed must be paid.

Φ **Disruptive solutions must allow reproducibility.** When dealing with disruptive solutions proposed by the XAI system, it is important that those solutions do not vary too much when tackling the same type of problem. This also applies to solutions that are not necessarily disruptive, in which the input-output space should be reasonable for all the proposed solutions. This certainly promotes *confidence in the system*. In this way the human can create a mental pattern on how the system works, thus *improving understanding and building trust* on the system more quickly and robustly.

6.3.4 Technical feasibility

- **Algorithms must prove unbiasedness to allow higher levels of automation in ATM.** The algorithms cannot systematically benefit or penalise the same airline, type of aircraft, route, etc. They must be impartial to guarantee fairness among all the airspace users. Even if such a bias is not explicitly implemented, it must be ensured and proved that the *algorithm has no unintentional bias*. This represents an inherent risk when using systems based on learning, therefore corrective actions during the training process are necessary to avoid them.

- **Automation can play an important role in fairness increase**. In CD&R scenarios, when an ATCO decides to implement a conflict resolution action, it is known that this may not always be the fairest solution. In order to implement the fairer actions, the ATCO might need to communicate to more than one pilot in some occasions, and this certainly takes time that is not available. Machines, though, could look for solutions that are fairer, even when these involve to multiple pilots. For instance, an adequate automation is potentially able to issue instructions to multiple pilots at the same time, minimizing the required time, thus increasing fairness at the same time.

- **Tools must be closely integrated.** This means that *they must consume the same real-time information to ensure data consistency*.

    - **In ATFCM.** In the actual operating environment and tactical/pre-tactical phase, the developed tools need to consider multiple parameters so that the proposed solution is efficient and feasible. If the tools are not properly integrated, the human is not able to select a subset of proposed solutions, implement them using a what-if simulation to obtain a partial solution, then re-consult the XAI to evaluate the resulting situation.

    - **In CD&R.** In these scenarios is important that tools are closely integrated, with a focus on information being exchanged using messaging protocols to allow the XAI to identify conflicts, make decisions and develop solutions in an interoperable manner. This allows the tools to be used in a realistic, real-time mode and to respond to unexpected conflicts which may have resulted due to other ATC or flight deck actions.

6.3.5 Efficiency (Key Performance Areas)

As indicated in [4], The main objective of the ATFCM validation exercises focuses on the ability of users to understand the issues that are identified by the XAI component and the solutions that are proposed or automatically implemented, therefore the analysis of KPA related to the efficiency of the solution are not considered to be of sufficient relevancy to derive principles and recommendations.

Consequently, as outlined in Section 5.2, specific experiments shall be devised, proposed, and conducted in order to extract relevant conclusions and extrapolating them into potential principles that might consider the impact of transparency and explainability on efficiency.

6.3.6 Safety (Key Performance Areas)

- **Level 2 of automation.** The operational experts indicated that applying the proposed solutions was straightforward, but issues relating to the use of 'what-if' tests and selection of only a sub-set of solutions, combined with the inability to then re-run the XAI to try to solve the remaining issues made use of level 2 more difficult. Nevertheless, the additional information available via the co-located VA display and its drill down features facilitated the identification of issues and comprehension of the proposed solutions, with little human effort.

- **Level 3 of automation**. At this level many more solutions were able to be applied in a very short execution time, and the VA support provided plenty of additional information to help humans to understand what actions had been taken. Some additional effort was still required to comprehend the effect of 'induced' Hotspots, created by previous actions, but in general, humans were able to consider many more solutions and *maintain good levels of situational awareness* even though those solutions had been 'automatically' applied. Consequently, solutions were considered to be *sufficiently safe*.

- Safety KPI was at an *acceptable level* (qualitatively) for levels 2 and 3 of automation.

6.3.7 Human Performance (Key Performance Areas)

- It is suggested that all the humans in the study felt *that situational awareness was maintained at all levels of automation* (with the support of the VA tools). They did not feel 'unexpected events' occurring, being clear that the FMP users *were able to perform the requested function at all levels of automation in a reasonable time, with an acceptable level of workload*.

- Human Performance KPI was at an *acceptable level* (qualitatively) for levels 2 and 3 of automation.

7 Principles and Recommendations for the Transparent Application of AI in ATM

This section presents knowledge that contains relevant aspects to consider when automating ATM activities in a transparent and explainable way, using technology based on XAI and VA tools. This way, the knowledge here presented is supposed to lay the foundations of the application of transparent and explainable AI technology for the automation of ATM activities.

The knowledge is presented by means of two different formats: a set of general principles of transparency, and a set of recommendations for the use and application of XAI and VA in the ATM domain. The principles contain knowledge that has been validated by adequately posed experiments, and therefore the information that it contains is supposed to rely on justified foundations. The recommendations, instead, provides a series of more general guidelines, practices and indications. These represent, perhaps, a more relaxed way of enclosing relevant knowledge, but more experimentation and validation might be required in some cases.

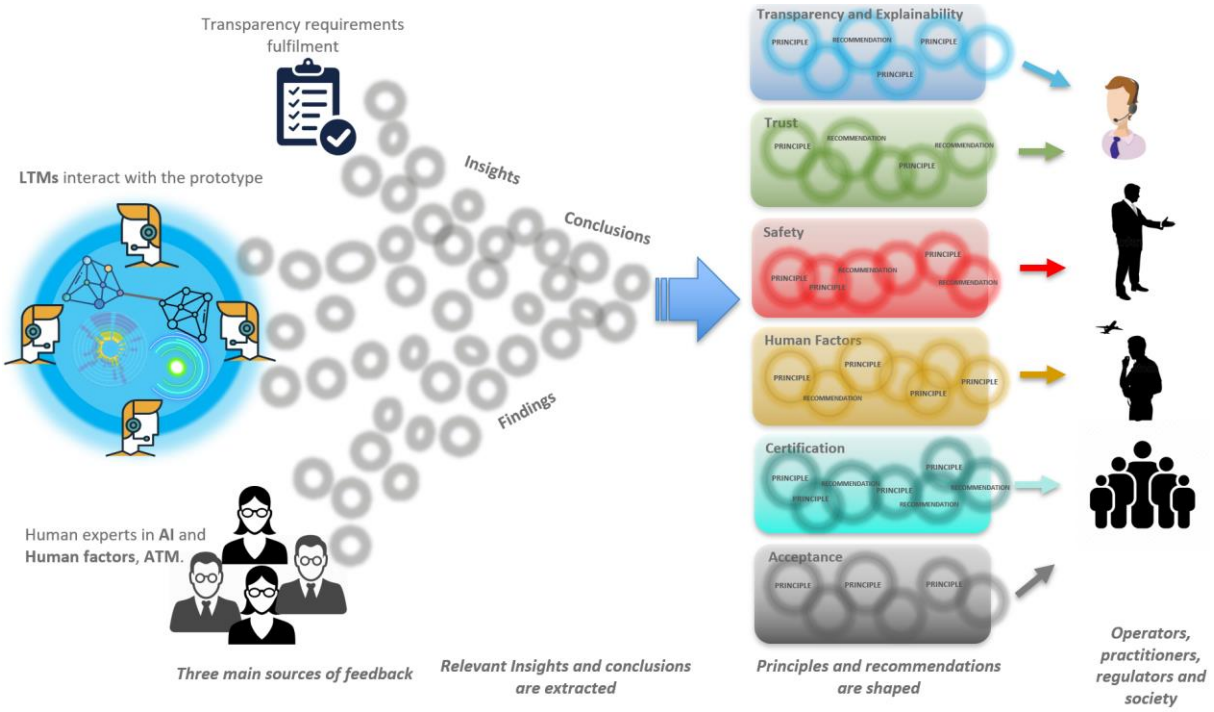

Figure 4. Illustration representing the flow of information in the derivation of principles and recommendations for the transparent application of AI in ATM.

Three main sources of information provide important ideas, conclusions and findings that are converted into actionable insights. These insights are classified, combined, and further elaborated. From the insights, new knowledge is inferred the shape of principles and recommendations, according

to different categories, that can be understood and used by operators, practitioners, regulators and society at large.

As highlighted in Section 3.1.1, it is worth mentioning that it is not a main objective of TAPAS to establish principles that might be regarded as 'universal truths' in any case. The principles and recommendations derived in this research are not intended to be regarded as *immutable truths*, as they are susceptible to refinement and maturation processes.

As shown in Figure 4, the overall idea is to use the three main sources of information used during the TAPAS project development which provide important ideas, conclusions, and findings regarding the use of transparent and explainable AI in different ATM scenarios. From these ideas, key insights are identified extracted, which can be later classified, combined, and further elaborated. Such insights allow new knowledge to be inferred in the shape of principles and recommendations that, according to different categories, can be understood and also are expected to be used by a wide target audience such as operators, practitioners, regulators and society at large.

Section 7.1 describes in a general way the process followed to combine the insights (marked with the Φ symbol) extracted in previous sections and finally synthesize the final principles and recommendations, and shows the format used to enclose knowledge and the main features of their visual appearance. Section 7.2 presents the principles and recommendations, which has been classified into a set of general categories according to their topic.

The knowledge presented through this section is supposed to lay the foundations of the application of *transparent* and *explainable* AI technology for the automation of ATM activities.

7.1 Shaping Principles and Recommendations

7.1.1 From Insights to Principles and Recommendations

An important process through the presented document is the extraction of relevant transparency and explainability insights from the different sources considered in the project, which are the fulfilment of the transparency requirements, the results of the validation activities for the ATFCM prototype, and other expert feedback.

What is an insight? Insights have been extracted through the document by means of a phi symbol (Φ) and they are small or basic pieces of information containing relevant knowledge and findings related to transparency and explainability. These can be seen as individual pieces of knowledge or highlights from which further elaborate.

These individual insights can be classified and categorized according to different topics. Once organized, such insights are related each other (sometimes) and combined into bigger and more elaborated pieces of information, and further developed in some cases. In some cases, the insights highlighted over the document are presented already in an elaborated or reworded way, for the sake of comprehension. This way, having available more complex pieces of information will eventually facilitate the synthesis of new knowledge in form of principles and recommendations. The process of combining and elaborating insights is internally conducted by the document's author and is not

presented here. For the sake of simplicity and legibility, only the final elaborated principles and recommendations are provided in this document.

7.1.2 Format and Content

The format used to present the transparency principles and recommendations is as follows. A main *header* presents the principle or recommendation, using a few words, that attempt to summarize what is the principle or recommendation about in a general way, while at the same time already providing some revealing information. A single glance at the header should give a quick and intuitive understanding of the subject matter addressed by the principle or recommendation.

Subsequently, a *subheading* is responsible for further developing the header by giving more details, as well as providing some intuitive keywords and key insights. The idea is for the reader to grasp the main aspects of the principle or recommendation just by reading the subheading. Finally, the content of the principle or recommendation is developed with the necessary level of detail by means of specialized vocabulary. See Figure 5.

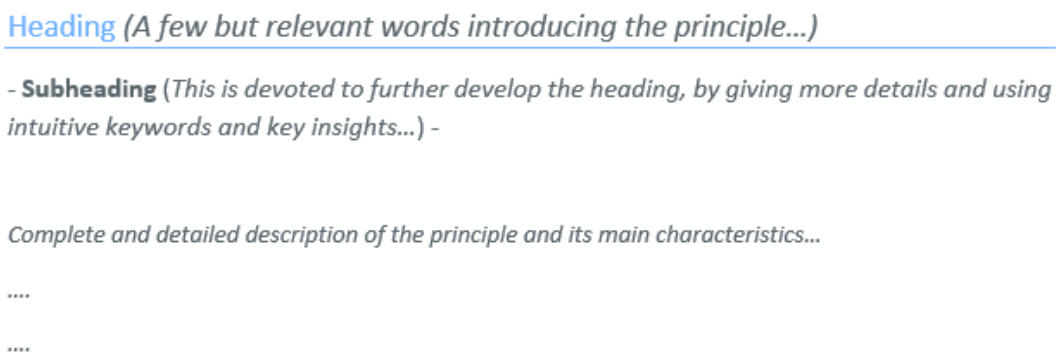

Figure 5. Format and content used in the definition of principles and recommendations.

Principles are presented using a *bluish* colour scheme, whilst recommendations are presented in an *orangish* colour scheme.

7.2 Principles

This section presents the derived principles, organized by category of application. The idea is to provide the reader an idea about the topic they belong to. The principles refer to the ATFCM and CD&R indistinctively, and unless they make explicit distinction inside, they should be considered applicable to both scenarios. This is intentional, and the objective is to provide principles that are wide enough and that have applicability, whenever possible, to the maximum number of ATM operational situations. Hopefully, some of these transparency principles for the use of Explainable AI in automation might even end up being extrapolated to another domain outside the ATM realm, a safety-critical domain which would share some operational characteristics and other similarities with ATM.

7.2.1 Transparency and Explainability

7.2.1.1 General Structure of Explanations

- Carefully structuring the information contained in explanations and presenting it in a clear way to humans is essential for understanding. -

Explanations, if not presented by an XAI in a clear and simple way, can act as a drag on human understanding and decision making. The explanations provided by an XAI must be kept *short* and *concise*, thus avoiding complex and intricate details which might prevent the human user to irretrievably lose the track of the provided solution. In some sense, if the explanations provided by the XAI still requires additional clarification to understand them, and even doing so it is hard to achieve a final comprehension, then the explanations are not optimal.

The information contained in any kind and shape of explanation, either visual, textual or by examples, must be sufficiently straightforward to access and trackable for the human user to make informed decisions based on them. Therefore, the adequate *traceability* of explanations is a key factor to foster transparency. The FMP needs, not only to see the final explanation of the solutions, but also to have a clear traceability of the elements related to each measure or solution. They highly appreciate the possibility of following the thread of certain solution down to the level of the flights to which it is related. This gives a clear transparency to the solutions or explanations provided, making it easier for the user to build trust in the system.

In addition, explanations must contain a *degree of abstraction* of the information and its complexity that favours and enhances human decision making. This degree of abstraction in the explanations must achieve a balance between providing representative details of the internal process for which give insight, presenting the content in a concise way and using a vocabulary and manners that are univocally understandable by the human operator, as well as taking care of the duration or length of the provided explanation. All these factors, if wisely crafted, are expected to rise a positive synergistic effect in informed human decision making in ATM.

Consequently, the explanations must include all the necessary and relevant information, but it is essential to structure their integrated information to eventually and present them a *simple* way so that the human user can access to it in a *clear* and *quick* way.

7.2.1.2 Level of Abstraction in Explanations

- Providing explanations with poor levels of abstraction to high complex solutions impacts real-time understanding. Finding an optimal level of complexity abstraction in explanations is essential. -

Humans sometimes struggle to fully understand real-time explanations given by an XAI. This difficulty can come from diverse sources. A major one comes when an explanation *does not abstract* the inherent complexity of the solution provided by the XAI in a sufficiently accurate and simple way.

It is common that the complexity of a solution offered by an XAI can be too complex for an average human to fully interpret it. This fact gets worse when it is necessary to understand solutions in real time operations. In these cases, it is important to provide the human with explanations that *organize* the information in several *layers of complexity*, *summarize information* and finally present it to the user in *simplified ways*, without losing important information during the process. Finding an optimal level of abstraction for an explanation is essential and requires careful devise and thinking. Similarly, highly complex solutions might significantly diminish the human capability to conduct effective oversight processes of AI/XAI systems, and therefore this needs to be considered when designing solutions.

Perhaps, a timely analogy on how important is abstracting and summarizing complex phenomena into simpler, interpretable, and actionable solutions can be found in Physics. The functioning of the universe is so complicated that human understanding is severely limited to find out how it works. When humans manage to find a model or equation that simplifies all that complexity into simpler means, we can conclude that we have an explanation for how it works, at least for one specific case. Such simpler and summarized means can be expressed as an equation or a model. The equation, indeed, (making an analogy: this would be the explanation provided by the XAI) manages to abstract humans from the complexity of reality and it gives *relevant* and *representative information* about the underlying physics phenomenon (let's imagine that this represents the complexity of our XAI solution). So, shall we somehow pursue offering explanations, which are something similar to accurate equations, that allow us to represent reality and at the same time abstract humans from its complexity? However, the challenge is not trivial.

Consequently, finding an optimal level of abstraction in explanations is essential, and this is supposed to largely impact *real-time understanding* and *decision making*. This might not become a major issue when dealing with ATFCM scenarios, in which the criticality of the action is supposed to be moderate, but this might certainly have a great impact on CD&R activities. Attention must be paid to this.

7.2.1.3 Off-line and On-line Transparency

- Transparency and explainability information can be applied in an on-line or off-line manner. These options depend on the time horizon of the ongoing action. -

Two major ways to tender transparency and explainability are identified, and these mainly depends on the time horizon in which the users operate. Generally, action can take place in three different time horizons, these are *-pre* operation, *operation* and *-post* operation.

Defining a proper *terminology* regarding the optimal timing to provide the users with transparency and explainability information is important. This way, transparency and explainability can be provided *off-line* or *on-line*:

- **Off-line transparency/explainability**. This involves providing any means of information -pre or -post the operational action. This includes:
  - -Pre. This mainly refers to the training process of operators in which they are instructed and taught on how to operate with specific XAI tools. This stage also involves system development and testing, verification, and certification.
  - -Post. This refers to the fact of providing transparency once a specific event has occurred and so the end-user requests some transparency in order to get insight on the outcome of a process. For example, in case of formal post-ops investigations after an operational incident.
- **On-line transparency/explainability**. This refers to the time horizon in which users actually operate using the offered tools to solve real problems. It involves providing any means of transparency during the operational action itself, and it can also be seen as a means for real-time transparency supply.

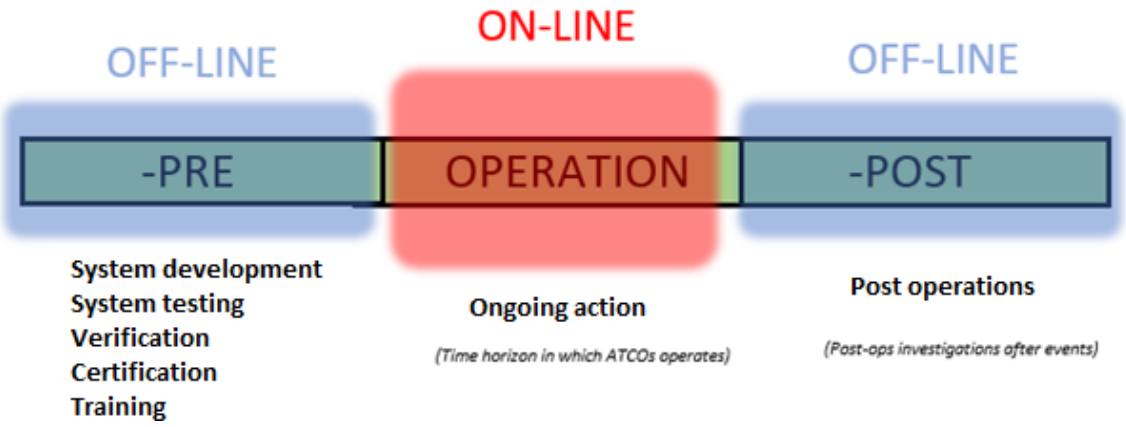

Figure 6. Two major ways to tender transparency and explainability depending on the ongoing action.

7.2.1.4 Ongoing Actions Define Transparency Needs

- ***The time horizon in which the action operates impacts the amounts of required explainability/transparency.*** -

The need for user to access to different amounts of information containing transparency/explainability aspects *varies depending on the time horizon* in which they operate. Interestingly, it has been proved that users require different amounts of explanations on specific solutions given by the XAI, depending on whether they are immersed in the operation or training (-pre) stages.

It turns out that during *operations*, users actually *do not need* to receive and see neither large amounts of *on-line* explanations nor lots of information in relation to the solutions proposed by the XAI. This might involve spending longer amount of time trying to understand all the information received, and they do not have such time, generally. It is stated that, in fact, users do not need to see the intermediate solutions the XAI algorithm considers and analyses before arriving to the final solution. Therefore, the amount of *on-line transparency* given during operational time must be carefully rationed and somehow reduced, as it is suggested that uncontrolled doses of transparency might have negative impact on human factors.

For example, in CD&R scenarios at partial automation, a very interesting fact suggested by ATCOs is that the level of required explanations should depend on the time horizon on which the proposed conflict resolution operates. This means that, if the proposed solutions only solve conflicts in a *short-term horizon*, without looking at the effects at medium or long term, no explanations would actually be needed. ATCOs tend to know almost all the possible solutions to a specific situation, they seem to control all the possible outcomes by their experience. ATCOs are fully capable of understanding conflicts rapidly due to their own experience, and they are able to evaluate and understand the potential impact of the proposed solutions with relative ease. Furthermore, ATCOs have no time for in-depth 'explanatory' information since little time is available in real-time CD&R operations. Actions need to be performed and verified *rapidly* and in *priority* order if more than one solution is available, before moving on to the next task. In addition, in cases where multiple actions are proposed to solve the same conflict, users need *clear indications* to distinguish these from other cases where a choice of independent solutions are provided. The most interesting thing comes now because it turns out that if the XAI system would propose solutions that look at a longer temporal horizon, thus involving *long-term situations*, then in this case more detailed explanations would be needed to yield light on those decisions. In other words, whenever the action transcends the ATCOs' human temporal, vision, and cognitive capabilities, then they need to fully understand the consequences of the actions that they are going to make. Here, explanations are needed.

During *training* stages, the issue is certainly different. Here, it seems that users are generally more open to interact and experiment with diverse *off-line* transparency and explainability. This is the phase where users do really need a greater number of off-line explanations on the algorithm and solutions provided by the XAI. This fact seems intuitive, because they need to fully understand the rationale and justification given by the XAI to each of the offered solutions. Once the proposed approach is sufficiently understood, users no longer need this information, as they already are instructed. They are

now highly knowledgeable on the tools to use, and they might have increased their confidence about the solutions being provided. They are ready to use the tools on the arena.

7.2.1.5 Necessity of Transparency

- There are different means for offering transparency, and these are influenced by the operational context, the user's prior training or the current human cognitive workload. -

Transparency can be offered and therefore accessed by users in multiple ways depending on the operational action and specific needs. The specification of different ways or levels to supply certain amounts of transparency/explainability might depend on aspects such as the operator's prior knowledge/experience about a process, the specific need (or lack of it) to obtain relevant insights on specific XAI's solutions/decisions, or even the operator's current cognitive workload.

The main needs for supplying transparency across the different operational scenarios can be covered by three different levels. These are '*transparency by default*', '*transparency on demand*' or '*self-transparent*'.

In transparency *by default*, the XAI system acknowledges the unusual or disruptive condition of a solution reached and then it provides, by default, transparency on it for the controller gain insights. This way of offering transparency is supposed to provide clear advantages specially in training phases, where the controller interacts with the tools in order to fully become familiar with them. Therefore, at these stages, having some degrees of transparency by default can be beneficial for the controller to understand as much as possible from the decisions taken by the XAI. This level might include also the option to '*disable*' or '*enable*' (on demand) the transparency, if requested by the human operator.

An interesting way of offering transparency is *on demand*. Here, the XAI may provide additional information to explain the solution reached, if requested by the human operator. For example, human operators should be able to ask for rationale at any time, or to request explanations on what happened when certain solutions go out of the acceptable frame. However, the decision to ask for further justification about the proposals or decisions of the system should remain with human operators, depending on the ongoing operational context. Furthermore, this level also encompasses the possibility of '*switching off*' or '*disabling*' the provided transparency, if the operator (for some reason) is not trusting it. For instance, it is expected that if the XAI offers excessive amounts of transparency that is provided all the time, additional workload will add up to the operator's cognitive task without a possible and clear benefit for the operator. These situations must be completely avoided, and this level of supplying transparency might facilitate keeping the human within reasonable margins of transparency for optimal processing.

Also, should the solution given by the system be explicative enough (*self-explanatory* or *self-transparent*) or it is known to be reliable by past experience, then no additional information is needed nor provided by the XAI system. This case can occur when the controller faces a typical or more routine operational situation in which the solutions given by the XAI are highly trustable and known by past experience when interacting with the system. In this case, no specific needs for providing additional

transparency on the solutions are needed since controllers already perfectly understand and agree with the reasons for specific decisions and solutions.

These three ways of providing transparency to human controllers could cover most of the transparency needs that may arise depending on some of the factors mentioned above.

7.2.1.6 Interfaces for Transparency

- Selecting efficient interfaces and adequately integrating them is paramount to the way transparency and explainability is conveyed and perceived. -

Using appropriate interfaces to convey key information in terms of transparency and explainability is an essential aspect. It stands to reason that the experience and interaction of the operator with the information provided by an XAI system may be perceived in radically different ways if the selected interfaces *are not adequate*.

For example, in ATFCM scenarios, the transparency/explainability provided by the XAI prototype, in combination with adequate VA displays and an FMP Client interface stand up as *relevant* and *efficient providers* of solutions involving transparency and explainability aspects. This original way of deploying and presenting solutions allows the controllers an optimal interaction, for example, by interrogating different scenarios, investigating the proposed solutions in an intuitive manner, as well as exploring possible reasons about why some solutions have been automatically implemented, depending on the ongoing automation level.

A well-designed FMP client is able to provide interactive demand and occupancy information for a particular airspace, in an efficient manner. To make a set of rapid access menus available in order to access and display key information related to sectors is also something valuable to the controller. At a partial automation level, an FMP client is also able to consume data related to traffic demand, Hotspots and proposed ATFCM solutions produced by the XAI component. In this way, the operators can easily view the information provided by the XAI in exactly the same format as the operational data held in the platforms used at lower levels of automation, and to utilize all of the interactive charting and flight list features supported by the XAI prototype. Additionally, as the paradigm used by the XAI to represent Hotspots and create Regulations is slightly different to that used by NM in today's system, providing an FMP client with adequate algorithms that allows combining the XAI Hotspots and Regulations and present them using the NM definitions can be useful. An FMP client should provide the controller functionalities as:

- Input new ATFCM actions (e.g., ATFCM regulations or re-routing measures) depending on the ongoing level of automation.
- Accurate information about hotspots identified by the XAI system
- Sector / traffic volume load and occupancy charts for any of the ACC in the region
- Interactive consultation of traffic lists for the flights which make up any of the charted periods
- Overloads identification on the load/occupancy charts in regard to the declared capacity thresholds

- Display of the solutions proposed by the XAI system to traffic for each Hotspot
- Simulation on the impact of hypothetical DCB measures by 'what-if' analyses
- Application (manually at level 2, and automatic and level 3) of real DCB measures in the operational environment, etc.
- And tools must be closely integrated. This means that they must consume the same real-time information to ensure data consistency. In the actual operating environment and tactical/pre-tactical phase, the developed tools need to consider multiple parameters (weather, flight updates, changes in the opened scheme, etc.) so that the proposed solution is efficient and feasible. If the tools are not properly integrated, the human is not able to select a subset of proposed solutions, implement them using a what-if simulation to obtain a partial solution, then re-consult the XAI to evaluate the resulting situation.

Also, components based on VA technology are required to optimally present the explainability and transparency needs provided by the XAI system. The way information is transmitted through VA displays must be therefore *carefully designed* to support explainability and transparency through a series of interactive views that the controller could work with to help understand how Hotspots are identified and why certain solutions are proposed. Also, it was shown that it is useful to make use of interactive VA displays showing relevant information related to analysis of the DCB scenarios being analyzed by the AI components, such as:

- Visual summaries for the variants of solutions
- Display of the evolution of the solutions over iteration steps of the simulation process
- Provision of details for each hotspot, sector, and time interval, including aggregated information about the flight delays
- Support for on demand such as features that justified delay decisions for selected flights, sectors, and time periods, etc.

In CD&R operations, the integration of XAI/VA tools with the platforms used is key. This must consider:

- *Interoperable consumption/exchange of data* between the platform and the automation/visualisation components.
- *Update/refresh rates* must be carefully designed to avoid confusion or time wasting on the part of the operators when actions had already been taken. This could also be improved using either status information (regarding the actions that are in progress or required) and/or a series of time-based steps that need to be followed by the operator.
- *Integration of visual components in the same display* if safety critical views are used by the operator. For example, the integration of the VA information with the CWP radar screen (instead of on an adjacent co-located display) helps the operator to *maintain their situational awareness* and to avoiding unnecessary deviations from the radar display. In other words, it has been seen that using this set up, ATCOs work *more efficiently* and are able to use more transparency-related information, which helps them to keep focus on the evolving traffic conditions in the sector, especially when traffic loads are high and complex.
- In these scenarios is important that tools are closely integrated, with a focus on information being exchanged using messaging protocols to allow the XAI to identify conflicts, make decisions and develop solutions in an interoperable manner. This allows the tools to be used in a realistic, real-time mode and to respond to unexpected conflicts which may have resulted due to other ATC or flight deck actions.

Adequately selecting, combining, and integrating some of these interfaces ensures a large number of benefits when it comes to deploying and effectively conveying transparency-based solutions.

7.2.1.7 Features

- 'VA Features' stick out as promising mechanisms to visually give explanation to how flights are gradually affected by DCB measures. -

In ATFCM scenarios, FMPs do need to univocally understand the reasons for the DCB measures proposed by the XAI system. This statement applies both at partial and full automation levels. In the first case the FMP takes an active role because he/she will directly implement the proposed DCB measure, whilst in the second case there is a more passive role because he/she supervises the measure automatically implemented by the system. In both cases, the FMP needs proper rationale to such proposals.

VA Features are elements that support classifying information that represents different characteristics of the situation of one flight over time, such as number of hotspots, number of delays, period of hotspot, etc. All this data helps the user to understand how this flight has been affected gradually by the measures until the final solution is reached. This way of enclosing relevant information seems to be useful for FMPs.

Once trained, FMPs are able and show competencies to validate DCB issues identified by the automation and to use the VA features which help develop an understanding the measures being proposed as well as the reasoning behind those solutions. In addition, VA 'drill down' features can also facilitate the investigation of the solutions in more detail and to develop a better understanding why those solutions are proposed by the XAI system.

Even though VA features require some effort on the part of the operator, this mechanism can be seen as a support method to understand transparency aspects, explanations and drill-down understanding of the DST processes. This is especially the case for those solutions that are automatically published and implemented without operator action. Additionally, it is suggested that using the complementary features in the FMP client and VA display, controllers are even able to *maintain situational awareness* at good levels.

Consequently, VA Features, if *improved and polished in terms of presenting the information in understandable ways*, show promising potential to be applied during operations in ATFCM operations.

7.2.1.8 Aggregating Information

- Endowing explanations with aggregated information stand out as a powerful mechanism to boost the interpretation of the solutions offered by the XAI. -

In ATFCM scenarios, FMPs agree that for levels 2 and 3 of automation providing the right transparency information is *useful*, *understandable* and that information is *easy* to access most of the times. However, they also identify a set of additional needs for transparency and explainability at such automation levels. The underlying reason for such proposal is the need for classifying the information in a more aggregated manner, as well as for being able to completely justify the decisions being made.

Specifically, it is suggested that the human interpretation of the information being presented can be largely enhanced by endowing explanations with additional data, *aggregated in diverse and multiple ways*. FMPs are highly interested and inquisitive to see information presented in an aggregated way, such as statistics on the potential impact on flights of implementing the proposed solutions, including a breakdown by region, sector, airline operator, minutes of delay saved, etc.

Providing the FMPs with mechanisms for dealing with and visualizing different ways of aggregated information is important and must be considered. They do see this functionality as a relevant way to describe and intuitively highlight the effect and impact of the proposed measures on the overall system when assessing the effectiveness of the XAI-based solutions. Consequently, aggregating information and including them into the offered explanations stands out as a powerful and promising way for FMPs to see their capabilities to interpret solutions enhanced at some extent, which might also represent a positive impact on final *decision making*.

However, for the CD&R use case, which is safety critical, the operator normally prefers to see the most relevant information only in a *straightforward manner* by default, as they have a limited amount of time to assess and solve the conflicts.

7.2.1.9 Information Assimilation and Understanding

- Transparency-based information must be offered in its simplest and clearest possible way, and including, by default, only the most relevant information. -

Generally, in ATFCM scenarios it has been proved that the effort for FMPs to scan VA displays is *low*. A possible reason for this might be that humans greatly rely on visual stimulus to quickly shape a mental representation of a particular phenomenon. Intuition suggests that multiple displays, if intelligently combined and arranged, are supposed to catalyze human understanding by offering different points of view and edges of a particular solution, at the same time.

The use of dedicated displays to present information through multiple customized views using VA tools can *support* and *stimulate* the reception of transparency and explicability at different levels of automation. Having multiple views on a display is proved to facilitate the consultation and interactive acquisition of highly detailed information and explanations, at any time, as well as helping human

operators to understand the issues that had been identified, as well as the reasons behind the proposed solutions. In particular, it is proven that FMPs are able to work efficiently with multiples views containing:

- *General views* showing information about the overall state of the sectors in terms of traffic demand, highlighting the detected hotspots by the XAI.
- *Specific views* which allow analyzing certain sectors, through which the user can see the flights crossing that sector in a 2D representation and analyze the declared hotspot and flights involved during that time.
- *Another series of views* exclusively devoted to present to the FMP the explanations given by the XAI to the different solutions.

Nevertheless, effort is still needed to improve those cases where the access to information is still *obscure* and so, it requires further *refinement and polishing of interfaces*.

In CD&R scenarios, however, transparency-based information is perceived as easily *accessible*, *clearly* presented, and it can be properly understood by ATCOs in a *timely manner*. Also, the level of detail of this information is useful, and it allow ATCOs to quickly comprehend situations, understand actions that are proposed or implemented by an XAI, to review options being proposed and select adequate solutions. For example, in relation to the proposal of conflict resolution actions by the XAI, having high-quality transparency information allows ATCOs to have a clear vision of such actions. This way, they are able to easily identify all the conflicts and they are not disturbed or overloaded by an excess of information.

Whenever solutions seem feasible, ATCOs barely need to make efforts to get a good understanding about what the automation is trying to do. It is easy for them to have a good global overview of the situation, to be ahead of the traffic and to be fully capable of planning and organise the work they need to do. In addition, ATCOs tend to request additional transparency-based information to the XAI and, if this is easily obtainable, the benefits are huge. Not only that, if such extra information about the conflicts and proposed solutions is representative enough to support understanding of the solutions, the actions proposed and foreseen consequences, this is highly beneficial for the maintaining of a *good situational awareness.*

Regarding the *CD&R platform* and its' use, ATCOs seem to be able to easily access the information provided in the VA support to understand solutions being proposed and, using features in the CWP, they are able to measure factors as impact and applicability in a painless manner. This type of information can also complement ATCOs' expertise, as well as the existing SACTA tool suite. Providing all of these elements certainly facilitates ATCOs to maintain very high levels *of situational awareness* and their level of understanding about what is operationally going on is high.

Experiments suggest that, if the platform and tools used are easy to understand and quickly to learn, ATCOs do not seem to need the help of technical support personnel to use them, at least at *partial automation*. At *full automation*, although ATCOs feel in general satisfied with the information provided and how the tools can be used, it is suggested that they tend to *rationalise* the solutions being applied by the automation tools, which may sometimes be confusing. This may be even worse considering the lack of some additional actions. Nonetheless, ATCOs manifest that if these issues are rectified, then little or no additional training is necessary.

Therefore, even considering the fact that ATCOs would probably conduct actions in a different way to how automation tools actually do, no extensive transparency-based information seems to be needed if this is presented in the *simplest* and *clearest* possible way, and including, *by default*, only the most relevant information.

7.2.2 Human Factors - Trust

7.2.2.1 Trust Volatility

- The trust in the XAI system that has been slowly crafted, can be quickly destroyed if something fails. -

It has been shown that the trust and confidence of FMPs in XAI tools is something that should be built *primarily in early stages*, that is, in the *training* phase. In these stages the human needs to see and understand that the solutions proposed and implemented by an XAI system are efficient in terms of the potential impact they will have on traffic, for example in the number of delays, hotspots resolved, flights affected, etc. This process of building trust is assumed to be slow, it requires time, maturation and the human capacity to understand the repercussions that the solutions (*innovative* or *disruptive*, perhaps, in some cases) will have, and most importantly, eventually accept them and learn from them as well. This paradigm of a human who is learning from under-the-hood mechanisms about XAI solutions, whilst assimilating and absorbing such information, is expected to provide them with new capabilities to operate more efficiently and confidently during real operations. The process seems to be slow, but it certainly promises countless benefits regarding the building of trust of a human with respect to the XAI system.

Bad news coming here: such slowly acquired trust in the XAI's solutions and way of functioning can be lost very rapidly. Trust and confidence can be *volatile*. As mentioned earlier, developing trust and confidence in a system does take a long time but also relies on that system providing, and continuing to provide, reliable solutions that the human accepts as being a valid response to a problem. In the event where something subsequently fails badly without the XAI system timely and adequately supplying convincing and explanations or complementary information on the exact failure reason, that confidence in the system will be lost very rapidly. To make it worst, *rebuilding* it can be hard.

A good example that represents how unexpected solutions can destabilize trust in the XAI system is found in disruptive solutions. These impactful solutions are expected to alter somehow the way a human understands a problem, and even change his/her way of tackling similar problems in the future. However, these solutions must allow certain degrees of reproducibility. This means that when dealing with disruptive solutions proposed by an XAI system, it is important that those solutions do not vary too much when tackling the same type of future problem. *Disruptive solutions must allow reproducibility*, and this certainly promotes confidence in the system. In this way the human will be able to create a mental pattern on how the system works without being afraid of major changes in future similar situations, thus improving understanding and building trust on the system more quickly and robustly.

7.2.2.2 On Trust Building through Explanations

- ***Trust is better built when relying on effective solutions and understanding their expected impact on the system, rather than by complex explanations.*** -

Users prefer trusting an XAI system, rather than obtaining intricate or in-depth explanations. Sometimes they do not even need explanations at all during operational time. It's time consuming and might end up harming trust building.

It is through the *constant*, *regular* and *extensive* use of the XAI system, especially during the training phase, that the human is able to gradually develop trust in the system and become familiar with the offered tools and technology. Different levels of trust are generally gained depending on how the XAI performed and the effectiveness of the solutions that were provided. In summary, the need for detailed level of explanations in ATM scenarios must be left for training and tool familiarization stages.

What users do really care about is about the *effectiveness* (or *efficiency*) and potential *impact* of the solutions offered by an XAI system. Users seem to build trust on the system by means of these two factors. For example, the users do need that the solutions are highly effective for them to ensure predictable outcomes for their actions, based on answers coming from XAI systems. It could be said that *trustworthiness*, if achieved, might be considered the best of the explanations. This is important, and it is expected to contribute to trust building, and such process for understanding how effective the solutions are must be conducted and assessed in training stages by means of robust accuracy metrics.

However, from experimentation it turns out that one of the most powerful means to generate confidence in the XAI system is to fully understand the *impact* that a potential solution *will cause on the overall system*, beforehand. This is extremely revealing, as well as a significant booster that stimulates human understanding and building of trust in a tool. In an ATFCM scenario, having the a priori knowledge of what results will cause the hypothetical application of a DCB measure, see the impact with a visual representation, whilst obtaining pertinent explanations of the expected impact sounds reveals itself as a very relevant mechanism for decision-making. This can be seen a decision-making enhancer, which at the same time contributes to trust building. And, indeed, this seems to be more valuable than the current explanations provided by the support tools to FMPs.

In CD&R tasks, it is also outlined that a situation in which ATCOs would acknowledge the provision of explanations is in case that the XAI gives a list of actions to *'not execute'*. For example, depending on the situation the XAI system could give the ATCO a proposal of not doing a flight level change, and complement this proposal with an explanation detailing the reasons for that, and expected consequences. ATCOs manifest that explanations in this context would be truly useful and may eventually promote confidence in the tool.

All of the above reveal, in general, the importance of the *technical accuracy* of the presented solutions as a main carrier of trust in the tools, without the need for detailed explanations. This is relevant to the point of asserting that some users highlight that seeing solutions which work, even if they are different to those that they may have chosen themselves, is sufficient to develop trust in the system and to accept how it is performing without the need for more explanations.

7.2.2.3 Trust Degradation by Unrealistic or Inaccurate Automations

- ***The solutions provided by XAI automation tools must be realistic, accurate and feasible or they are at risk of acting as a trust builder blocker.*** -

Accurate, realistic and feasible automation solutions *promote trust building*. Specially in CD&R scenarios, different aspects such as a combination of unrealistic (from a perspective of human way of thinking) solutions, or solutions that may lead to more complex issues further downstream, along with the lack of additional actions that are also considered to be an integral and necessary part of the conflict resolution process may contribute to a significant reduction in trust and confidence in the automation tools.

For example, if an XAI-based automation only focuses on actions that resolves the conflicts that are detected, and then it proposes actions that are limited to the establishment of suitable separation in accordance with the minima required for the region in which the aircraft are operating, limitations in confidence toward the tool may arise. It has been proved that it would be highly beneficial for trust building including actions such as clearances to help aircraft recover their original plans, and other aspects such as directional flight level strategies, aircraft types/performance characteristics, proximity to departure/arrival airport etc. Not including the aforementioned aspects, apart from being inconsistent with actual practices, may lead to a significant reduction of confidence in the tool.

For instance, in real scenarios it was seen that an automation that provides clearances that may result in additional and sometimes more critical downstream conflicts, or solutions that requested traffic to climb to solve a conflict when a descent would have been better, or scenarios in which the XAI is not able to detect all the conflicts, and some of them are not solved in the most efficient manner or using ‘open loop’ manoeuvres (e.g. a heading change). Furthermore, ATCOs tend to question the system's solutions and ask for reasons when conflicts are partially solved, or unresolved. In other words, if solutions lead to further issues which the ATCOs can avoid based on their own experience, then not optimal automation can act as a trust blocker.

In addition, from experimentation it can be outlined that a factor that would considerably increase confidence in decision-making is to offer several and alternative solutions to the human controller in order for him/her to compare the available solutions based on their potential impact. This simple action would likely enhance decision-making allowing the human to choose the most suitable one. Giving humans the possibility to select a solution from several options, each one presenting their likelihood to solve a problem, and implement it by their own is also expected to gradually increase trust in such solutions.

Proposing accurate and variate solutions is so important, that even providing transparency and explainability for a solution (if this does not solve a problem in first place) could not be by itself enough to promote trust XAI-based solutions. *The accuracy of a solution must come before any means of transparency*.

Finally, ATCOs manifest that, the culmination of trust building would be a *hypothetical monitoring situation* in which they could stay for hours by simply looking at the display and doing nothing. Just by observing how the automation is able to accurately solve the conflicts without enforcing them to take over. That would be the definitive trust builder.

7.2.2.4 'What-if' Mechanisms

- On how forecasting the impact of potential decisions fosters transparency and enables trust building across different automation levels in ATFCM scenarios. -

What-if mechanisms are recognized by FMPs as a powerful and extremely useful means to foresee the impact of DCB measures proposed by the XAI system, whilst conveying key transparency aspects on the elaboration of such measures by the XAI. This way, FMPs can perform what-if analysis of potential actions to see how they would affect the current situation, *without actually changing the real traffic situation*. This is mighty.

Once the FMPs understand the expected impact and agree with the outcome of a what-if scenario, the same actions can be published in the real scenario as the proposed solution to a given DCB problem. The fact of having the confidence of controlling the consequences will act as an enabler of trust over time. This is probably because every time that humans see their expectations translated into materialized actions, they are *positively reinforced* and they feel they have the situation *under control*.

This way of working takes especial relevance at level 2 of automation, *partial automation*, in which the operator selects one or more of the solutions proposed by the XAI and initiate the action. In this sense, the what-if functionality is recognized as very useful and complements decision-making. However, careful attention must be paid on the way to provide information regarding the impact of solutions. It is suggested that presenting the information in an aggregated way would be beneficial because it would allow the FMP to more effectively and intuitively assess the appropriateness of such solutions. However, it is worth highlighting that the use of 'what-if' functions may be limited in time constrained safety-critical applications (e.g., CD&R).

At automation level 3, *full automation*, where the operator is in charge of reviewing the solutions that had been automatically applied by the platform, this paradigm is different. At this level there should already be a certain level of confidence previously established in the tool. Therefore, progressive trust building would continue to take place, albeit at a different pace. Now, the operator can access transparency and explainability information about the solutions proposed and selected by the XAI. He can understand the rationale behind these proposals. In addition, the operator is able to monitor the automatic implementation of these solutions by the system. Consequently, having the ability to monitor the entire process, from the proposal of solutions to the automatic implementation of measures, and the verification of the final effect on the system (if everything works as expected), will lead to a very positive effect of *reaffirmation* and *confidence* in the tool.

7.2.3 Human Factors - Acceptance

7.2.3.1 Unbiasedness and Fairness

- ***A wider acceptance of XAI-based automations relies on fairer and more unbiased solutions***.
-

There exist factors that might act as barriers if social and operational acceptance when automating ATM scenarios by XAI are to be achieved. These barriers must be broken out, and providing humans with revealing forms of transparency about how such automation occurs and why some solutions are selected, is expected to be a key factor to achieve so. Regardless of how long it may take. Specifically, there are two simple but important concepts that may largely contribute to a greater human acceptance of XAI-based ATM automation.

For instance, the algorithms used in the automation of ATM scenarios must prove high levels of *unbiasedness*. As an example, it was highlighted that such algorithms cannot systematically benefit or unfairly penalize the same airlines, type of aircraft, routes, airports, etc., without providing strong reasons. Algorithms must be impartial in order to guarantee fairness among all the airspace users. Even if such bias is not explicitly assigned or implemented, mechanisms must be included to ensure that the algorithm has *no unintentional bias*. This also represents an inherent risk when using systems based on learning, therefore specific corrective actions during the training process might be necessary to avoid them. These concepts are important that there seems to be a general agreement on that greater acceptance and confidence in the solutions would be achieved if the solutions implemented by the automation were reasonable and well balanced across all airspace users. Unbiasedness seems a good mechanism to ensure this. Otherwise, this may pose a barrier for trust and acceptance, initially for the FMP operators and very quickly thereafter for airspace users and other stakeholders.

Likewise, in terms of *fairness* increase, an adequate automation may play an essential role. For example, in CD&R scenarios, when an ATCO decides to implement a conflict resolution action, it is known that this may not always be the fairest solution. In order to implement the fairer actions, the ATCO might need to communicate to more than one pilot in some occasions, and this certainly takes time that *is not available*. Efficient automations, though, could look for solutions that are fairer, even when these involve to multiple pilots. For instance, an adequate automation is potentially able to issue instructions to multiple pilots at the same time, minimizing the required time, thus increasing fairness at the same time. Giving some transparency for the user to fully grasp why the proposed solutions are fairer would definitely represent a boost for acceptance.

In addition, it is generally outlined by human operators that if the transparency and explainability aspects provided by an automation are not proved to be sufficient in terms of *usefulness*, *reliability*, *accuracy* and *comprehensibility*, this might represent a serious blocker for acceptance as well. In this case, further research and improvement in the XAI-based mechanisms employed to convey transparency and explainability may be needed.

Although more issues are expected to be identified regarding the acceptance of XAI-based solutions, these are some of the dilemmas to be faced and resolved in the path towards social and operational acceptance of XAI-based solutions in automation.

7.2.3.2 Closing the Gap Between Humans and XAI systems

- What if the paradigm was flipped to bring humans closer to XAI systems and not the other way around? -

XAI systems can function in a very different way to the way a human naturally does. Historically there has been a clear, and somewhat logical, tendency and efforts to bring the machine closer to the human, rather than the other way around. Our natural tendency makes us want to devise and develop models that are based on or reminiscent of human behavior, or even reproduce human biological functions (e.g., models based on biological neural networks). However, it should be noted that this may not always be the most efficient way.

Sometimes, systems that do not work in a way that resembles, or is reminiscent of, the way a human would do it are precisely those that show the best final performance. Focus must be brought to this fact.

As an example, the ATFCM XAI prototype turns out to work *very differently* to how human operators work. The methodology applied by the XAI system attempts to solve all of the problems encountered at once (a.k.a., all in one go), which radically contrasts to how FMPs work today by solving problems one by one in a more traditional paradigm. They are completely different approaches not easily subject to reconciliation. For example, some of the immediate issues that this new paradigm introduces, at partial and full automation levels, is that human operators sometimes struggle when trying to focus on the problem as a whole, rather than single problems. This, as one might have guessed, is mainly due to the way that they work nowadays, solving one problem at a time, therefore on some occasions they are surprised by induced events that they actually do not expect. This can be seen in, for example, when trying to detect problems in specific sectors. Sometimes they miss such problems, even if shown in the display, and the main reason turns out to be that they are not fully focused on at that time. Also, they can make mistakes when trying to identify solutions proposed to certain sectors and flights to solve a hotspot in another different sector. It is suggested though that this new way of solving 'all problems in one go' could be accepted as a new working paradigm if it is demonstrated that the algorithm does not contain biases.

Something similar, and very interesting, occurs with the CD&R XAI prototype. Some ATCOs, during experiments, suggest that if the solutions provided by the automation were more similar to those actions that *they would have chosen* for the resolution of a specific conflict, they would *accept them more easily*.

All these aspects might result counter-intuitive at first and perhaps a very instinctive (and natural) human reaction is to implement changes to the recently created XAI system in order to bring it closer to the human's way of functioning. But also, perhaps, an alternative and more optimal way to go may be to provide *adequate training* for the humans to fully understand that new paradigm, which may be recently arising. Of course, to achieve that, the solutions although exotic for some users, should demonstrate to be extremely accurate, but making efforts to bring humans closers to XAI systems

could be one way to go. Otherwise, we could be at risk of seriously constraining or limiting the new paths for technology development, due to a somewhat selfish or simply a reason that make humans feel more comfortable.

So, why don't we allow new paths to technology development to come even if this involves *extensive human training* to properly *understand*, *use* XAI systems, *control* and naturally *accept* their outcomes? Otherwise, the potential for achieving XAI's performances beyond human capabilities might be seriously stuck.

7.2.4 Safety

7.2.4.1 Automation Levels and Situational Awareness

- How does switching between automation levels affects to human situational awareness and understanding? -

This might be one of the most concerning issues when it comes to introducing automation roles in the ATM domain. Initial experiments in ATFCM scenarios highlight some remarkable insights. Generally, users are capable of maintaining a *controlled level of situational awareness* during operations, as well as being able to describe what the automation consists of, or why certain solutions have been proposed by the XAI tool. This fact gets reinforced as the users gain more knowledge on the tools, over time. And these results seem to be valid for levels 2 and 3 of automation.

However, at higher automation levels there are also limitations. It turns out that some users experience difficulties when trying to focus on the problem as a *whole*, rather than single problems. Due to this, users are sometimes surprised by unexpected events, which they do not expect according to their operational experience. The reason for this is due to the fact that they work nowadays solving one problem at a time, therefore on some occasions they can be surprised by induced events that they do not actually expect. Besides this, users may fail to detect some problems that are displayed on the screen, in sectors that they were not focused on at that time, as well as some problems and solutions proposed to certain sectors and flights in order to solve a hotspot in a different sector. To make it worst, it has been seen that they are even at *risk of forgetting things*. Once again, the reason is that they are used to focusing on problems individually, one after another, and the functioning of the XAI tool represents a paradigm change that requires time, and training. Still, somehow it was seen that FMPs are generally able to *maintain a controlled level of situational awareness* during operations. At the same time, users prefer to remain humble and do not pronounce on their ability to keep things under control, specially at level 3 of automation.

In CD&R scenarios, extracting clarifying insights is a bit fuzzier. At level 2, users do not seem to agree in relation to how well they understand the actions proposed by the XAI tool to solve the occurring conflicts. There are clear *discrepancies* in this regard and first investigations suggest that this may be a consequence of the actual clearances being proposed by the XAI, which seem to differ significantly to those naturally proposed by humans, as well as the *lack of confidence* in the applicability of those solutions. To make it harder, at level 3, users express even more difficulties to adequately understand why the tool provides certain solutions and to describe what the automation is actually doing. This might have a simple explanation. For example, it could be that such difficulties are somehow influenced by the choices that were made by the automation. Some these solutions, if not considered to be realistic or efficient enough by the users, might build a gap between the human way of reasoning when solving conflicts and the automation paradigm itself. In addition, the inherent human monitoring role for level 3 might even convert the user into a mere spectator with very constrained capabilities for intervention and control of such an ill-fated situation. All of these factors certainly have the potential to negatively influence how users perceive and understand the solutions provided by XAI-based automation.

Despite all of these situations, in terms situational awareness itself, users manage to agree on that even if the automated resolutions were not ones that they would have chosen themselves, they still

seem to be able to understand the expected impact and consequences of those actions, and therefore they can still maintain a *good situational awareness.* However, further investigation and stronger experimentation would be revealing in this sense.

7.2.4.2 Intervention and Control Recovery in Automation

- ***On the feasibility to deploy full automation levels in real world scenarios.*** -

The full automation level 3 [14] implies, almost by definition, a human supervision or *monitoring role*. In this level, automation can initiate actions for some tasks but the human operator is still in the loop, and his/her presence is needed. This is, therefore, the first automation level in which automation is allowed to initiate actions by itself.

If one thinks about the natural consequences of this, in such a critical scenario as CD&R, safety comes first into the game. Perhaps, the most immediate (and reasonable perhaps) thing to consider would be to provide the automation with mechanisms for human intervention in case something goes wrong during the actions conducted by the automation tools. Some initial experimentations attempting to test these concepts highlight, though, some current limitations. Apparently, there exists concern of experimented users regarding the deployment of full automation, who instead express some preferences to merely envisage automation mechanisms as tools that support human focused solutions such as those seen at level 2, partial automation.

In particular, there are some highlights after testing several level 3 CD&R scenarios that included some situations where the automation did not capture all the conflicts or could not successfully provide a suitable conflict resolution action. In these cases, the users were invited to *recovery control* and *intervene* and to provide suitable clearances of their own, or to identify conflicts that cannot been found for some reason. In these circumstances, it has been observed that the ATCOs are aware of the situation and are able to provide suitable actions that would resolve the issue when requested. Not only that, their situational awareness seems not to be degraded, even when performing a monitoring role at level 3.

However, and despite these preliminary results, there is still huge room for experimentation in this regard by using different scenarios and situations. What it draws attention is that, even at this primitive experimentation stage, there is already an *overall agreement* between ATC experts in one important aspect. Full automation level 3 *does not seem feasible* to implement and deploy nowadays in the real world. Experts think that if automation of this type, involving *control recovery*, was deployed in real scenarios, and given the safety critical nature of the CD&R functions, there are strong reasons to think that safety issues would be seriously limiting. More specifically, there would be a large risk that over a sustained period, ATCOs may become *'de-skilled'* which could lead to issues when trying to recover following unexpected failures of the automation. As an example, performing a simple monitoring task alone may result in ATCO loss of expertise in the controlling tasks and in case the XAI fails, and the ATCOs might not have the capability to recover control in complex situations in a safe manner. This would also have a significant negative impact to situational awareness, which could lead to issues when trying to recover following unexpected failures of the automation. This situation brings into focus the debate about 'human-automation teaming' concepts compared to a full 'human out of the loop'

situation. At this stage, it is clear that the TAPAS research is focused in the former environment and looks to determine how explainability can help the human work in a better fashion with automated support tools, particularly if those tools are also 'able to learn and adapt on the job' – a key characteristic of AI and ML applications. Nevertheless, both this and a longer-term full automation scenario are highly interesting topics for future research.

For example, there exists a risk that this *'de-skilling'* of the operators in higher levels of automation could be produced in quite short periods of time following deployment. Therefore, in the aforementioned scenarios, there are real chances that, after several months of such a major paradigm change, and/or for traffic situations beyond a certain threshold/complexity, the human would no longer be able to step back in and regain control.

This certainly would represent a disruptive change to the existing ATC/ATM process and, if implemented, will surely require that an entirely *new set systems* and/or processes are developed to provide a safe and reliable back-up.

7.2.4.3 Safety at Partial and Full Automation

- The explainability and transparency functionalities provided by the combination of the XAI prototype and the VA tool seem to offer acceptable levels of safety. -

In ATFCM scenarios, in general, *safety remains at an acceptable level*, in a qualitative way, for levels 2 and 3 of automation. However, due to their different nature, there are remarkable differences in how FMPs perceive such safety in when working with partial or full automation.

For example, in level 2 of automation, FMPs seem to experience some difficulties when applying diverse 'what-if' tests and finally selecting only a sub-set of solutions, combined with the inability to then re-run the XAI to try to solve the remaining issues. However, some transparency-related elements such as the additional information available via the co-located VA display and its drill down features facilitate the correct identification of issues and comprehension of the proposed solutions, *with little human effort*. In addition, for the rest of operations, the operational experts generally perceive applying the proposed solutions as something straightforward, which certainly contributes to *safety*.

Working with full automation, at level 3, the paradigm changes. It turns out that, at this level, a larger number of proposed solutions can be applied in a very short execution time, as the implementation time of a solution does not mostly rely on a human capability. In addition, the use of the VA display can accurately support with plenty of additional information to help humans understand what actions are taken.

Even though some additional effort is sometimes required to comprehend the effect of 'induced' Hotspots, created by previous actions, humans are generally able to consider many more solutions and maintain *good levels of situational awareness*, even considering the fact that those solutions are automatically applied by the XAI. This paradigm of working and the solutions provided are perceived by FMPs to be sufficiently *safe*.

Regarding the CD&R scenarios, *ensuring safety* is even more critical than in the previous use case. It was observed that the functionalities provided by the prototype are acceptable to maintain an adequate level of safety. In particular, the integration of the prototype with the ATC platform allows users to access information in real time, thus *increasing their situational awareness* on the evolution of flights and the set of possible conflicts.

Still, some improvements are needed in the prototype, like the need for a specific action to resume the FPL after a resolution action. In general, the prototype is able to provide all the needed information for the ATCO to perform their tasks and guarantee that no collision is provoked.

7.2.5 Certification

7.2.5.1 Tailoring Explanations for Certification

- The depth and level of detail in transparency and explainability aspects required for certification is not trivial, and this must be carefully devised. -

Certification is a required step towards the wide use and application of explainable and transparent AI in ATM. This aim must always be present when designing and working with this type of technology. What implications can we foresee from this statement?

An important aspect to consider is the different level of explanations and transparency that is required at different stages. For example, it is logical to think that different levels and depth of explanations will be required depending on whether it is a training stage for controllers on the use of tools or real-time operations. In these cases, it will be useful to provide a type of transparency tailored to maximise the absorption of this information, as well as a more specific, and more detailed, language and vocabulary.

However, the explanations and transparency intended to work towards certification may be completely different. For certification processes, it will probably not be as relevant to provide explanations or transparency with very specific and concrete details, at a very low level. The language used will also be different, and explanations are likely to focus on other aspects, such as possible consequences and repercussions of impact on safety issues.

Therefore, different layers of abstraction in providing explanations and transparency would be useful depending on what the main region of work is. This would in a way mean tailoring explanations to these needs. Clearly, this adjustment process will also have a cross-cutting impact on the way transparency requirements are defined, depending on the aspects mentioned above. In other words, the main changes to be made to allow for these different types of explanations will *have to be made to the initial transparency requirements*. So, explicitly asking for different levels of depth in the transparency requirements would eventually allow this tailoring process.

For example, if the proposed solutions are self-explanatory, this means that the explanations will not provide much value in terms of transparency and explainability. Little to note here. Now let us assume that we are working in an operational phase. Here, it is expected that the proposed solutions, although complex in many cases, can be understandable and explainable to a human to some extent. However, if we are working at a *certification/testing* stage, this could also include cases where the proposed solutions are too complex for an average human to process and understand correctly. It is in these cases that the human could end up transcending their inherent need for explanations. Explanations would no longer be so necessary and what the operators would really need would be to have full confidence in the tools they are using. If we get to these cases, then we might be close to glimpsing that the solutions we are dealing with are close to being certifiable.

7.3 Recommendations

7.3.1.1 Seeking Accuracy in Transparency Requirements

- ***Transparency requirements must seek to be unambiguous, concise and clear.*** -

In general, for a transparency requirement to be *unambiguous* and not give rise to confusion in its reception and interpretation, it must contain univocal, clear and concise information and/or instructions. Good transparency requirements must mostly contain information on transparency and explainability functionalities to be implemented.

These should focus on concisely describing how to provide transparency about a given process. For instance, this can be achieved by providing the user with:

- Certain *parameters* and *values*
- Representative *indicators* for obtaining a solution
- Lists or *rankings* of importance
- High impact *variables* or *features*
- Etc.

All these elements will significantly contribute to obtain more details about the inner functioning of a model, or the decision-making process behind it. Some other times an *explanation*, if *simple* and *clear* enough, will give rationale about processes more effectively and can justify the reason for a system's decision. Sometimes humans are overwhelmed with lot of details, information, and parameters. It is certainly easy to get lost in the details, and in some situations an explanation will contribute in a greater extent to understanding than having a multitude of available parameters informing about a process. The ultimate goal is to *stimulate* and *enhance* human understanding, not to damage it.

7.3.1.2 Not Mixing Up Requirements Definition with Validation Aspects

- ***Transparency requirements must merely contain technical/functional aspects, and not those to be later acquired during experiments and validation.*** -

A good general practice on the definition of transparency requirements is that these must not contain information about aspects that will be demonstrated later after validation of the implemented prototype. For example, it should not be stated nor assumed in the definition of the requirement how the recipient of the requirement will eventually interact or feel respect to it. The reason is because drawing conclusions on how useful or transparent the functionality provided by a requirement will be perceived by a human, must solely come from the validation results.

Suppose that a requirement contains a technical functionality to be implemented, and in turn some instructions on how the transparency should be received by the end user. For example, it would be equivalent to say: *'provide explanations about X so that the end user quickly understands Y*'. What would happen if the final prototype really provides that functionality (explanations) but the user is

unable to understand it? Can we consider in this case the requirement as partially, or completely fulfilled? However, if only information related to technical functionalities is included in the definition, then the evaluation of compliance with the requirement will be more straightforward, and it will not depend on aspects to be demonstrated during validation. Mixing in this case is not a good idea.

7.3.1.3 Vocabulary Matters

- ***Words and expressions that must be avoided in the general definition of transparency requirements.*** -

The use of certain vocabulary in a transparency requirement should be avoided. In particular, terms such as '*trustable*', '*interpretable*', '*understandable*', '*comprehensive*', '*quickly stimulate*', 'clearly', etc; have demonstrated not to be appropriate nor useful when defining transparency requirements.

The main reason to this assertion is that demonstrating that something provides such characteristics is not straightforward, it can lead to ambiguity, and requires extensive validation processes which are in fact also subject to a great deal of *subjectivity*.

However, if any of these terms were going to be used, it must be explicitly complemented by adequate *descriptions* and adequately supported by *quantifiable criteria*. These terms are highly susceptible to subjective interpretations and their definitions must be univocal.

Transparency requirements should seek to avoid ambiguities and personal interpretations, as this will greatly facilitate their technical implementation and final verification. Vocabulary is important.

7.3.1.4 Refining and Iterating on Transparency Requirements

- ***Refining, iterating and updating transparency requirements is sometimes necessary and a good practice.*** -

This is one of the major lessons extracted from TAPAS and this certainly will benefit XAI prototype maintenance and continuous enhancement. At the beginning of this research, the concept of transparency requirement was something very new. To the best of this document's author knowledge it did not exist before the development of TAPAS project. Therefore, the definition of a new concept from scratch is highly subjected to error, revision, and continuous improvement.

Generally, over the development of a project many lessons are learned. This project is no exception and many interesting concepts and ways of improving processes have arisen as we experimented with prototypes and talked to diverse experts. What we learned is when it comes to the definition of transparency requirements, in case that these have been proved as not being relevant enough, *requirements must be refined and improved*.

Transparency requirements are a powerful means to promote transparency in the solutions given by an XAI automation tool. If the sought transparency functionalities do not provide enough quality, in shape or content, *refinement is needed*.

If we end up realizing that there are missing transparency functionalities that should have been integrated and deployed in a XAI system, *new transparency requirements must be designed and included for new implementation*.

Therefore, the maintenance and enhancement of XAI prototypes *should always be subject to potential updates* by means of a tailored refinement of the originally defined transparency requirements that allowed the prototype's conception and birth.

7.3.1.5 Transparency Requirements Must be Ultimately Understood

- ***Ensuring that transparency requirements are understood by developers is essential.*** -

Mainly focusing on explanations, by experimentation it is suggested that, if these were to be requested in transparency requirements, they must be accompanied by a mechanism to ensure that not only the explanation is offered, but also *understood by the recipient*. Such a mechanism shall be sufficiently robust to ensure that, if a minimum level of interpretability/actionability is not achieved by the offered explanation, the transparency requirement should not be considered as fulfilled because the explanation still requires refinement.

Otherwise, the explanations are at risk for being *irrelevant* or *meaningless*. A greater interpretability could be achieved if the explanation is adequately complemented with any *metric* capable of *quantifying* the degree of satisfaction of the recipient with respect to the explanation. Consequently, good transparency requirements should include not only the functionality of offering the explanation, but also *ensuring* that this meets a minimum degree of interpretability.

These mechanisms are not implemented in TAPAS due to the limited time assets available for the project, but this lesson is worth considering for future developers.

7.3.1.6 Evaluation of the Level of Compliance

- ***Defining accurate compliance criteria is key to assess the level of transparency requirements compliance achieved by an XAI prototype.*** -

A major learnt lesson is regarding the evaluation of the fulfilment of the transparency requirements achieved by the implemented prototypes. Ideally, this process should be as objective as possible, and based on univocal and well-defined criteria.

It is outlined that the definition of a set of *compliance criteria* would facilitate this process. Therefore, along with the definition of a transparency requirement, some corresponding compliance criteria must

be provided. These, once the XAI prototype is implemented, will allow to effectively assess if such transparency aspect is fully or partially met, and even unmet by the prototype. Particularly, these criteria must thoroughly specify under what conditions a transparency should be considered as fully or partially met, and unmet. Also, depending on the nature of every transparency requirement, each one might need different evaluation or compliance criteria. Compliance criteria must be carefully designed and tailored to each of the transparency requirements.

Defining these criteria in advance, before the implementation of the prototype, will also benefit the objectivity of the process, as the final decision of fulfilment would not only rely on a human's assessment (sometimes subject to undesired biases), but on well-defined criteria, values and parameters. In general, conducting this transparency requirements verification process this way will naturally result in a *clearer*, *simpler* and more *efficient* manner.

Understanding adequately the final compliance with transparency requirements is essential for XAI prototype iteration, maintenance and improvement. Attention must be paid to this process.

7.3.1.7 Automation Level and Transparency Requirements

- Should the automation level be considered as impact factor when defining requirements?-

In this regard, we have not identified and extracted significant information to conclude anything relevant yet. The idea that transparency requirements should include more details and capabilities (or even quantity) when requesting explanations whether it is a partial automation (where decision-making depends on the human) or full automation (where supervision is the key), may be certainly appealing but more experimentation is still needed.

There is no strong evidence in this regard. However, it is glimpsed that requirements must focus on requesting more transparency and explainability aspects when these are aimed to be used as part of a tool functionality during *training* and *familiarization* stages of such tool. The automation level itself, therefore, is not proved yet to be an impact factor when defining transparency requirements, and specific experiments are needed to draw conclusions.

7.3.1.8 Transparency, Explainability and Automation Recommendations

- Some other recommendations regarding transparency, explainability and automation. -

These are some other recommendations collected from experienced operators during experiments. They mostly refer to transparency, explainability and automation, and therefore they attempt to serve as general guidelines to be considered by future XAI developers.

- ✓ *Ranking the best proposed solutions is something valuable*. When multiple solutions are possible for the same issue, a clear ranking of the solutions from best to worst is very valuable to the user. This is particularly important in time-constrained, safety critical situations typically seen in CD&R scenarios.

- ✓ When actions are being automatically performed by the XAI, *their current status should be clearly communicated*. For example, whether those actions are pending, in progress or completed.
- ✓ *If multiple actions have the same ranking it should be clarified whether both actions must be performed or not*. Sometimes two actions are too many (e.g. to communicate and monitor) and other times two actions with less impact to each recipient can be of added value in terms of capacity / efficiency / equity etc.

8 Conclusions

This work presents a *framework* (See Appendix C) that contains instructions on the use and application of explainable AI when the current levels of automation in the ATM domain are increased. Its main contribution is a set of principles for the transparent use of this type of technology in highly automated scenarios. More specifically, these principles contain new knowledge, purely derived from experimentation in real scenarios, which tries to lay the foundations, serve as a guide, and pave the way for a correct and efficient application of XAI technology in ATM. The operational scenarios chosen for experimentation are ATFCM and CD&R. These principles are addressed to a wide target audience, such as XAI developers, end users and operators, designers to certification agencies.

Throughout the process followed to derive such principles, numerous lessons have been learned. First of all, it was necessary to design, implement and deploy prototypes that, using XAI and VA technology, would allow increasing the automation of different ATFCM and CD&R scenarios. In addition, this automation had to be carried out in the most transparent way possible to the user. The initial premise was that if the user is able to understand how this automation had been carried out internally, and why certain solutions were offered, this would ultimately make it easier to build trust in the use of these tools. In safety-critical applications it is essential to strongly trust the tools being used.

In order to guide the technical development of these XAI/VA prototypes so that the provision of transparency to the user was optimal, requirements were needed. We named these requirements as '*transparency requirements*', as they attempt to indicate the prototype developers what information needs to be provided in order for the user to understand the reasons behind the solutions and decisions offered by the automation tools. A total of 94 requirements were provided to guide the needs in terms of provision of transparency by the two built prototypes, the ATFCM and CD&R. Therefore, the *first contribution* of this document is to provide guidelines for the definition of accurate transparency requirements. This includes not only the transparency requirements themselves, but a set of *lessons and recommendations,* obtained from our experiences, which might be worth considering by future XAI developers. The process of understanding how to write optimal requirements was not trivial, full of iterations, changes and discussions. To the best of our knowledge, the concept of transparency requirements was new in XAI so it had to be built from scratch, and this involved research, extensive trial and error, discussions with experts and continuous improvement. Due to the exploratory research nature of this project and the limited time available, this process of definition of transparency requirements still may need maturation and improvement. However, it might already serve as a first basis for future additions.

Principles, in almost all the domains of knowledge, are built from experimentation. TAPAS project was not exception and, in order to generate valuable principles for the use and application of transparent automation in ATM, experimentation was needed. Our experiments consisted in selecting a group of human experts, skilled and experienced in real ATFCM and CD&R operations and making them interact with the developed prototypes. From these experiments a huge amount of results and valuable information were collected. Due to the volume of the information collected from the experiments and the feedback given by experts, an efficient way of information organization was needed. Principles needed to be envisioned and generated from these results, and at the beginning everything looked like a big cloud of messy concepts. *The biggest challenge of this delivery has been information classification and synthesis of new knowledge*. Consequently, to remove redundant information and to adequately order, categorize, simplify, relate and combine and form more complex pieces of

information, an efficient mechanism was needed. Therefore, the *second contribution* of this document is the process followed to combine basic pieces of information, or '*insights*' as we called them, into bigger and more complex information pieces. This would eventually facilitate the synthesis of new concepts from the available information extracted by the experiments. Although we do not explicitly provide the process followed to combine the insights, we do present the insights themselves so that the reader can track them and better understand from where the content of the principles exactly comes from. That is, by simply reading the final principles, the reader should be able to relate specific paragraphs or sentences, to some of the presented insights over the document. This process was not perfect, but it allowed us to classify information in a very efficient manner.

The third, and last contribution of this document, is a set of *principles* for the use of transparent AI in ATM automation. Principles in this document are meant as a way to *capture new knowledge* that has been synthetized from experimentation and validation. As it happened with the transparency requirements, we first had no clue on how to define principles, how a principle should look like, its format and presentation to the user, and even if they could be an efficient way to transmit knowledge. We invented a way to define principles based on three elements: a main header, a sub-header and the content itself. Each of the elements attempts to give a different level of information about the content of the principle. This was intentionally done in order to make the read as efficient and pleasant as possible. Principles are organized in 5 categories: *transparency and explainability, trust, acceptance, safety,* and *certification*. This classification is simply done based on the content of the principles and its possible application to these fields.

The writing of the principles, based on the extracted insights, has been tricky. It was sometimes arduous to use an optimal *vocabulary*, and to choose the right *syntax* that allows connecting words in such a way that can give the sensation of what a principle really wants to transmit. A principle is meant to convey findings, to stablish new ways that describe how a phenomenon works or behaves under specific conditions, or to define rules or laws. The extracted insights were not always in the right vocabulary nor the optimal syntax to express what a principle should, so it sometimes required extensive rephrase and reading.

These principles must be considered for further completion and future additions. Due to the limited duration of TAPAS project there are still interesting aspects to analyze in the future. A relevant one is to better understand how the fact of providing transparency and explainability about some automation processes might have real consequences on human efficiency. In other words, understanding the real effects and quantifying the impact of the quality of the transparency provided on human performance. For example*, how much the human efficiency can improve when using XAI respect to when only AI or other automation technologies are used*? *Can we design experiments and define metrics to measure that difference in performance?.* It is natural to think that the fact of offering transparency to users should have some positive effects and consequences on human efficiency, otherwise *what would be the real benefits of using XAI?* It would be extremely valuable to conduct specific experiments to test this and translate the results and findings into adequate principles to set some foundations on *how transparency impacts human efficiency*. This should perhaps be on the horizon as one of the major breakthroughs to achieve.

Finally, the extrapolation of this framework to other domains would be something important. The results obtained here might certainly seem highly dependent on the domain. However, if we scratch a little bit, it can be glimpsed that there may be some concepts on the use and application of transparent AI in safety-critical scenarios that could be applied and be relevant to other domains. These domains

should perhaps share some basic characteristics with the safety-critical nature of ATM, and changes would be needed, but hopefully this work could be a considered a basis for future applications to diverse domains.

9 References

[1] D2.1 TAPAS Use Cases Description.

[2] D2.2 TAPAS Consolidated Requirements and Functional Roadmap.

[3] D3.1 TAPAS Use Cases Transparency Requirements.

[4] D5.2 TAPAS Validation Report.

[5] Newton, Isaac, 1642-1727. 'Philosophiæ Naturalis Principia Mathematica'. Londini :Apud G. & J. Innys, 1726.

[6] J. M. Steele, University of Toronto, (review online from Canadian Association of Physicists) Archived 1 April 2010 at the Wayback Machine of N. Guicciardini's "Reading the Principia: The Debate on Newton's Mathematical Methods for Natural Philosophy from 1687 to 1736" (Cambridge UP, 1999), a book which also states (summary before title page) that the "Principia" "is considered one of the masterpieces in the history of science".

[7] Wicander, R. and Monroe, J.S. (2006) Essentials of Geology. 4th Edition, Thomson Brooks/Cole, Belmont, 239.

[8] Monroe, J.S. and Wicander, R. (2012) The Changing Earth, Exploring Geology and Evolution. 6th Edition, Brooks/Cole CENGAGE Learning, Belmont, 253.

[9] Blaise Pascal, Traitez de l'Equilibre des Liqueurs (Treatise on the Equilibrium of Fluids), Paris, 1663.

[10] Georg Christoph Lichtenberg. 'Philosophical Writings'. 223 pages. 2012.

[11] Heisenberg, W., "Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik", Zeitschrift für Physik, 43 (3–4): 172–198, 1927.

[12] Cambridge Dictionary. https://dictionary.cambridge.org/es/. Last time accessed: 06/08/2021

[13] Oxford Languages. https://languages.oup.com/google-dictionary-en/. Last time accessed: 06/08/2021

[14] European ATM Master Plan, Edition 2020.

Appendix A - Schemas

Validation Objective	Sub-Focus	Sucess Criterion	Result
OBJ 1 - Identify principles for Transparency of AI-based solutions	1.1 Determine how much additional information is needed at automation levels 2 & 3 to ensure that the human operator is able to make informed decisions to help solve ATM problems.	VA and explanatory support information is clear and understandable and the tools are able to provide the required information at the right time.	Partially OK
	1.2 Identify when support information is required, what level of detail is needed and how should it be provided in a timely manner.	Key data that can be easily understood by the human has been identified that supports transparency needs and is provided in the required time frame and at an appropriate frequency	Partially OK
	1.3. Evaluate areas where the levels of transparency may need to be improved	Information that is unavailable but could help during the use of the proposed XAI has been identified and catalogued for future analysis.	OK
	1.4.Propose suitable methods by which the level of understanding and trust in the AI automation can be measured	Questionnaires, 'over-the-shoulder' observation and debriefing analysis metrics have been identified to support the necessary measures.	OK
OBJ2 - Develop prototype XAI/VA methods for ATM use cases to address transparency at various levels of automation	2.1 Produce customised VA views to support transparency and explanatory information to the human operator at different levels of automation.	VA display tools are able to consume data provided by the XAI component to support interactive drill down views for the human operator	OK
	2.2 Assess how the VA methods can help enhance operator understanding and trust in AI-based automation	Elements provided in the VA provide clear visual evidence related to the actions being performed by the XAI tools	Partially OK
	2.3 Evaluate the effectiveness of the transparency solutions being deployed	Human operators are able to use the visualisation to interrogate the on-going scenario and solutions being considered	OK
	2.4 Determine the different needs for transparency at different automation levels	Human operators classify the information being provided and confirm that it is sufficient to explain the decisions being made	Partially OK
	2.5 Evaluate the level of understanding and situational awareness of the human as the automation proposes / implements solutions	Human operators are able to describe what the automation is doing and why solutions have been proposed	OK
	2.6 Verify that the human can successfully take over and recover control of the situation if the automation fails for any reason	The human was able to either take over and complete the current task when automation failed,	NOK
	2.7 Ensure that the human is able to identify and resolve any remaining issues at the end of the XAI process, if present.	The human operator was able to identify and to complete any remaining issues that were not successfully solved at the end of the process	NOK
	2.8 Demonstrate how transparency can promote operational and social acceptance of 'black-box' AI solutions	The operator confirms that the solutions provided by the XAI were fit for purpose	Partially OK
	2.9 Assess shortfalls and areas where transparency can be improved in future solutions	Operational experts identify areas where information was insufficient to support understanding	Partially OK
	2.10 Identify opportunities for additional training	Additional training or processes to enhance the ability for the XAI/VA to assist the human in understanding the process at different automation levels has been identified by the team	OK

Figure 7. Summary of the TAPAS ATFCM validation results per validation objective and criterion. Achieved sub-focuses are in green, partially achieved in orange, and non-achieved in red.

Appendix B - Schemas

Validation Objective	Sub-Focus	Sucess Criterion	Result
OBJ 1 - Identify principles for Transparency of AI-based solutions	1.1 Determine how much additional information is needed at automation levels 2 & 3 to ensure that the human operator is able to make informed decisions to help solve ATM problems.	VA and explanatory support information is clear and understandable and the tools are able to provide the required information at the right time.	OK
	1.2. Identify when support information is required, what level of detail is needed and how should it be provided in a timely manner.	Key data that can be easily understood by the human has been identified that supports transparency needs and is provided in the required time frame and at an appropriate frequency	OK
	1.3. Evaluate areas where the levels of transparency may need to be improved	Information that is unavailable but could help during the use of the proposed XAI has been identified and catalogued for future analysis.	OK
	1.4.Propose suitable methods by which the level of understanding and trust in the AI automation can be measured	Questionnaires, 'over-the-shoulder' observation and debriefing analysis metrics have been identified to support the necessary measures.	OK
OBJ2 - Develop prototype XAI/VA methods for ATM use cases to address transparency at various levels of automation	2.1 Produce customised VA views to support transparency and explanatory information to the human operator at different levels of automation.	VA display tools are able to consume data provided by the XAI component to support interactive drill down views for the human operator	OK
	2.2 Assess how the VA methods can help enhance operator understanding and trust in AI-based automation	Elements provided in the VA provide clear visual evidence related to the actions being performed by the XAI tools	Partially OK
	2.3 Evaluate the effectiveness of the transparency solutions being deployed	Human operators are able to use the visualisation to interrogate the on-going scenario and solutions being considered	OK
	2.4 Determine the different needs for transparency at different automation levels	Human operators classify the information being provided and confirm that it is sufficient to explain the decisions being made	OK
	2.5 Evaluate the level of understanding and situational awareness of the human as the automation proposes / implements solutions	Human operators are able to describe what the automation is doing and why solutions have been proposed	Partially OK
	2.6 Verify that the human can successfully take over and recover control of the situation if the automation fails for any reason	The human was able to either take over and complete the current task when automation failed,	Partially OK
	2.7 Ensure that the human is able to identify and resolve any remaining issues at the end of the XAI process, if present.	The human operator was able to identify and to complete any remaining issues that were not successfully solved at the end of the process	Partially OK
	2.8 Demonstrate how transparency can promote operational and social acceptance of 'black-box' AI solutions	The operator confirms that the solutions provided by the XAI were fit for purpose	Partially OK
	2.9 Assess shortfalls and areas where transparency can be improved in future solutions	Operational experts identify areas where information was insufficient to support understanding	OK
	2.10 Identify opportunities for additional training	Additional training or processes to enhance the ability for the XAI/VA to assist the human in understanding the process at different automation levels has been identified by the team	OK

Figure 8. Summary of the TAPAS CD&R validation results per validation objective and criterion. Achieved sub-focuses are in green, partially achieved in orange, and non-achieved in red.

Appendix C – TAPAS Framework for the transparent use of AI/ML automation in ATM

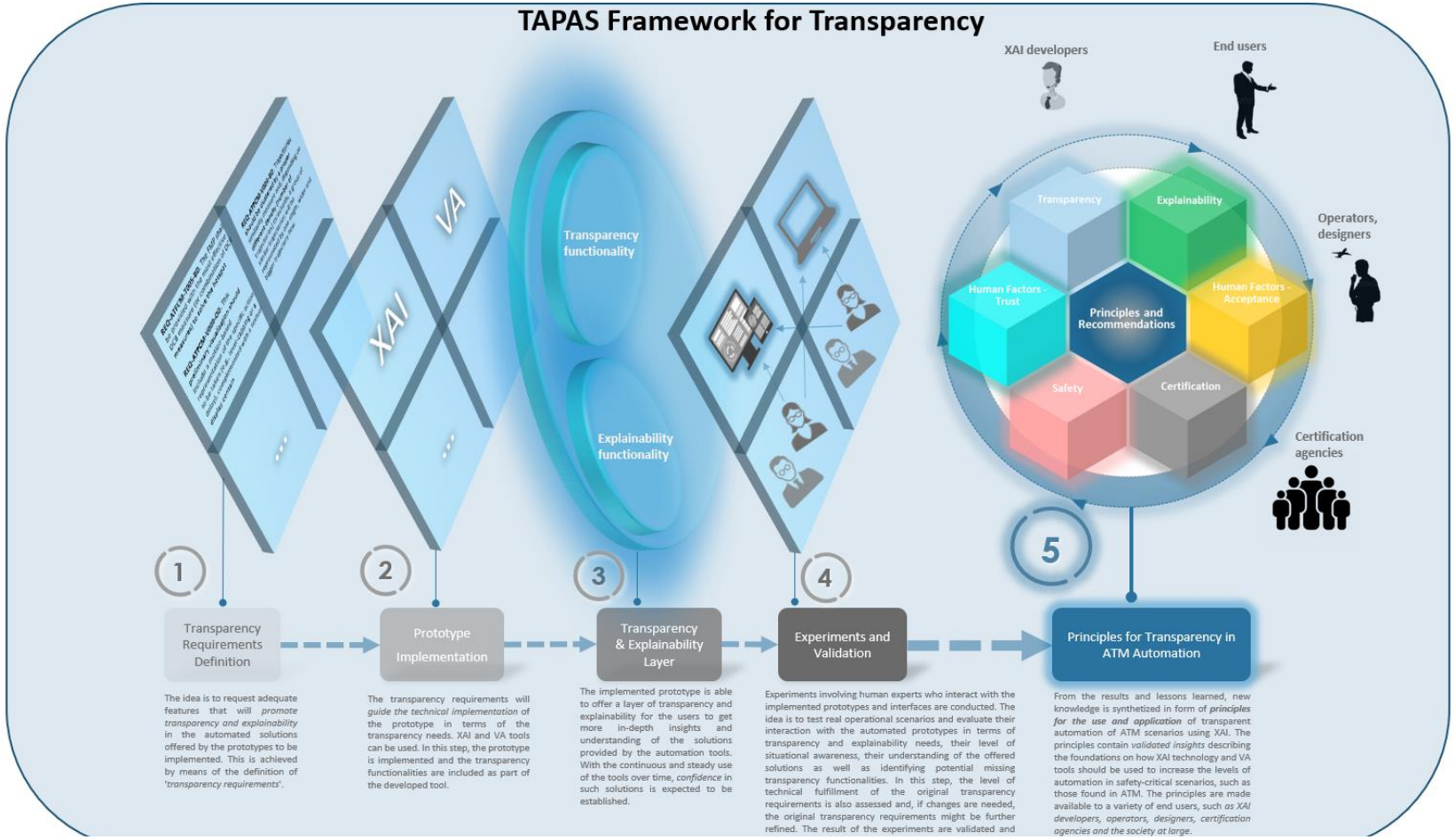

Figure 9. TAPAS framework for the transparent use of AI/ML automation in ATM.