Transparency & Explainability in higher levels of automation in the ATM domain

Natividad Valle, M Florencia Lema, José Manuel Cordero,
Enrique Iglesias , Rubén Rodríguez
Centro de Referencia I+D+i en ATM (CRIDA)
Madrid, Spain
nvallef, mflema, jmcordero, eiglesiasm, rrodriguezr
@e-crida.enaire.es

Gennady Andrienko, Natalia Andrienko
Fraunhofer Institute IAIS
Sankt Augustin, Germany
gennady.andrienko, natalia.andrienko
@iais.fraunhofer.de

George A. Vouros, Theocharis Kravaris,
George Papadopoulos, Alevizos Bastas, Georgios Santipantakis
University of Piraeus Research Center Piraeus, Greece
georgev@unipi.gr

Ian Crook, Sandrine Molton
ISA Software
Paris, France
ian, sandrine
@isa-software.com

Antonio Gracia-Berna
Boeing Research & Technology Europe
Madrid, Spain
antonio.graciaberna@boeing.com

Abstract— **This paper presents findings, lessons learnt and guidelines for the use of explainable and transparent Artificial Intelligence (AI)/Machine Learning (ML) in ATM. The paper focuses on the results obtained from validating two AI/ML prototypes for Conflict Detection & Resolution (CD&R) and Air Traffic Flow and Capacity Management (ATFCM) problems. These two prototypes are representative of the type of advanced automated systems that can support respectively the tactical and the pre-tactical operational phases The aim is, shifting the paradigm of human-AI teaming, providing full explainability and operational transparency. The major question is: when and how explanations should be provided for systems to be acceptable and trustworthy by operators?**

Transparency; Explainability; AI/ML; CD&R; ATFCM

I. Introduction

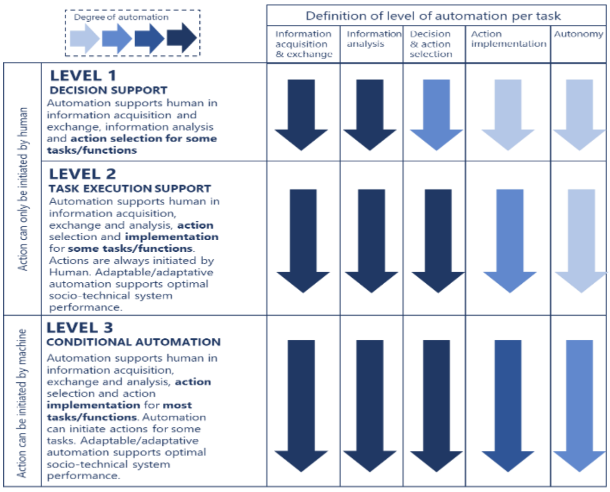
With the advances in computing power that have been experienced in the last 5-10 years, the application of AI and ML techniques is becoming commonplace for solutions where automated support is concerned. This paper describes the process followed to address the effectiveness of introducing AI/ML solutions to increase the levels of automation in ATM, considering the operator relinquishes certain tasks to the system. The main objective is to explore AI/ML explainability/transparency for automated systems to be acceptable and trustworthy by ATM operators. In so doing, this article provides details on the exploration followed, as well as on findings and lessons learnt by the SESAR exploratory research project TAPAS [1]. This includes the presentation of two use cases (CD&R and ATFCM), the description of the AI/ML prototypes developed for each one of them, the details on the validation activities performed, and the most remarkable conclusions gathered from the conducted verification tests and Real Time Simulations (RTS).

This study aims at paving the way for the deployment of AI/ML technologies in ATM environments, in particular, in automation levels 2 and 3 as expressed in the successive editions of the European ATM Master Plan [2]. In that sense, the paper provides not only principles and criteria for explainable/transparent systems, but technical lessons learnt as well regarding the selection and application of AI/ML methods in the context of the two use cases.

II. Preliminaries

A. Levels of Automation: Goals, issues and questions

TAPAS project considers as basis the levels of automation as defined in the European ATM Master Plan and adopted by the SESAR programme [2]. Although the objective is not to amend these, the work done provides some insights on this topic.



Degree of automation
Definition of level of automation per task
Information acquisition & exchange
Information analysis
Decision & action selection
Action implementation
Autonomy
Action can only be initiated by human
Action can be initiated by machine

LEVEL 1
DECISION SUPPORT
Automation supports human in information acquisition and exchange, information analysis and **action selection for some tasks/functions**

LEVEL 2
TASK EXECUTION SUPPORT
Automation supports human in information acquisition, exchange and analysis, **action** selection and **implementation** for **some tasks/functions**. Actions are always initiated by Human. Adaptable/adaptative automation supports optimal socio-technical system performance.

LEVEL 3
CONDITIONAL AUTOMATION
Automation supports human in information acquisition, exchange and analysis, **action** selection and action **implementation** for **most tasks/functions**. Automation can initiate actions for some tasks. Adaptable/adaptative automation supports optimal socio-technical system performance.

Figure 1. Extract of Automation levels defined in the European ATM Master Plan [2].

ATFCM (Air Traffic Flow and Capacity Management) and CD&R (Conflict Detection and Resolution) automation developments have two automation levels (level 1 and level 2) focusing on increasing the level of system support, while the initiation of actions always remain with the human. The breakthrough happens in automation level 3, when higher automation levels put the human in a monitoring position instead of a leading position in the decision-making process.

The applicability and acceptance of such automation systems is currently limited by their lack of explainability of decisions and actions to humans, especially in the certification and training phases. The work presented in this paper aims at contributing to this in ATM operational environments, to help building principles to facilitate the adoption of these technologies in the ATM domain.

B. *Explainability, Transparency and Trustworthiness in AI/ML*

The concept of trustworthiness in AI/ML has special relevance in the context of ATM. In June 2018, the European Commission (EC) set up a High-Level Expert Group on AI with the objective of supporting the implementation of the European strategy on AI. Later, in April 2019, this group of experts proposed the following seven key requirements for trustworthy AI [3]: accountability, robustness and safety, oversight, privacy and data governance, non-discrimination, environmental well-being and transparency.

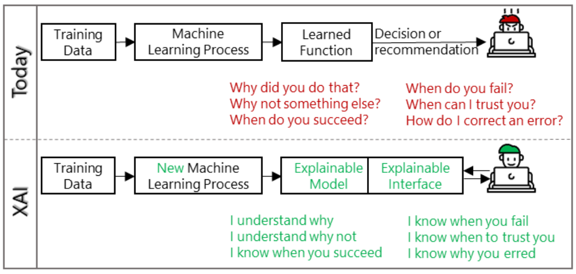


Figure 2. EXplainable AI (XAI) vs todays' automation without Explainability.

This last requirement of transparency is the focus of this paper. In particular, the paper goes beyond transparency and explores explainability, which goes a step further contributing to the interpretability of an action given by an AI system. Going deeper into definitions [4], explainability seeks to provide valuable information to the user on the inner mechanisms of AI-based models. Such explanations may provide comprehensible insights on aspects such as what the system is doing, why it made certain decisions over others, as well as give intuitive rationale for certain solutions that might seem counterintuitive, at first, to the human operator. From these concepts, it could be concluded that achieving higher explainability levels can be straightforwardly related to a higher comprehension and trust to the outcomes given by the system. But there is a further step: Transparency, that relates to the AI system's ability to produce effective explanations by means of proper visualization, text or examples, making such explanations comprehensible to humans with diverse expertise and support their reasoning w.r.t. operational constraints.

C. *ATFCM and CD&R use cases*

Two different use cases are explored: ATFCM and CD&R. The first use case deals with a pre-tactical timeframe (one day before the operation day, D-1) and focuses on the Flow Management Position (FMP) role of an ACC. Whilst the CD&R use case stays in the tactical horizon (the day of operation, D) and focuses on the role of the executive air traffic controller (EC). The selection of those use cases allows the exploration of different time horizons, along with different necessities regarding safety and time criticality.

In the ATFCM exercise the prototype supports the FMP in the detection and solution of the hotspots (imbalances between demand and capacity). In contrast with the current operation paradigm, here the AI/ML system tries to solve all hotspots at the same time by applying demand measures (regulations, level capping), instead of using capacity measures (new sectorization) alongside with the demand measures and analyzing a single sector at a time.

In the CD&R exercise, the prototype supports the EC in the detection and resolution of conflicts, and monitoring of non-conformances in a similar way to today's operation.

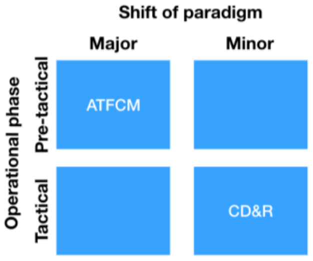
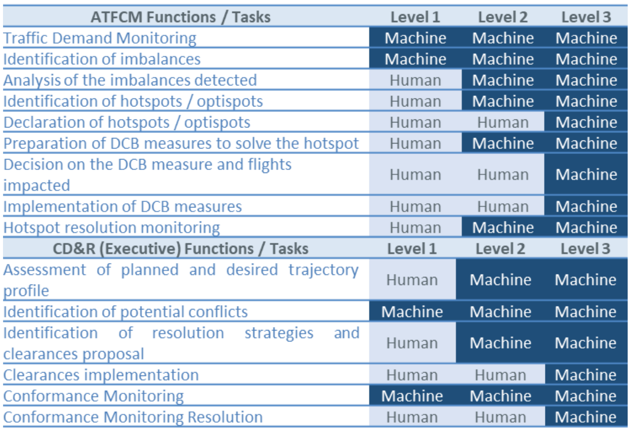


Figure 3. Space explored.

Depending on the automation level, some tasks are initiated by the human and others by the machine, as shown in *Figure 4*.



ATFCM Functions / Tasks	Level 1	Level 2	Level 3
Traffic Demand Monitoring	Machine	Machine	Machine
Identification of imbalances	Machine	Machine	Machine
Analysis of the imbalances detected	Human	Machine	Machine
Identification of hotspots / optispots	Human	Machine	Machine
Declaration of hotspots / optispots	Human	Human	Machine
Preparation of DCB measures to solve the hotspot	Human	Machine	Machine
Decision on the DCB measure and flights impacted	Human	Human	Machine
Implementation of DCB measures	Human	Human	Machine
Hotspot resolution monitoring	Human	Machine	Machine
CD&R (Executive) Functions / Tasks	**Level 1**	**Level 2**	**Level 3**
Assessment of planned and desired trajectory profile	Human	Machine	Machine
Identification of potential conflicts	Machine	Machine	Machine
Identification of resolution strategies and clearances proposal	Human	Machine	Machine
Clearances implementation	Human	Human	Machine
Conformance Monitoring	Machine	Machine	Machine
Conformance Monitoring Resolution	Human	Human	Machine

Figure 4. Tasks allocation roadmap between machine and human.

III. Prototype Systems

To test the previous task allocation roadmap, as well as to extract a set of principles and criteria for transparency and explainability, different prototypes were developed for both use cases, ATFCM and CD&R. Those prototypes were built according to technical and operational requirements that were defined with the aid of operational experts and further refined through the conduction of several workshops with them in an iterative approach.

A. Air Traffic Flow and Capacity Management (ATFCM)

The overall architecture of the ATFCM prototype is shown in *Figure 5*. It comprises an AI/ML module enhanced with functionality for the provision of explanations, and two other components: the Visual Analytics (VA) tool, that provides data exploration facilities, presenting the proposed ML solutions and explanations on those solutions; and the FMP client, a Demand & Capacity Balance (DCB) tool that allows the FMP to monitor the sectors, define hotspots, create demand measures, and perform what-if query on those hotspots. This FMP tool is also connected to the Innovative Network Operations Validation Environment (INNOVE) [5] platform where the sectorization and flight data is uploaded. The FMP tool provides information on the sectors, demand charts (OCC, HEC) and allows the creation of hotspots and demand measures.

The data preparation component processes the source data sets to provide the trajectories reported in flight plans, associated with contextual information (sectorization, entry/exit points and times for each crossed active sector). This information together with the airspace configurations, their activity intervals and sectors' capacity thresholds are then feed into the AI/ML and the VA module.

The AI/ML module [6] deals with identifying the imbalances/hotspots (capacity exceeded 110%), preparing the DCB measures (regulations/ground delays, level capping) and selecting the flights impacted by those measures. To do that, the component implements a Deep Reinforcement Learning (DRL) [7] method using Deep Q-Networks (DQN) [8]. According to this agent-based methodology, agents are the flights that decide on additional minutes of ground delay to be taken at every time step of the simulation; the system simulates each 24 hours scenario with all flights, to detect all hotspots occurring. In the trials, the time step of the simulation was set to 10 minutes and the decision of agents on additional delay a number between 0 and 10 minutes. Agents may increase at any time step their delay by adding additional minutes, until they reach their maximum delay parameter that is set as a constrain. To decide on additional delay at every time step, each agent participating in the scenario (i.e., each flight crossing the airspace at any time during a 24 hour interval) calculates the demand per active sector that it crosses in the airspace, identifies the imbalances and hotspots, prepares the types of measures to be taken to resolve any hotspot, and finally, it individually decides on the DCB measures to be taken. Therefore, the problem is tackled as a whole and from the flight perspective, not from the individual view of the opened sector where there is an imbalance, which is the current operating method.

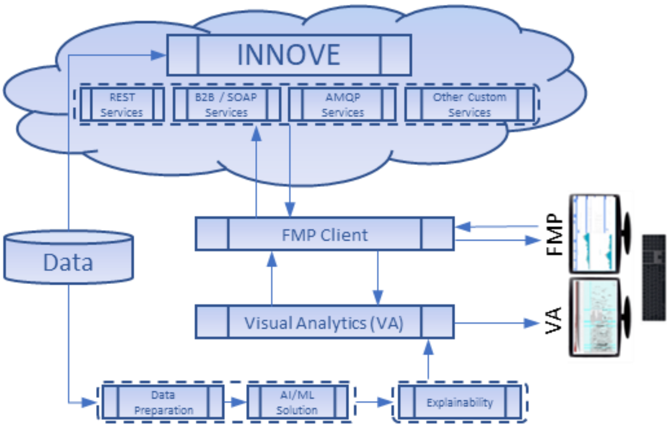


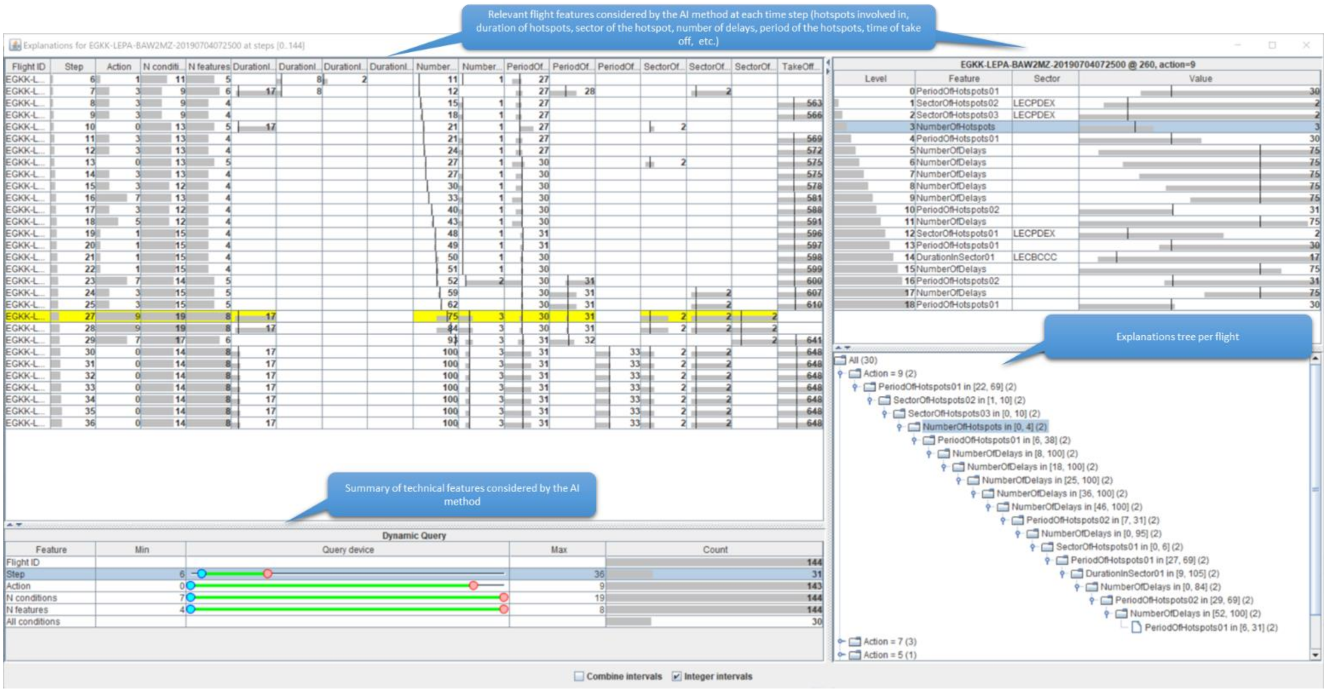
Figure 5. ATFCM prototype overall architecture.

To train the AI/ML module, information from the Spanish airspace during 2019 was used. In particular, the algorithm was provided with flight plan trajectories [9] from DDR (in the ALLFT+ format); the sector configuration including volumetry and declared capacity of sectors; the regulations applied, and flights affected by those demand measures over the Spanish airspace.

The solutions are provided to the Explainability component, which following the mimicking paradigm implements an XAI method through a Stochastic Gradient Tree [10], emulating the decisions taken from the DQN method. This module 'learns' how the DQNs is solving the problem and presents an inherently interpretable method by exploiting a decision tree. At this point the explanations include arguments on what is important for the decision of an agent at a specific time point during the simulation, and counterarguments on what it considers important to take an alternative decision.

These explanations and the enriched flight information go into the VA component (*Figure 6*). It presents to the operator an overview of the current situation (demand charts highlighting imbalances and delays) and allows the user to compare two or more scenarios involving the same set of flights, understand the process of the solution development (it provides the solutions for all simulation timesteps) and investigate the details for scenarios, sectors and intervals, including the decision tree on the explanations for a specific flight/agent.

This VA component runs in a secondary screen with the FMP tool running in another screen. Both consume the same data from the XAI algorithm.



Relevant flight features considered by the AI method at each time step (hotspots involved in, duration of hotspots, sector of the hotspot, number of delays, period of the hotspots, time of take off, etc.)

Explanations for EGKK-LEPA-BAW2MZ-20190704072500 at steps [0..144]

Flight ID	Step	Action	N conditi...	N features	DurationI...	DurationI...	DurationI...	DurationI...	Number...	Number...	PeriodOf...	PeriodOf...	PeriodOf...	SectorOf...	SectorOf...	SectorOf...	TakeOff...
EGKK-L...	6	1	11	5		8	2		11	1	27						
EGKK-L...	7	3	9	6	17	8			12		27	28			2		
EGKK-L...	8	3	9	4					15	1	27						563
EGKK-L...	9	3	9	4					18	1	27						566
EGKK-L...	10	0	13	5	17				21	1	27			2			
EGKK-L...	11	3	13	4					21	1	27						569
EGKK-L...	12	3	13	4					24	1	27						572
EGKK-L...	13	0	13	5					27	1	30			2			575
EGKK-L...	14	3	13	4					27	1	30						575
EGKK-L...	15	3	12	4					30	1	30						578
EGKK-L...	16	7	13	4					33	1	30						581
EGKK-L...	17	3	12	4					40	1	30						588
EGKK-L...	18	5	12	4					43	1	30						591
EGKK-L...	19	1	15	4					48	1	31						596
EGKK-L...	20	1	15	4					49	1	31						597
EGKK-L...	21	1	15	4					50	1	30						598
EGKK-L...	22	1	15	4					51	1	30						599
EGKK-L...	23	7	14	5					52	2	30	31					600
EGKK-L...	24	3	15	5					59		30	31			2		607
EGKK-L...	25	3	15	5					62		30	31			2		610
EGKK-L...	27	9	19	8	17				75	3	30	31		2	2	2	
EGKK-L...	28	9	19	8	17				84	3	30	31		2	2	2	
EGKK-L...	29	7	17	6					93	3	31	32				2	641
EGKK-L...	30	0	14	8	17				100	3	31		33	2	2		648
EGKK-L...	31	0	14	8	17				100	3	31		33	2	2		648
EGKK-L...	32	0	14	8	17				100	3	31		33	2	2		648
EGKK-L...	33	0	14	8	17				100	3	31		33	2	2		648
EGKK-L...	34	0	14	8	17				100	3	31		33	2	2		648
EGKK-L...	35	0	14	8	17				100	3	31		33	2	2		648
EGKK-L...	36	0	14	8	17				100	3	31		33	2	2		648

Summary of technical features considered by the AI method

Dynamic Query

Feature	Min	Query device	Max	Count
Flight ID				144
Step	6		36	31
Action	0		9	143
N conditions	7		19	144
N features	4		8	144
All conditions				30

□ Combine intervals ☑ Integer intervals

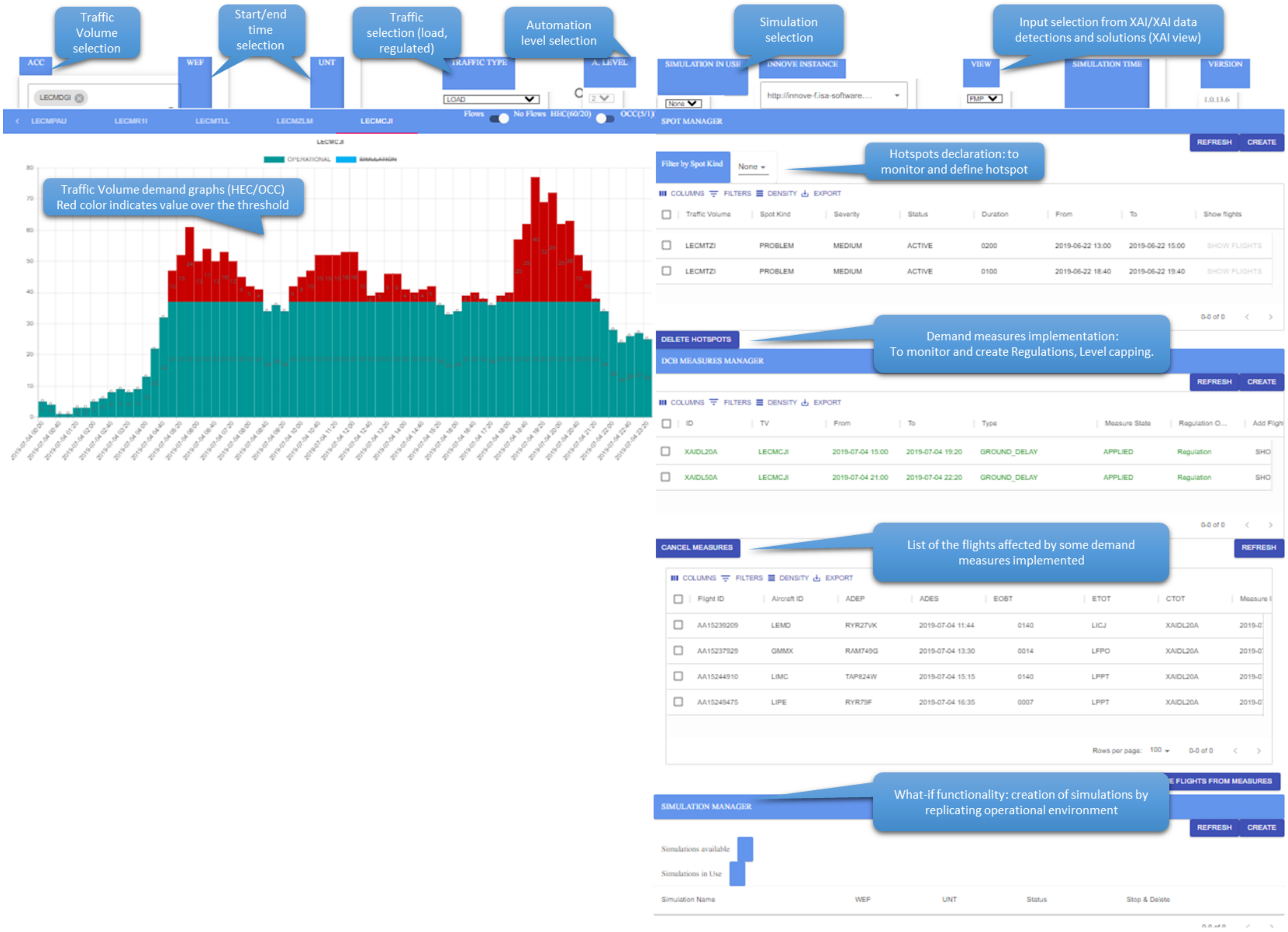
EGKK-LEPA-BAW2MZ-20190704072500 @ 260, action=9

Level	Feature	Sector	Value
0	PeriodOfHotspots01		30
1	SectorOfHotspots02	LECPDEX	2
2	SectorOfHotspots03	LECPDEX	2
3	NumberOfHotspots		3
4	PeriodOfHotspots01		30
5	NumberOfDelays		75
6	NumberOfDelays		75
7	NumberOfDelays		75
8	NumberOfDelays		75
9	NumberOfDelays		75
10	PeriodOfHotspots02		31
11	NumberOfDelays		75
12	SectorOfHotspots01	LECPDEX	2
13	PeriodOfHotspots01		30
14	DurationInSector01	LECBCCC	17
15	NumberOfDelays		75
16	PeriodOfHotspots02		31
17	NumberOfDelays		75
18	PeriodOfHotspots01		30

Explanations tree per flight

```
All (30)
  Action = 9 (2)
    PeriodOfHotspots01 in [22, 69] (2)
      SectorOfHotspots02 in [1, 10] (2)
        SectorOfHotspots03 in [0, 10] (2)
          NumberOfHotspots in [0, 4] (2)
            PeriodOfHotspots01 in [6, 38] (2)
              NumberOfDelays in [8, 100] (2)
                NumberOfDelays in [18, 100] (2)
                  NumberOfDelays in [25, 100] (2)
                    NumberOfDelays in [36, 100] (2)
                      NumberOfDelays in [46, 100] (2)
                        PeriodOfHotspots02 in [7, 31] (2)
                          NumberOfDelays in [0, 95] (2)
                            SectorOfHotspots01 in [0, 6] (2)
                              PeriodOfHotspots01 in [27, 69] (2)
                                DurationInSector01 in [9, 105] (2)
                                  NumberOfDelays in [0, 84] (2)
                                    PeriodOfHotspots02 in [29, 69] (2)
                                      NumberOfDelays in [52, 100] (2)
                                        PeriodOfHotspots01 in [6, 31] (2)
  Action = 7 (3)
  Action = 5 (1)
```

Figure 6. Explanations view for a particular flight for the ATFCM use case.



Traffic Volume selection

Start/end time selection

Traffic selection (load, regulated)

Automation level selection

Simulation selection

Input selection from XAI/XAI data detections and solutions (XAI view)

Traffic Volume demand graphs (HEC/OCC) Red color indicates value over the threshold

Hotspots declaration: to monitor and define hotspot

Demand measures implementation: To monitor and create Regulations, Level capping.

List of the flights affected by some demand measures implemented

What-if functionality: creation of simulations by replicating operational environment

□	Traffic Volume	Spot Kind	Severity	Status	Duration	From	To	Show flights
□	LECMTZI	PROBLEM	MEDIUM	ACTIVE	0200	2019-06-22 13:00	2019-06-22 15:00	SHOW FLIGHTS
□	LECMTZI	PROBLEM	MEDIUM	ACTIVE	0100	2019-06-22 18:40	2019-06-22 19:40	SHOW FLIGHTS

0-0 of 0

□	ID	TV	From	To	Type	Measure State	Regulation O...	Add Fligh
□	XAIDL20A	LECMCJI	2019-07-04 15:00	2019-07-04 19:20	GROUND_DELAY	APPLIED	Regulation	SHO
□	XAIDL50A	LECMCJI	2019-07-04 21:00	2019-07-04 22:20	GROUND_DELAY	APPLIED	Regulation	SHO

0-0 of 0

□	Flight ID	Aircraft ID	ADEP	ADES	EOBT	ETOT	CTOT	Measure I
□	AA15239209	LEMD	RYR27VK	2019-07-04 11:44	0140	LICJ	XAIDL20A	2019-0
□	AA15237929	GMMX	RAM749G	2019-07-04 13:30	0014	LFPO	XAIDL20A	2019-0
□	AA15244910	LIMC	TAP824W	2019-07-04 15:15	0140	LPPT	XAIDL20A	2019-0
□	AA15249475	LIPE	RYR79F	2019-07-04 16:35	0007	LPPT	XAIDL20A	2019-0

Rows per page: 100 0-0 of 0

E FLIGHTS FROM MEASURES

Simulation Name	WEF	UNT	Status	Stop & Delete

Figure 7. FMP Client Showing Potential Hotspots.

B. Conflict Detection & Resolution (CD&R)

The overall architecture of the CD&R integrated prototype developed is shown in *Figure 8*. This prototype comprises the operational ATC platform SACTA (the Spanish ATC platform developed by INDRA) [11] and the XAI system, together with the visualization and user interface (Vis&UI) component.

In particular, the ATC platform facilitates the user/ATCO to monitor the flights through a radar screen, where they can see the actual and planned trajectory of all the flights under their responsibility. The platform also provides updates on radar tracks and Flight Plans (FPLs) every 30 seconds to the XAI module to allow it to compute the solutions and provide them to the user through the Vis&UI interface.

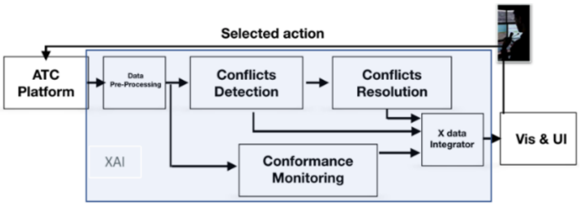


Figure 8. Overall CD&R Prototype System

The XAI module integrates three different components: conflict detection; conflict resolution; and conformance monitoring. The first component detects conflicts in the horizontal (separation infringements of less than 5NM) and vertical plane (separation infringement of less than 1000 ft), by projecting the trajectory into the future, 10 minutes ahead (***t_h***) (Figure 9). The module distinguishes whether the flight follows its FPL (a) or if it deviates from it (b). In the latter case, the trajectory projection is estimated according to the deviation of flight’s course.

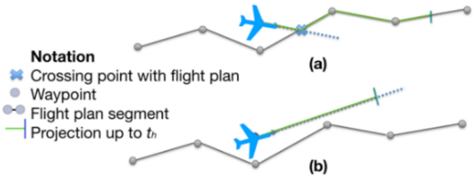


Figure 9. Projection of flight trajectory for the detection of conflicts.

The conflict resolution module, using an enhanced graph convolutional cooperative reinforcement learning method [12] [13] , decides on the actions (and their duration) to be taken for each flight to avoid a conflict. Those actions, decided in each timestep of the computation, every 30 seconds, include: change of altitude [±1 FL]; course change [±20, ±10, 0]; speed change [±7 knots;]; and go direct to a waypoint of the FPL.

Finally, the conformance monitoring component oversees if an aircraft follows the resolution actions as prescribed by the AI/ML module, by comparing the desired trajectory against the actual trajectory.

This AI/ML module is trained with real operational data from the entire 2019 year of the Spanish Barcelona ACC en-route sectors. The dataset comprises the FPL data, along with radar tracks of those flights and ATC events detected using ATON [14]. These ATC events include the clearances performed by the ATCOs to all the flights under their Area of Responsibility (AoR).

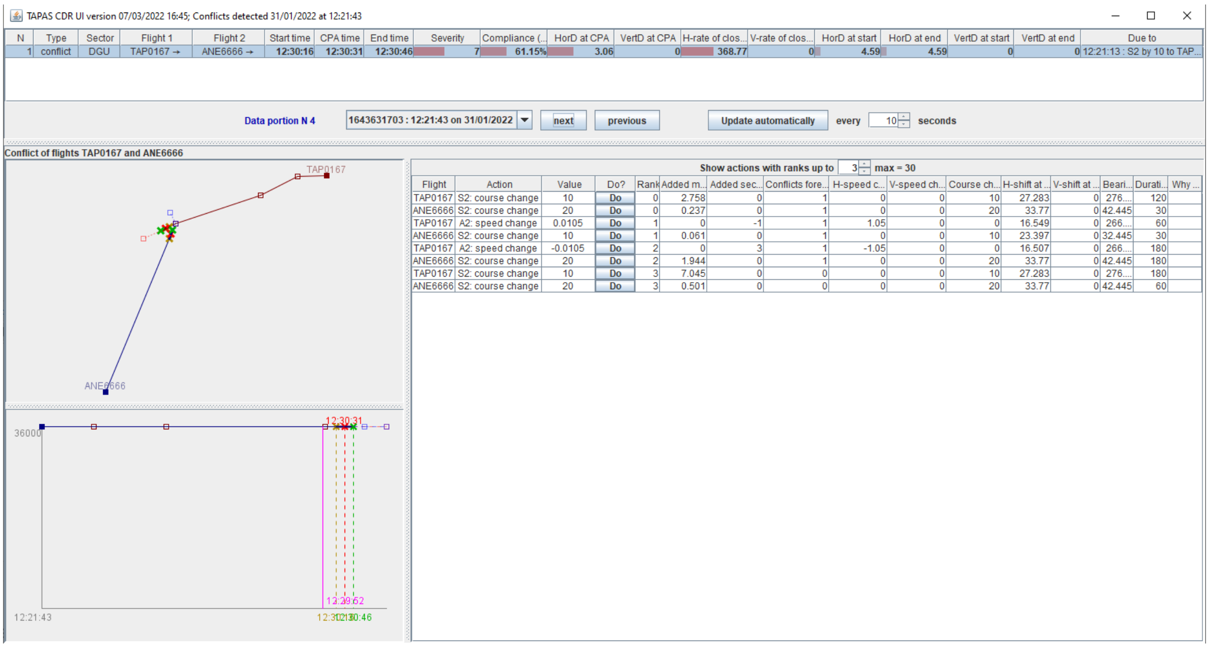
The CD&R prototype does not have any distinct explainability component for providing explanations of instructions decided by the system. Thus, a different explainability paradigm is followed from the one from ATFCM case. All relevant parameters that drive system’s decisions are provided, offering more transparency on decision making (i.e., making transparent the situations the system foresees), focusing on operational concerns, rather than on explainability of how decisions are taken from an AI/ML model.

Transparency is achieved through the use of a Vis&UI interface. The design of this interface respected the safety and time criticality aspects of the CD&R use case. Therefore, minimal information, but sufficient for understanding the problem and suggested solutions was presented; and a simpler and intuitive view was provided.

This translates into a representation of the conflicts and resolution actions in a tabular form as shown in *Figure 10*. The conflict is described in the upper part, including aircraft involved; aircraft altitude; separation minima violated; previous conflicts and resolution actions, if any, that provoked the new conflict; time at the start/end of the conflict and at the Closest Point of Approach (CPA); and a severity metric on the conflict. This severity metric, highlighted in red colored bars, is calculated as the sum of two scores: compliance measure (MoC, as the percentage of compliance with the separation minima required) and rate of closure of the flights (how the aircraft are getting close to each other). All this information is configurable, and the user can select to disclose or hide any of the columns. Additionally, the tool shows a 3D map of the conflict, separating the 2D view from the vertical one. This is a complementary view to the one provided by the ATC platform. Another view shows in a table list the potential ATC clearances for each flight involved in a conflict, ranking them according to their likelihood to solve the problem.

IV. Prototypes Validation

To explore and define the main principles and criteria for explainable and transparent AI/ML, several RTS were conducted. Those trials were performed using operational staff from the Spanish ANSP ENAIRE. The tests consisted of different runs conducted in the context of the three automation levels explained in Figure 1 and according to defined roadmaps. The idea behind these validation activities was to reproduce a realistic environment, in which to integrate the developed prototypes, prior to their verification tests, and extract the feedback of the involved operational users.


TAPAS CDR UI version 07/03/2022 16:45; Conflicts detected 31/01/2022 at 12:21:43
Data portion N 4
1643631703 : 12:21:43 on 31/01/2022
next
previous
Update automatically
every
10
seconds
Conflict of flights TAP0167 and ANE6666
TAP0167
ANE6666
36000
12:30:31
12:28:52
12:21:43
12:30:16
12:30:46
Show actions with ranks up to
3
max = 30

N	Type	Sector	Flight 1	Flight 2	Start time	CPA time	End time	Severity	Compliance (...	HorD at CPA	VertD at CPA	H-rate of clos...	V-rate of clos...	HorD at start	HorD at end	VertD at start	VertD at end	Due to
1	conflict	DGU	TAP0167 →	ANE6666 →	12:30:16	12:30:31	12:30:46	7	61.15%	3.06	0	368.77	0	4.59	4.59	0	0	12:21:13 : S2 by 10 to TAP...

Flight	Action	Value	Do?	Rank	Added m...	Added sec...	Conflicts fore...	H-speed c...	V-speed ch...	Course ch...	H-shift at ...	V-shift at ...	Beari...	Durati...	Why ...
TAP0167	S2: course change	10	Do	0	2.758	0	1	0	0	10	27.283	0	276....	120	
ANE6666	S2: course change	20	Do	0	0.237	0	1	0	0	20	33.77	0	42.445	30	
TAP0167	A2: speed change	0.0105	Do	1	0	-1	1	1.05	0	0	16.549	0	266....	60	
ANE6666	S2: course change	10	Do	1	0.061	0	1	0	0	10	23.397	0	32.445	30	
TAP0167	A2: speed change	-0.0105	Do	2	0	3	1	-1.05	0	0	16.507	0	266....	180	
ANE6666	S2: course change	20	Do	2	1.944	0	1	0	0	20	33.77	0	42.445	180	
TAP0167	S2: course change	10	Do	3	7.045	0	0	0	0	10	27.283	0	276....	180	
ANE6666	S2: course change	20	Do	3	0.501	0	0	0	0	20	33.77	0	42.445	60	

Figure 10. User interface of the VA component for the CD&R use case.

A. *Air Traffic Flow and Capacity Management (ATFCM)*

This first exercise focuses on the validation of ATFCM use case. It was assumed that tasks allocation was according to the roadmap described (*Figure 4*) and operators involved were fully familiar with the NM pre-tactical planning and DCB process.

Additionally, the dataset used for this validation activities was not included in neither the training nor in the testing datasets for the AI/ML component. Thus, the airspace of study focused on Madrid ACC and the traffic samples used were selected from the busiest days of 2019, where more imbalances between demand and capacity were found.

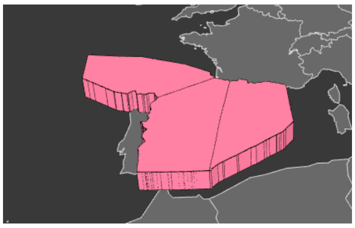


Figure 11. ATFCM simulated region.

During a three-day exercise, the operational users participated in different runs with the three automation levels. First day was focused on training with the new tools, and the rest of the days were oriented to validate the tools and extract inputs from them. To collect the feedback from the operational users, the validation team used over the shoulder observations, along with questionnaires on situational awareness, automation and workload, and debriefing sessions.

From the ATFCM experiments, it was concluded that both, FMP client and XAI/VA tool, were useful for the task at place. The users confirmed that in general they were able to access to the information in an easy way and that all the relevant information (monitoring data, solutions, explanations) and functionalities (what-if, creation of measures) that they would need were included in the tools.

With regards to the VA tool, the participants considered it helped them to maintain the situational awareness, although sometimes the access to the explanations were somehow obscure. Moreover, they also declared that the tool provided too much information (e.g.: how the algorithm arrives to the solution in each timestep of its calculation) that in the operation timeframe was not needed and that they would not consult, but in the certification and training phase such detail would be probably useful. The explanations provided by the tool were also considered appropriate. However, they also recommended to have the solutions' impact from a sector perspective rather than a flight perspective, as well as to include some statistical metrics on that impact (e.g.: delays saved, number of flights impacted, etc.). This was also related to the fact that the paradigm of solving hotspots was tackled by the AI/ML component in a different way than today's operating method, from a sector perspective to a global approach. This was especially relevant in automation level 2, where the human must decide to implement or not the solutions proposed by the system. Since those solutions involved different outcomes, a solution could, for instance, try to solve more than one hotspot at a time in different sectors. Then the operator was unable to partially implement a solution and would need to implement all solutions or none. Here the operator also declared the importance of avoiding any biases in the computation of the solutions (e.g.: the AI/ML method should not prioritize any company over another, any country over another). In any case, this different approach was not rejected by the users, as they considered it could be also applicable as long as that fairness or the 'game rules' are guaranteed.

The operators consulted agreed on the feasibility of automation level 3, with no need of too many explanations from the tool. Instead, they expressed the need for the tool to be effective and, thus, trust is built from the continued use of it. Especially since this use case takes place in the pre-tactical phase and the FMP are nowadays used to EUROCONTROL's CASA algorithm, which allocates automatically the delays implied by the regulations. This algorithm is a 'black box' for the operators, but they trust it as it solves the imbalances. During trials for automation level 1, which reflects todays' operating method, operators did not have any CASA algorithm. Therefore, regulations should be calculated manually by the user, leading to a lower automation level.

All these conclusions and lessons learnt from this use case were very valuable and served as an input for the next CD&R use case, which was conducted several months later.

B. Conflict Detection & Resolution (CD&R)

For the CD&R also a three-day RTS with actual ATCOs was conducted. The trials focused on two en-route sectors (above FL345) of Madrid ACC: Domingo upper (LECMDGU) and Toledo upper (LECMTLU) sectors. Each run focused on one sector at a time as the prototype was developed in that way, that is, to detect conflicts inside the active sector and the borders of the downstream sector.

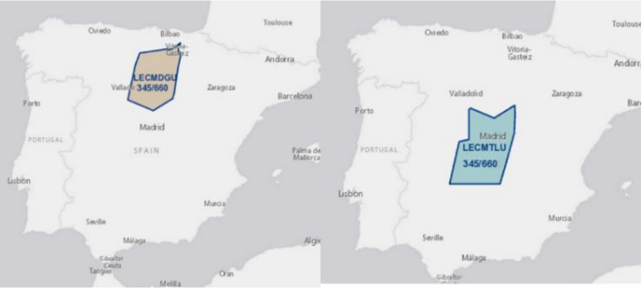


Figure 12. CD&R Airspace considered for the RTS.

The traffic samples used were historical data from 2019 (the 25th and 30th of June, and the 4th of July 2019). This traffic samples were loaded into the SACTA platform, and its traffic simulator provided both FPL and radar track to the radar screen and VA prototype. At first, the VA tool worked in a secondary screen, but later it was integrated into the radar screen. This was highlighted as very useful by the experts involved, since this CD&R use case is safety critical, thus, deviating the line of sight from the radar screen is not preferable. The trials also involved a pseudo-pilot emulating the pilots aboard the aircraft crossing the ATCO AoR and their communications. It also should be noted that the trials were performed with no planner controller aiding and supporting the executive role.

For this specific use case, the VA tool provided information focusing on the transparency rather than explanations, since this was a safety critical experiment, where explanations of the proposed solutions and detected conflicts were self-explanatory and there would not be as much time to consult them.

The participants were Spanish ATCOs, familiar with the airspace of the study. They quickly assimilated the information provided by the tool regards the conflict detection and the analysis of the proposed solutions. Some features of the display were considered less useful than others. For example, users indicated that the graphical display of the conflict trajectories in the VA display did not provide useful information beyond what was already available in the radar screen. However, other information that was provided related to conflict alerts and the proposed actions, was considered very useful and allowed the users to quickly understand the conflict and traffic involved, as well as the solutions that were proposed, even if they did not always agree with the priorities given by the AI/ML method.

Users responded that the system was easy to use and understand with little or no assistance from technical support personnel. They also indicated that little or no additional training was needed. Additionally, the VA component allowed certain degree of configuration, as the user could hide some of the fields shown. They stated that feature was very useful and that probably they would not need all the fields present, just mainly the flight ID of the aircraft involved in the conflict, time to conflict, separation minima at the CPA, the best solution to solve the conflict and the point and/or alert where/when the flight after the resolution action should resume to its FPL.

Regarding the solutions, even though the tool provides a ranking, the users stated that by default they would prefer a single solution presented, but also have the option to search for more. As for the solutions themselves, they should consider the performance of the aircraft and other basic rules (e.g.: ceiling of the aircraft, descend a flight that plans to descend).

In cases where the users disagreed with the clearances being proposed, or automatically implemented at level 3, they still tended to accept the solutions if they solved the conflicts, even though the solutions differed from those that they would have applied themselves. But most users indicated that due to the often very short lead times for conflicts to be identified and solved, including the instructions given to traffic, offering more information than was already provided by VA tool would not necessarily have changed the understanding that they could usually acquire due to their own experience and expertise in the domain. Therefore, users agreed that the level of information provided was sufficient for their needs in the CD&R use case.

The users also stated the importance of the visual aids used in this use case. The VA tool highlighted through red colored bars those conflicts with more severity. However, the view was simplified, and an extensive use of colors to catch the ATCO attention would be desirable.

The participants also considered that automation levels 1 and 2 were feasible in terms of operational applicability as they were implemented in the trials. Nonetheless, there were mixed outputs regarding automation level 3. In general, users declared that relegating the human to a monitoring role would imply that over time the ATCO would lose its expertise and capabilities to take over, if necessary. In those scenarios it would be extremely important to have a degrading mode, where the system alerts on its malfunctioning, and adequate and sufficient training has to be completed to allow the human to regain control.

Regards to the later issue, tests including the malfunctioning of the prototype in automation level 3 were performed. In all these runs, the ATCO was able to detect that malfunction and take over to solve the remaining conflicts. However, this was expected since the participants were well trained ATCOs.

V. CONCLUSIONS

All these experiments conclude in the following findings, insights and remarks on explainability and transparency. Further research is also advised especially regarding Human Performance assessment with use of extensive indicators apart from the ones used in the simulations through questionnaires, debriefing sessions and over the shoulder observations.

First, rather than having explanations, the user needs to trust the system. This was especially true for the CD&R use case, where aspects such as robustness and safety are more critical than transparency & explainability. Through the constant use of the system, especially during the training phase, the human can develop trust in the system through how it performs and the solutions it is 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 explanations provided by the support tools to the users.

Confidence and trust can be volatile. Developing trust and confidence in an AI/ML system takes a long time and depends on the system providing reliable solutions that the user accepts as a valid response to a problem. When something subsequently fails badly, even after trust has been achieved, that confidence in the system can be lost rapidly and rebuilding it can be hard. This is especially critical in the CD&R domain, and therefore close attention must be paid to the reliability and suitability of the proposed solutions. Additionally, disruptive solutions, solutions leading to more complex issues later on, and the lack of a complete resolution process (e.g., resumption of flight plan) may contribute to a reduction in trust and confidence.

Varied levels of explainability are necessary according to the time horizon considered. During the operation, the users do not need to see all information or explanations related to the proposed solutions by the AI/ML system (it may require a time they do not have). In contrast, during training, explanations and solutions provided by the AI/ML are needed and appreciated by the user, but once the approach being used was understood, users did not really interrogate this information further.

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. In particular, in ATFCM scenarios 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.

Complexity of the solutions limits the capacity of the human to understand the explanations in real time. Although these are provided, in cases where the solution is too complex the human will have neither the time nor the ability to understand them. However, more than having explanations, the user wants to see the impact of the solutions proposed by the system.

For the safety critical use case of CD&R, the acceptance of automation level 3 requires further research. In the trials, human experts discussed extensively that performing a monitoring task may result in ATCO loss of expertise in the controlling tasks and whenever the AI/ML system 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.

In CD&R scenarios the importance lies on providing solutions that work well and are accurate, rather than focusing on explanations. Users 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 prioritization of choices is sufficient to allow them to rapidly understand the proposals and the consequences of those actions.

ACKNOWLEDGMENT

This work has received funding from SESAR Joint Undertaking (JU) within TAPAS project under grant agreement No 892358. The JU receives support from the European Union’s Horizon 2020 research and innovation programme and the SESAR JU members other than the Union. Also, we thank ENAIRE’s operational staff for their support in the use cases validation.

REFERENCES

[1] TAPAS website – Towards and Automated and exPlainable ATM System

[2] SESAR Joint Undertaking, “EUROPEAN ATM MASTER PLAN. Digitalising Europe’s Aviation Infraestructure. Executive view”, Edition 2020.

[3] European Commission, High Level Expert Group on AI, “Ethics Guidelines for Trustworthy AI”, April 2019.

[4] George Vouros, “Explainable Deep Reinforcement Learning: State of the Art and Challenges”, ACM Computing Surveys.

[5] Innovative Network Operations Validation Environment (INNOVE). EUROCONTROL

[6] T. Kravaris, et. al, “Explaining Deep Reinforcement Learning Decisions in Complex Multiagent Settings: Towards Enabling Automation in Air Traffic Flow Management”. Applied Intelligence, 2022.

[7] R. Barto and A. Sutton, “Reinforcement learning: An introduction”, MIT Press, 2019.

[8] V. Mnih and e. al., “Human-level control through deep reinforcement learning,” Nature 518(7540), 2015.

[9] Demand Data Repository (DDR). EUROCONTROL

[10] H. Gouk, B. Pfahringer and E. Frank, “Stochastic Gradient Trees,” ACML 2019, also in arXiv:1901.07777, 2019.

[11] ENAIRE, INDRA, “SACTA III. SACTA Architectural Model”, April 2006

[12] J. Jiang, C. Dun and L. Zongqing , “Graph Convolutional Reinforcement Learning for Multi-Agent Cooperation,” ArXiv, 2018.

[13] George Vouros et. al., “Automating the Resolution of Flight Conflicts: Deep Reinforcement Learning in service of Air Traffic Controllers”, PAIS IJCAI, 2022.

[14] ATON - Automatización TOmas Norvase),CRIDA A.I.E website.