Proceedings of the 2021 AIAA/IEEE 40th Digital Avionics Systems Conference (DASC), 3-7 October, San Antonio, USA
DOI:10.1109/DASC52595.2021.9594341

An Explainable Artificial Intelligence (xAI) Framework for Improving Trust in Automated ATM Tools

*

Carolina Sanchez Hernandez
School of Aerospace, Transport and Manufacturing
Cranfield University
Cranfield, U.K.
c.sanchez-hernandez@cranfield.ac.uk

Samuel Ayo
School of Aerospace, Transport and Manufacturing
Cranfield University
Cranfield, U.K.
samuel.s.ayo@cranfield.ac.uk

Dimitrios Panagiotakopoulos
School of Aerospace, Transport and Manufacturing
Cranfield University
Cranfield, U.K.
d.panagiotakopoulos@cranfield.ac.uk

***Abstract*—With the increased use of intelligent Decision Support Tools in Air Traffic Management (ATM) and inclusion of non-traditional entities, regulators and end users need assurance that new technologies such as Artificial Intelligence (AI) and Machine Learning (ML) are trustworthy and safe. Although there is a wide amount of research on the technologies themselves, there seem to be a gap between research projects and practical implementation due to different regulatory and practical challenges including the need for transparency and explainability of solutions. In order to help address these challenges, a novel framework to enable trust on AI-based automated solutions is presented based on current guidelines and end user feedback. Finally, recommendations are provided to bridge the gap between research and implementation of AI and ML-based solutions using our framework as a mechanism to aid advances of AI technology within ATM.**

Keywords— Air Traffic Management, Artificial Intelligence, Machine Learning, Trust Framework

I. INTRODUCTION

The aviation industry is currently facing challenges such as the need for improved profitability, fuel efficiency, environmental sustainability, airspace utilisation and safety. In addition, there is the urgent requirement to enable digitalization and automation to support the seamless and safe integration of new entrants such as Uncrewed Aerial Systems (UAS) and operations such as Urban Air Mobility (UAM).

These challenges are also set in context of the overall transition towards Performance-Based Regulations (PBR) and Performance-Based Operations (PBOs) [1] [2], which aim to enable integration of more automated, autonomous and intelligent systems into an industry that has historically been based on deterministic systems with known risks. PBR is being introduced to allow regulatory approval of novel solutions (that may include Artificial Intelligence (AI) and Machine Learning Techniques (MLT) that may have unknown hazards and unproven controls, by measuring performance with and without the use of traditional standards [3]. Within this context, clear guidelines on acceptable levels of performance and trustworthiness for increasingly automated and autonomous systems that use AI or ML techniques, particularly in safety-critical operations, are extremely important.

In order to make AI and ML solutions implementation successful, we have to focus on trust assurance frameworks encompassing many different elements that range from technical robustness to transparency to security and safety [41]. A recent review of existing literature on Machine Learning Techniques (MLT) in aviation indicates clear advances in this area of research in various Air Traffic Management (ATM) applications. However, due to recent developments on guidelines needed to ensure the trustworthiness of an AI solution such as those of EASA [4], more consideration to these upcoming regulations is needed in research projects henceforth.

This paper explores all the above as part of Fly2Plan project. Fly2Plan is part of the Innovate UK's Future Flight Challenge, a research and development programme aiming to encourage development of new and sustainable modes of air travel and to support the airspace and aviation systems of the future. The challenge brings together established leaders in aviation, academics and SMEs to research Air Traffic Management (ATM) and Uncrewed Traffic Management (UTM) integration as well as advance automation innovation.

A component of this project is a novel framework for Trust Assurance of automated solutions in ATM powered by AI. The framework is developed taking into account a variety of guidelines from regulators and feedback from industry experts through an AI Trust Assurance survey that was conducted as part of the Fly2Plan project. It aims to bridge the gap between AI research projects and the assurance needed for the implementation of these AI solutions within ATM.

II. AI AND ML IN AVIATION AND AIR TRAFFIC MANAGEMENT

Over the years, the aviation ecosystem has evolved into an environment of trust underpinned by stakeholders communicating mainly by voice. Trust in based on the knowledge that competent and certified agents are on the

other side of that link and utilizing mostly deterministic automated decision support tools. With the increased use of intelligent automation, tending towards full-autonomy and inclusion of non-traditional entities such as new UAS and UAS Traffic Management (UTM) service providers, stakeholders need assurance that collaborating entities in the airspace are trustworthy. Therefore, there is a need to develop robust solutions to manage, measure and assure trust between humans and automated entities. The solutions also need to fulfil the same safety levels as currently experienced in aviation.

In this context, the European Union Aviation Safety Agency (EASA) is working very actively towards the understanding of challenges and creation of guidelines to facilitate the safe implementation of AI solutions in the near future. In order to start preparing for all the changes in the Aviation ecosystem, EASA [4] has released guidelines for Trust Assurance to orient choices in the development of AI and ML solutions. This does not however constitute either definitive or detailed means of compliance. These guidelines apply to any system developed using AI and ML techniques and are intended for use in safety-related applications.

At the same time, interest in research of the applications of AI and ML in ATM has soared in recent years, as stakeholders realise the potential of leveraging the data they collect to optimise their processes [5] [6].

Conducting a search in Google Scholar and Scopus with the keywords “Machine Learning” and “Aviation” showed an increasing number of publications related to these topics in the last five years. Papers that were not specifically related to ATM were discarded (i.e. engine fault detection, aircraft maintenance, passenger demand). Although this does not represent an exhaustive search in terms of detailed algorithms or areas of application, the general trend shows that significant body of work is added every year to literature, averaging over 200 publications in the last two years only.

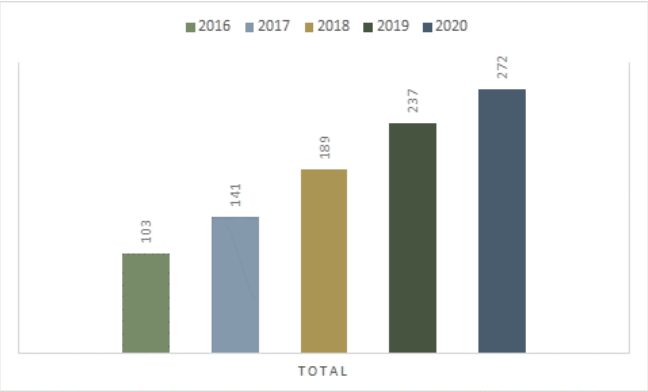

Figure 1 Google Scholar and Scopus publications for MLT in Aviation

Some of recent ATM areas where ML algorithms have been the focused of research are represented in the table below:

Machine Learning Algorithms	ATM Applications	Example of References
Support vector machines	Data mining, Forensics of a flight Aviation accident research, Accident Risk ,Conflict detection ,Hazard identification	[7],[8][9]
Random forests, X-boosting, LightGBM	Departure Delay, Arrival Delay, Fuel consumption estimation, Trajectory prediction, Scheduling, Feature importance	[10][11][12][13]
Bayesian regression	Delay propagation, Fuel efficiency, Predicting airport acceptance rates	[14] [15]
Clustering	Cluster aircraft trajectories before landing, insights on aircraft approach phase, Anomaly detection on flight events	[16][17][18][26]
Ensembles and hybrid models (parallel or sequential combinations of different algorithms)	They are more suitable for real-time data and non-linear problems, Risk prediction in Aviation safety	[19][20][21][22]
Reinforcement learning	Process control, Operational decision making, Traffic optimization, Demand capacity balance	[23][24][25]
Common neural networks algorithms	**ATM Applications**	**Example of References**
Long short-term memory (LSTM)	Forecasting traffic flow Forecasting aircraft trajectory, Fuel consumption, Aircraft landing speed prediction, Visibility prediction at airports, Delay forecasting	[27][28][22]
Convolutional neural network (CNN)	Anomaly detection, CyberSecurity, Audio processing, Runway occupancy prediction	[29][30][31]
Feed forwards neural networks (FFNN)	Predicting and forecasting, Fuel estimation, Trajectory prediction, Flight-delay prediction	[32][33][34]
Extreme learning machine (ELM)	Anomaly detection, Boarding time prediction, Multi-aircraft conflict resolution, Forecasting weather and visibility	[35][32]

Table 1 Review of recent ML applications in ATM

Despite the abundance of research in recent years, our review revealed that there is a gap between the increase on research papers and the actual implementation of these solutions in an operational environment. This is due to several challenges.

The first challenge is that aviation is a very conservative field with a focus on deterministic systems that continuously builds on known risks and often on lessons learnt from recorded incidents, near-misses and always from accidents.

This means that currently, available aviation design and development assurance methods using safety assurance standards are not fully suited for regulatory assurance of autonomous systems or AI or ML applications. Furthermore, the potential growth UAS and Urban Air Mobility (UAM) technology is prompting new questions about the level of automation of existing systems to be able to incorporate ATM and UTM operations and flexibility to integrate new entrants and new types of data. The challenge for applicants and regulators is to assure system safety and trustworthiness of new solutions using AI [36]. Current work by regulators and a roadmap for future standards for AI solutions are being addressed by working groups such as SAE/EUROCAE WG114 [37] which objectives are to establish industrial best practices for the development and the certification of AI embedded into aerial vehicles and ground equipment, providing standards for Qualification of Aeronautical Systems embedding AI.

Generalisation is another challenge that might impede the implementation of AI and ML research solutions within ATM. For example, many of the research papers that we reviewed focused on specific operational scenarios and datasets such as using a limited set of routes, time of year, aircraft types and meteorological conditions. This means that the predicted results will only be acceptable for scenarios for which the timeseries, aircrafts or route characteristics are similar to the training data making their generalisation difficult. There can be very valid reasons for this, such as specific weather patterns, location and structure of airports runways, flight and trajectory mix complexity. However, this makes the wider applicability and scalability of the methods and algorithms described extremely challenging and thus so as their full deployment in order to advance automation within ATM.

There are papers that recognise this challenge and have focused on areas where a methodology can be developed and potentially generalised and scaled-up. For example, [38] evaluated feature identification for flight risk and they proposed a step-by-step methodology to down-select a reduced set of parameters that can be used for safety analysis. As millions of flights are flown every year and as the size of the data grows, knowing which parameters analysts need to pay attention to is becoming very critical. They de-identified Flight Operational Quality Assurance (FOQA) data obtained from commercial airline routine operations and used clustering algorithms in order to identify common characteristics of safety events during the approach phase. After different clustering algorithms were applied and evaluated, the one-way analysis of variance (ANOVA) was performed on selected clustering results to identify parameter significance. Their methodology helps focus similar research in the parameters that are of importance, aiding the computational efficiency and facilitating the generalisation and comparison of results in different applications.

In [11] the authors proposed a framework leveraging flight data from the approach phase between certain approach altitudes in order to train an offline model to predict the landing true airspeed and ground speed using a Random Forest regression algorithm. This model developed offline was then used to predict landing performance metrics online. They used data obtained from commercial airline operations that contained thousands of flight records from the FOQA datasets. They concluded that their model had the ability to predict the true airspeed and ground speed at 300 feet above touchdown to within a few knots providing the basis for decision making by the Air Traffic Controller to decide about stability in a potentially hazardous situation, regardless of location specific data.

Finally, another important challenge is interpreting AI and ML algorithms, especially those that use deep learning techniques, as this has become one of the main obstacles with regards to their practical implementation. The inability to explain or to fully understand the reasons why AI and ML algorithms perform as they do is a real issue for trust assurance in aviation, as such high safety environment requires full traceability between system outputs and their input parameters. According to the Dependable and Explainable Learning project (DEEL) [39], there is a wide consensus in the AI scientific and industrial community on the need to have the capability to explain the behaviour of a model produced by these technologies in order to be certified and implemented in safety critical operational systems. Recent research in explainability such as explainable AI (xAI) techniques, user-centric explanations and auditability of algorithms are trying to address this gap [40] and increase trust in the solutions.

This paper focuses on two of the above described challenges: addressing trustworthiness and assurance.

III. Meeting the challenges of AI implementation

In order to accelerate the validation and implementation of all the promising research that is being developed, there are a number of initiatives that aim to improve trustworthiness and drive the development of AI-based solutions. According to the EU High Level Group of Experts on AI (AI HLGE) [41] there are three components which should be met throughout the system's entire life cycle to create a trustworthy AI solution: it should be lawful, ethical and robust.

The above translate into a series of areas that should be considered when building trust in an AI solution. Not all of them might be considered in all cases, depending on the area of development and the solution itself. These areas include: (i) Technical Robustness and Safety (ii) Privacy and Data Governance (iii) Transparency (iv) Diversity, non-discrimination and fairness (v) Societal and Environmental wellbeing (vi) Accountability (vii) Human Agency and Oversight [41].

Furthermore, the EU AI Act proposal [50] lists a series of more concrete points that would need to be fulfilled for high-risk AI implementations (which includes critical infrastructure such as aviation):

1. Using high-quality training, validation, and testing data
2. Using documentation and design logging feature that ensure continuous documentation

3. Ensuring transparency and informing the user about the application of AI systems
4. Ensuring human oversight throughout the process
5. Ensuring accuracy, robustness, and cybersecurity of the system

Finally, the European Union Aviation Safety Agency (EASA) strategy which is in the Policy and Regulation side of ATM, embraces this approach from an aviation perspective and has committed to participate in the testing and improvement of these guidelines [4]. For that it has developed a high-level AI Trust Framework for Aviation that takes into account the AI HLEG guidelines and translated them into three main blocks:

- Learning Assurance
- Safety Risk Mitigation
- Explainability

In order to attain our goal and bridge the gap between research projects and future implementation, we have taken into account all the above guidelines as a baseline, so we are sure that we future proof our solution by incorporating the fundamental requirements of potential future regulatory and safety standards. In addition to this, we considered necessary to emphasize the importance of experts' views within the field of ATM and assess specific areas that are most relevant to an aviation safety driven environment. In order to do this, we carried out an AI Trust Survey as part of the Fly2Plan project. The results of this survey will be discussed in detail in the following section.

IV. AI Trust Assurance Survey

The aim of our AI Trust Assurance survey was to engage with stakeholders and experts within the Fly2Plan project both in ATM and UTM operational environments, in order to assess the level of trust in solutions and automation provided by Artificial Intelligence within decision support tools. The questions were based on:

- The taxonomy developed by SESAR on Levels Of Automation (LOAT) [42] to assess what level of automation the stakeholders were comfortable with and whether having a solution provided by an AI-based system would make any difference in the level of trust they have in it
- Based on EASA [4] and the AI HLEG guidelines [41] for trust assurance, we formulated questions to assess what elements involved in the design, development and implementation of a solution were more important to the potential end user
- Based on literature review, we explored what components of explainability and communication with an AI were more relevant for the end user
- Finally, we assessed if the area of operation (planning, pre-tactical or tactical) had any influence in the level of trust on an AI solution

The results are based on responses from a total of 34 experts. The spread of expertise was quite wide including air traffic controllers, airport operations specialists, air traffic management consultants and software developers as the main groups.

Key findings included:

- Trust tends to decrease as the level of automation increases
- Trust is lower when the automation is provided by an AI in areas such as action selection and action execution as opposed to information acquisition and information analysis where trust seems to be higher.
- Safety and security are the most important elements of an AI solution in ATM, followed by accuracy and reliability
- Within the area of safety and increasing trust in the solution, alerts and safety nets are of upmost importance
- False positives or negatives have a high influence on lower trust in the solution. Knowing the reason why they happen would help increase trust but operationally might still be an issue to deal with, so it is preferable not to have them
- Loss of trust is difficult to solve and requires time to rebuild, providing explicit evidence that the issues that caused the loss of trust have been address is very important in order to start regaining trust
- Explanations are preferable based on main factors that influence the solution, specific examples and visuals
- Explanations are important around why the solution performs well or badly and understand when (in what situations or point in time) it is likely to perform badly
- There is a willingness to work with the AI in an "human augmentation" manner, and learning and complementing each other (AI-Human cooperation)
- The area of operation where the solution might be implemented (planning, pre-tactical, tactical) does not influence the trust in the solution.

For example, the SESAR LOAT taxonomy [42] is grouped by the four cognitive functions:

- Information Acquisition
- Information Analysis
- Decision and Action Selection
- Action Implementation

In the case of the area of action implementation which implies that the AI could execute an action, experts were asked to choose the level of automation they felt more comfortable with. The results based on the automation classification by SESAR LOAT [42] were:

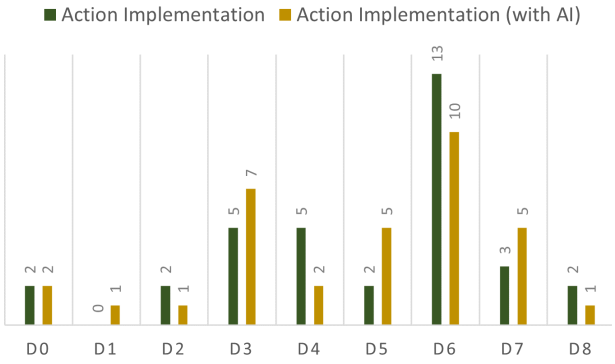

Figure 2 Action implementation levels of trust with AI and without AI (with D0 being manual solutions and D8 being fully automated solutions)

Option D6 was the level of automation most voted by over a third of respondents and depicts a medium level action sequence automation (the system initiates and executes automatically a sequence of actions and a human can monitor all the sequence and can interrupt it during its execution). This highlighted that human oversight is still key for trust assurance in automation, independently of the use of AI (and concurs with the suggested guidelines by the EU AI Act proposal, AIHLG and EASA mentioned earlier). This is to be expected as levels of automation are still low in ATM and thus we would still expect humans to be overseeing the increasing levels of automation as confidence is gained of the correct and safe implementation and operation of more intelligent and AI-based systems. This is particularly evident in safety-critical decision making, such as executive decisions to be made by Air Traffic Control (ATC) and pilots, that has direct and immediate impact on safety.

We also asked the participants to rank, in order of importance (1 being the most important), what would be the main elements to have trust in an automated solution provided by AI/ML. The ranking that resulted from the votes is as follows:

Ranking	Options
1	The safety and security of the solution
2	The accuracy of the solution
3	The reliability of the solution
4	The resilience of the solution
5	The reproducibility of the solution
6	The auditability of the solution
7	The fairness of the solution

Table 2 Ranking of elements of trust on AI in ATM

Safety and security of the solution got 58% of the votes as first in importance, followed by accuracy (26.5% votes as first choice) and reliability (8% as first choice) which once more reflects the upmost importance of safety and security within the aviation industry. Furthermore, as part of national critical infrastructure, ATM and UTM solutions will be classified as High Risk for AI implementations by the new EU AI Act and as such a high level of auditing and compliance will be needed to assure safety.

When asked specifically about safety and what knowledge would help increase trust in an AI automated solution, the answers in order of importance were as follows:

Ranking	**Options**
1	Knowing that there is an alert and safety net when there is an anomaly in the data
2	Knowing that there is an alert and safety net when there is a change in performance
3	Knowing that the data quality is monitored
4	Knowing that there is a human in the loop
5	Knowing that the performance of the algorithm is monitored

Table 3 Ranking of safety measures and trust

Safety nets came on top of the reassurance on the solution when there is an anomaly or change or performance with 47% of respondents voting for it as first in the ranking.

In terms of explainability methods in order to increase trust in the solution, the resultant raking in order of importance was as follows:

Ranking	Options
1	Explanation of the main factors influencing the algorithms decisions
2	Explanations through specific examples to understand the reasons for an algorithm decision
3	Explanation through visuals that represent the functioning of the algorithm and the solution
4	Causal explanations. What can be changed about an AI/ML algorithm or its input that results in an impactful change in the output
5	Explanation of the overall life cycle and design of the AI/ML solution
6	Counterfactual explanations. Why the answer is A instead of B

Table 4 Ranking of explainability techniques

Explanations are fundamental in order to gain the trust of the end user, especially those that can demonstrate factors that influence the final decisions, and this will resonate with end users mental processes. The level of detail and the way the explanation is delivered has been the subject of research for years [52] but it has become even more relevant as AI and ML lack of traditional transparency become a cause of distrust.

Finally, in terms of area of operation and automation, we asked the level of trust on an AI solutions in planning, pre-tactical and tactical operations. Tactical operations are the ones that pertain more risk and therefore have more stringent safety assurance requirements. In this case, the "somewhat likely" and "likely" options were voted by the majority of respondents but their accompanied comments indicated that this trust would be conditioned by trust assurance methods that should include explainability, extensive tests, validation, guaranteed technical robustness

and safety performance. Nevertheless, this demonstrated a positive attitude towards automation and AI solutions within this high safety risk operational environment.

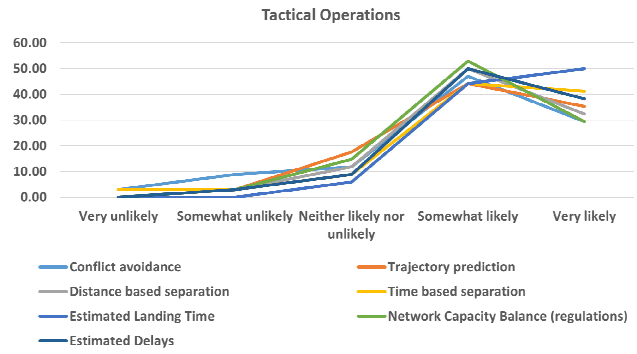

Figure 3 Trust on an AI solutions in tactical operations

V. ATM XAI TRUST FRAMEWORK

Based on all our research indicated above, we designed an xAI Trust Assurance framework that encompasses all the necessary elements in order to bridge the gap between ATM AI research projects and capture the basic elements for future regulatory approval and operational deployment. Our proposed building blocks are:

1. Purpose of the AI Solution
2. Technical Robustness and Learning Assurance
3. Safety and Security Assurance
4. Transparency and Explainability

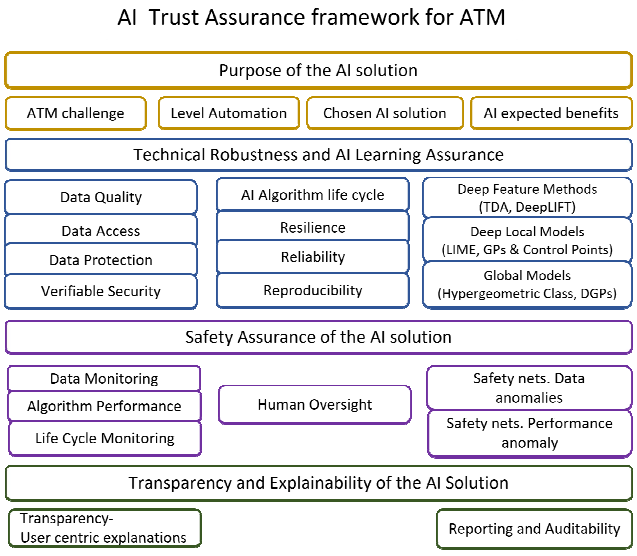

Figure 4 AI Trust Assurance Framework for ATM solutions with the different blocks and sub blocks of the framework. Colours depict different areas of focus.

A. *Purpose of the AI Solution*

In an industry with high safety standards, it is extremely important to understand what challenge the solution is trying to solve and the reason why a particular solution was chosen instead of another (i.e. why an ML algorithm was chosen to solve this issue). Many research papers focus on a comparison with current practices to highlight the advantages of their proposed solution using ML [48] [27] [32] and others concentrate on the novelty of the algorithm itself [46] [47]. The results and advantages of these algorithms are mainly focus on a specific performance metric (i.e. accuracy, RMSE) [55] [56]. Although the challenge addressed is normally mentioned in the introduction of many papers, such as improving operational efficiency, experts and end users feedback emphasize the need for a more thorough description of the purpose and advantages of these algorithms within current practices.

Secondly, our AI Trust Assurance survey highlighted that the level of automation the solution will operate in, is also very important. Trust seems to decline greatly the more the solution approaches full automation. As mentioned earlier, this is to be expected as there are low automation levels currently in ATM and thus need to have a gradual change and evolution to higher automation and more intelligent systems Researchers need to consider that the higher the level of automation they propose, the lower the trust of the end user might be and the need for more trust assurance mechanisms such as evidence of traceability of inputs and outputs. For example [49] suggested some uses of their research could be providing a preferred runway assignment in a multiple runway configuration when aircraft enter the terminal airspace, which will make the solution a decision and action selection automation level of SESAR LOAT. This is an important piece of information for an end user and implementation perspective.

Finally, from an end user point of view, the expected benefits of the solution need to be clear from the start whether these are operational, financial or environmental. In the SESAR ATM Masterplan [72] some of the targets for improvements by 2050 are for example, up to 30% reduction in departure delays, up to 10% more flights landing in congested airports or up to 10% reduction in fuel burn. The adoption of more AI solutions would benefit greatly if research projects could relate their results to achieving performance targets.

B. *Technical Robustness and Learning Assurance*

These two equally important elements of a solution are a combination of guidelines from HLEG [41], EU AI Act proposal [50] and EASA [4]. Based on these and the results of our survey, the elements that we consider most relevant for our Trust Assurance framework are:

- Data sources (quality, access, integrity, protection and security) (Learning Assurance/Technical Robustness)
- Algorithm Life cycle (Learning Assurance)
- Resilience, Reliability and Reproducibility (Technical Robustness)
- Algorithm Interpretability/Explainability (Learning Assurance/Technical Robustness)

Our AI Trust Assurance survey verified that both technical robustness and learning assurance are extremely important factors for trust and for future implementation.
Example of papers that we found that explored technical robustness of an AI algorithm in an ATM operational environment were [11] and [12]. The authors used Gradient Boosting Decision Trees to predict if an aircraft will miss its optimum runway exit (called its procedural exit) by building a model based on historical data at Vienna airport in order to alert the operator accordingly. The tool was subsequently assessed through Real-Time Simulations (RTS) in order to further validate its potential in a real operational scenario. Although presenting interesting benefits, as the predicted information was not 100% reliable, controllers stated that they would use the information presented to check and monitor a situation more closely, as opposed to issuing an executive decision which highlights the need to involve end users, if possible, during research development to refine the algorithms and assess practical challenges.

Another aspect of AI technical robustness that has been the focus of recent research is the field of adversarial attacks. A comprehensive review of these methods can be found in [60] [61] [62] [63] [64]. The aim of the researchers is to introduce in their models' layers of robustness such that the models are not misled by out of distribution examples, known or unknown attacks and targeted or untargeted attacks. This is of upmost importance because it guarantees the accuracy of such models while safety is taken into consideration. A very good example in an ATM application is [73]. Here the authors tested adversarial data in different machine learning algorithms for trajectory prediction. This was done by producing samples similar to the original ones but which led to significant mistakes in the models. They gained a valuable amount of evidence on how the algorithms behaved in these situations and explored how to make them more robust by introducing adversarial data in training. They concluded that although there was a tradeoff between accuracy of predictions using adversarial examples as part of the training and robustness of the solution the results were extremely insightful. In a data-driven and safety critical operational environment, this type of research is fundamental and we are sure there will be more of it to come to understand robustness and resilience of the different AI solutions.

Regarding learning assurance of the algorithm life cycle, this is becoming a fundamental requirement for future regulations as per EASA´s guidelines [4] and recently also an important part of the EU Proposal for AI Act [50]. In this sense, data quality assurance is the first step of the learning assurance process and should be fundamental to all research projects. It covers the identification of the various datasets used for training and evaluation and the dataset preparation (including collection, labeling and processing). These aspects are normally well represented in research papers [54]. But it should also address considerations on the representativeness of the datasets. This includes for example acknowledgement of "known unknowns", this is, situations that might occur in operations that are not reflected in the datasets used for training, testing and validation and how this might affect the performance of the algorithm (i.e. rare events, system failures, emergency landings or unusual weather hazards). Finally, it should cover objectives on the independence between datasets and an evaluation of the bias and variance inherent to the data (for example the data might include an unusual travel season due to restrictions or weather events). In this sense, many of the papers reviewed describe in detail the datasets used in training and testing [49] [51] [53] [57] but some have chosen not include variables such as weather data or patterns that could affect greatly prediction results, generalization and scalability and this is acknowledge by the authors.

The actual tuning of the algorithm is something that is normally mentioned in many of the research papers reviewed, with [58] being a particularly good example that explores and specifies the selection and validation of key elements such as the activation function, the loss function, the initialization strategy, and the training hyperparameters of different algorithms, which all have the potential to influence the result of the training in terms of performance. The learning life cycle of the algorithm is of upmost importance and one of the key blocks of the EASA trustworthiness guidelines [4]. Furthermore, the EU AI Act proposal [50] suggests that an immutable log of algorithm training and life cycle should exist, in a similar way that deterministic ATM systems' design assurance is currently being documented to support their end-user acceptance, regulatory approval and subsequent operational deployment. These are all new considerations to be taken into account in future research projects.

Concerning learning assurance, xAI techniques are a key element to address the explainability challenge and they have been the subject of many research papers in recent years. Of the papers that we reviewed, [43] [12] and [44] used model-agnostic techniques for feature importance, and others such as [45] used tree-ensemble post-hoc explanation for simplification and feature relevance. [40] have done a really extensive and comprehensive review of xAI methods and many ATM research papers reviewed are starting to incorporate such techniques in order to explain the performance of their chosen algorithms.

C. Safety Assurance

According to the results of our AI Trust Survey, Safety and Security of the AI solution where the most important factors in order to trust any results and implementation of new AI-based solutions. From the perspective of trust assurance of the algorithms we consider safety as the main element of our framework as elements of security such as cybersecurity are not different from those already covered by current software design and security standards. Considering guidelines and feedback, three main elements of safety should be considered:

- Human Oversight
- Performance monitoring
- Safety Nets

The answers and additional comments from our survey highlighted that one of the most important means of

assuring trust and monitoring performance of AI solutions is human oversight. From an end user point of view, this helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach. This aspect has also been introduced as necessary for high-risk AI implementations such as those in critical infrastructure by the EU AI Act proposal [50].

The element of performance monitoring with respect to the safety requirements of safety critical systems was presented in [75] where they discussed a novel set of measures that can be used for the evaluation of the safe performance of ML algorithms. They argued that the challenge of safety performance using ML algorithms can be solved following a commonly used safety principle, namely safety reserves [74], which can be used to define safety margins where the predictions of the algorithm are guaranteed to be correct. These are a Safe True Positive (TP) threshold and a Safe True Negative (TN) threshold, where the first specifies a threshold that any observation with scores higher than it, is sufficiently guaranteed to be TP, and the last specifies a threshold that any observation with scores lower than it, is sufficiently guaranteed to be TN. Accordingly, observations outside these thresholds cannot be guaranteed to be correct. They referred to such observations as No Prediction (NP) and can not be used for safety critical decisions. Another approach is [75] where the authors described four types of monitoring: (1) input monitoring, for checking whether inputs are within acceptable bounds before they are given to the ML model; (2) environment monitoring, for checking that the observed environment matches any assumptions made during the ML workflow; (3) model internal monitoring, to protect against the effects of single event upsets; (4) output monitoring, by replicating a traditional system safety approach in which a high-integrity monitor is used alongside a lower-integrity item. These research papers are extremely important going forwards within the context of ATM as any changes in performance of an algorithm could result in a safety event.

During our survey, safety nets came as the most important safety measure within an ATM environment. In the examples above, if there were to be any changes in any of the monitoring measures that indicated a drop in performance or an anomaly in the output, a safety net would need to be in place. Depending on the area of operation and the role of the AI or ML solution, the safety net should consist of a warning or alert and time available to the operator to either override the system or ignore the outputs of the algorithm until the situation is investigated. Current examples of safety nets are for instance, Short Term Conflict Alert (STCA), which assists the controller in preventing collision between aircraft by generating a timely alert of a potential or actual infringement of separation minima or Area Proximity Warning (APW), which warns the controller about an unauthorised intrusion of an airspace volume by generating a timely alert of a potential or actual infringement of the required spacing to that airspace volume. Going forwards, we anticipate the research and development of safety nets for AI solutions will be an important focus their implementation in operations.

D. Transparency and Explainability Assurance

Finally, an exceptionally important assurance block in our framework that joins together all the others is the one that addresses the transparency and explainability of the solution. This is crucial from the point of view of regulators and end users. As we have mentioned previously, it is critical to understand the solution, its advantages and its limitations in order to trust it and to use it. In this sense, it also needs to be auditable so it can satisfy regulators and potential investigations as well as provide in detailed explanations to experts end users.

In terms of explainability, the work done on xAI techniques, as mentioned earlier, are primarily looking at global or local explanations selecting which features in the model are important to understand an outcome. However, even though the research on xAI techniques is currently abundant and feature exploration is being investigated by many authors ([65][68][69]), a major limitation of existing work is that the explanations are designed based on the intuition of researchers rather than focusing on the demands and understanding of end-users. From an end user's point of view, the goal of a good explanation is to understand and trust the functioning and outcome of an algorithm. Therefore, for researchers it is of upmost importance to evaluate what makes an explanation user-oriented and user-friendly and present results in a way that is clearly understandable [67].

Finally, auditability of the model is also an area that is gaining importance. This is especially true with current proposals and guidelines for regulations and possible safety certification of systems that embed an AI algorithm [4] [37] [50]. Auditing of AI is shaping itself as an imperative tool as AI may bring unprecedented and unpredictable consequences [70]. Auditability entails the enablement of the assessment of algorithms, data and design processes and as such, it depends greatly on the learning assurance and technical assurance processes as well as the safety assurance methods in our framework. As more progress is made in these areas the research projects would need to tune in their design and development to make sure all the elements of auditability and trust are present. We believe that our framework presents a very good baseline to do so.

VI. SUMMARY AND RECOMMENDATIONS

In this paper we have set up the basis of an xAI Trust framework in order to address the gap between research and implementation solutions within an ATM environment. We have highlighted current guidelines and recommendations by regulators for trustworthy AI and we have addressed what constitute trust on AI automated solutions in ATM for end users through an AI Trust Survey answered by stakeholders of the Fly2Plan project. Through a literature review we have identified the areas that need more research

on elements of AI trustworthiness in ATM such as those reflected in our framework.

Due to recent developments on guidelines and to truly bridge the gap between research and operational implementation, we consider that, from now on, there should be more consideration on demonstrating the trustworthiness of research project outcomes in a more holistic way.

We believe that our framework provides a strong basis to do so and recommend its use in setting up future ATM research projects. The framework would also provide a base for comparison of different techniques and applications in similar scenarios in order to assess advantages and disadvantages of each in an operational environment. As next steps, we will be applying our framework to a practical case scenario in AI research to test its efficacy and subsequently evaluating it and obtaining end user feedback to improve it further.

ACKNOWLEDGMENTS

This work has been completed as part of the Future of Flight UK challenge sponsored by the UK Research and Innovation.

REFERENCES

[1] EASA Report: "A Harmonised European Approach to a Performance-Based Environment (PBE)", August 2014.

[2] FAA Performance-Based Operations Aviation Rulemaking Committee Charter, August 2018

[3] U. D. Ferrell and A. H. A. Anderegg, "Applicability of UL 4600 to Unmanned Aircraft Systems (UAS) and Urban Air Mobility (UAM)," 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC), pp. 1-7, 2020.

[4] EASA 2020, Artificial Intelligence Roadmap Human-centric approach to AI in aviation

[5] ICAO Doc 9971 - Manual on Collaborative Air Traffic Management, 2014.

[6] Z. Wang, M. Liang, and D. Delahaye, 'Data-driven Conflict Detection Enhancement in 3D Airspace with Machine Learning', in 2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT), pp. 1–9, 2020.

[7] D. Zhou, X. Zhuang, H. Zuo, H. Wang, and H. Yan, "Deep Learning-Based Approach for Civil Aircraft Hazard Identification and Prediction", IEEE Access, vol. 8, pp. 103665–103683, 2020.

[8] D. Guijo-Rubio, P. A. Gutiérrez, C. Casanova-Mateo, J. Sanz-Justo, S. Salcedo-Sanz, and C. Hervás-Martínez, "Prediction of low-visibility events due to fog using ordinal classification", Atmospheric Research, vol. 214, pp. 64–73, 2018.

[9] Y. Liu, M. Hansen, A. Pozdnukhov, and D. Zhang, "Using machine learning to analyze air traffic management actions: Ground delay program case study", Transportation Research Part E: Logistics and Transportation Review, vol. 131, pp. 80–95, 2019.

[10] T. G. Puranik, N. Rodriguez, and D. N. Mavris, "Towards online prediction of safety-critical landing metrics in aviation using supervised machine learning", Transportation Research Part C: Emerging Technologies, vol. 120, p. 102819, 2020.

[11] C. C. Morgan, M. Ellejmi, F. Herrema, and R. Curran, "Validation of the Runway Utilisation concept", 9th SESAR Innovation Days, 2019.

[12] D. Martinez, S. Belkoura, S. Cristobal, I. Research, F. Herrema, and P. Wachter, "A Boosted Tree Framework for Runway Occupancy and Exit Prediction", 8th SESAR Innovation Days, 2018.

[13] R. Marcos, R. Herranz, R. R. Vázquez, P. García-Albertos, and O. G. C. Ros, "Application of Machine Learning for ATM Performance Assessment – Identification of Sources of En-Route Flight Inefficiency", 8th SESAR Innovation Days, 2018.

[14] A. Sternberg, J. Soares, D. Carvalho, and E. Ogasawara, "A Review on Flight Delay Prediction", Transport Reviews, pp. 1–30, 2020.

[15] D. Rios Insua, C. Alfaro, J. Gomez, P. Hernandez-Coronado, and F. Bernal, "Forecasting and assessing consequences of aviation safety occurrences", Safety Science, vol. 111, pp. 243–252, 2019.

[16] C. Ma, Q. Cai, S. Alam, and V. N. Duong, "Airspace Capacity Overload Identification Using Collision Risk Pattern", in 2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT), pp. 1–9, 2020.

[17] E. Mangortey, T. Puranik, O. Pinon, and D. Mavris, "Prediction and Analysis of Ground Stops with Machine Learning". AIAA Scitech Forum, 2020.

[18] K. Sheridan, T. Puranik, E. Mangortey, O. Pinon, M. Kirby, and D. Mavris, "An Application of DBSCAN Clustering for Flight Anomaly Detection During the Approach Phase". AIAA Scitech Forum, 2020.

[19] D. Bari, "Visibility Prediction Based on Kilometric NWP Model Outputs Using Machine-Learning Regression", in IEEE 14th International Conference on e-Science (e-Science), Oct. 2018, pp. 278–278, 2018.

[20] X. Zhang and S. Mahadevan, "Ensemble machine learning models for aviation incident risk prediction", Decision Support Systems, vol. 116, pp. 48–63, 2019.

[21] B. Yu, Z. Guo, S. Asian, H. Wang, and G. Chen, "Flight delay prediction for commercial air transport: A deep learning approach", Transportation Research Part E: Logistics and Transportation Review, vol. 125, pp. 203–221, 2019.

[22] A. Fernández Llamas, D. Martínez, S. Cristóbal, F. Schwaiger, J. Nuñez, and J. Ruiz, "Flight Data Monitoring (FDM) Unknown Hazards detection during Approach Phase using Clustering Techniques and AutoEncoders", 9th SESAR Innovation Days, 2019.

[23] E. Andres, D. Gonzalez-Arribas, M. Sanjurjo-Rivo, M. Soler, and M. Kamgarpour, "GPU-Accelerated RRT for Flight Planning Considering Ensemble Forecasting of Thunderstorms", 10th SESAR Innovation Days, 2020.

[24] G. Hondet, L. Delgado, and G. Gurtner, "Airline Disruption Management with Aircraft Swapping and Reinforcement Learning", 8th SESAR Innovation Days, 2018.

[25] D. Pham, N. P. Tran, S. K. Goh, S. Alam, and V. Duong, "Reinforcement Learning for Two-Aircraft Conflict Resolution in the Presence of Uncertainty", in IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF) 2019.

[26] E. Mangortey, D. Monteiro, J. Ackley, Z. Gao, T.G. Puranik, M. Kirby, O. J. Pinon and D.N. Mavris, "Application of Machine Learning Techniques to Parameter Selection for Flight Risk Identification", AIAA Scitech 2020 Forum, 2020.

[27] R. Dalmau, F. Ballerini, H. Naessens, S. Belkoura, and S. Wangnick, "Improving the predictability of take-off times with machine learning: a case study for the Maastricht upper area control centre area of responsibility", in 9th SESAR Innovation Days, 2019.

[28] W. Deng, J. Xu, and H. Zhao, "An Improved Ant Colony Optimization Algorithm Based on Hybrid Strategies for Scheduling Problem", IEEE Access, vol. 7, pp. 20281–20292, 2019.

[29] C. Chilson, K. Avery, A. McGovern, E. Bridge, D. Sheldon, and J. Kelly, "Automated detection of bird roosts using NEXRAD radar data and Convolutional Neural Networks, Remote Sensing in Ecology and Conservation", vol. 5, no. 1, pp. 20–32, 2019.

[30] R. Mori and D. Delahaye, "Simulation-Free Runway Balancing Optimization Under Uncertainty Using Neural Network", 9th SESAR Innovation Days, 2019.

[31] D. Martınez, A. Fernández, P. Hernández, S. Cristóbal, F. Schwaiger, J. M. Nunez, & J. M. Ruiz, "Forecasting Unstable Approaches with Boosting Frameworks and LSTM Networks". 9th SESAR Innovation Days, 2019.

[32] Z. Wang, M. Liang, and D. Delahaye, "Automated data-driven prediction on aircraft Estimated Time of Arrival", J. Air Transp. Manag., vol. 88, 2020.

[33] M. Poppe, R. Scharff, J. Buxbaum, and D. Fieberg, "Flight level prediction with a deep feedforward network", in 8th SESAR Innovation Days, 2018.

[34] C. E. Verdonk Gallego, V. F. Gómez Comendador, F. J. Sáez Nieto, G. Orenga Imaz, and R. M. Arnaldo Valdés, "Analysis of air traffic

control operational impact on aircraft vertical profiles supported by machine learning", Transportation Research Part C: Emerging Technologies, vol. 95, pp. 883–903, 2018.

[35] W. A. Khan, H.-L. Ma, X. Ouyang, and D. Y. Mo, "Prediction of aircraft trajectory and the associated fuel consumption using covariance bidirectional extreme learning machines", Transportation Research Part E: Logistics and Transportation Review, vol. 145, p. 102189, 2021.

[36] Z. Wang, M. Liang, D. Delahaye and W. Wu, "Learning Real Trajectory Data to Enhance Conflict Detection Accuracy in Closest Point of Approach", 13th USA/Europe ATM R&D Seminar, 2019.

[37] SAE G-34/EUROCAE WG-114 Fact Sheet May 2021

[38] E. Mangortey, D. Monteiro, J. Ackley, Z. Gao, T.G. Puranik, M. Kirby, O. J. Pinon and D.N. Mavris, "Application of Machine Learning Techniques to Parameter Selection for Flight Risk Identification", AIAA Scitech 2020 Forum.

[39] F. Mamalet, E. Jenn, G. Flandin, H. Delseny, C. Gabreau, et al.. White Paper Machine Learning in Certified Systems. [Research Report DEEL Project] IRT Saint Exupéry; ANITI. 2021.

[40] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, "Explainable Artificial Intelligence (xAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", Information Fusion, vol. 58, pp. 82–115, 2020.

[41] Independent High-Level Expert Group on Artificial Intelligence (AI HLEG), Ethics Guidelines for Trustworthy AI, 2019.

[42] SESAR, Automation in Air Traffic Management, Long term vision and initial research roadmap, 2020.

[43] M. C. R. Murça, R. J. Hansman, L. Li, and P. Ren, "Flight trajectory data analytics for characterization of air traffic flows: A comparative analysis of terminal area operations between New York, Hong Kong and Sao Paulo", Transp. Res. Part C Emerg. Technol., vol. 97, no. October, pp. 324–347, 2018.

[44] X. Wang, A. E. I. Brownlee, J. R. Woodward, M. Weiszer, M. Mahfouf, and J. Chen, "Aircraft taxi time prediction: Feature importance and their implications", Transportation Research Part C: Emerging Technologies, vol. 124, p. 102892, 2020.

[45] G. Hondet, L. Delgado, and G. Gurtner, "Airline disruption management with aircraft swapping and reinforcement learning", in SESAR Innovation Days, 2018.

[46] S. Li, C. Ma, Q. Li, J. Zeng, and L. Wang, "Application of Improved Ant Colony Algorithm in Flight Path Planning", in 2020 IEEE International Conference on Information Technology,Big Data and Artificial Intelligence (ICIBA), vol. 1, pp. 763–771, 2020.

[47] M. F. Yazdi, S. R. Kamel, S. J. M. Chabok, and M. Kheirabadi, "Flight delay prediction based on deep learning and Levenberg-Marquart algorithm", Journal of Big Data, vol. 7, no. 1, 2020.

[48] R. Alligier and D. Gianazza, "Learning aircraft operational factors to improve aircraft climb prediction: A large scale multi-airport study", Transp. Res. Part C Emerg. Technol., vol. 96, pp. 72–95, 2018.

[49] C. S. Bosson and T. Nikoleris, "Supervised learning applied to air traffic trajectory classification", in AIAA Information Systems-AIAA Infotech at Aerospace, 2018.

[50] EU Proposal for Artificial Intelligence Act, 2021

[51] F. Herrema, R. Curran, S. Hartjes, M. Ellejmi, S. Bancroft, & M. Schultz 2019, A machine learning model to predict runway exit at Vienna airport. Transportation Research Part E: Logistics and Transportation Review, 131, 329-342.

[52] T. Miller, 2017, ''Explanation in artificial intelligence: Insights from the social sciences.'' [Online]. Available: https://arxiv.org/abs/1706.07269

[53] Z. Wang, M. Liang, and D. Delahaye, "Automated data-driven prediction on aircraft Estimated Time of Arrival", J. Air Transp. Manag., vol. 88, 2020.

[54] Chakrabarty N. "A data mining approach to flight arrival delay prediction for american airlines". In 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON). New York: IEEE; 2019.

[55] Meel P, et al. "Predicting flight delays with error calculation using machine learned classifiers". In 7th International Conference on Signal Processing and Integrated Networks (SPIN). New York: IEEE; 2020.

[56] Ye B, et al. "A methodology for predicting aggregate flight departure delays in airports based on supervised learning". Sustainability. 12(7):2749, 2020.

[57] Shao W, et al. "Flight delay prediction using airport situational awareness map". In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2019.

[58] A. Maheshwari, N. Davendralingam, D. DeLaurentis, "A Comparative Study of Machine Learning Techniques for Aviation Applications", 10.2514/6.2018-3980, 2018.

[59] M. Henne, A. Schwaiger, K. Roscher and G. Weiss, "Benchmarking uncertainty estimation methods for deep learning with safety-related metrics", SafeAI@ AAAI, pp. 83-90, 2020.

[60] E.R. Balda, A. Behboodi and R. Mathar, "Adversarial Examples in Deep Neural Networks: An Overview" in Deep Learning: Algorithms and Applications". Studies in Computational Intelligence, Cham:Springer, vol. 865, 2020.

[61] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," arXiv preprint arXiv:1810.00069, 2018

[62] H. Chacon, S. Silva, and P. Rad, "Deep learning poison data attack detection," in 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, pp. 971–978, 2019.

[63] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in 2017 Ieee symposium on security and privacy (sp). IEEE, pp. 39–57, 2017.

[64] Qiu, Shilin; Liu, Qihe; Zhou, Shijie; Wu, Chunjiang. "Review of Artificial Intelligence Adversarial Attack and Defense Technologies" Applied Sciences; Basel Vol. 9, Iss. 5, 2019.

[65] X. Wang, A. E. I. Brownlee, J. R. Woodward, M. Weiszer, M. Mahfouf, and J. Chen, "Aircraft taxi time prediction: Feature importance and their implications" Transportation Research Part C: Emerging Technologies, vol. 124, p. 102892, 2020.

[66] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, "Explainable Artificial Intelligence (xAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", Information Fusion, vol. 58, pp. 82–115, 2020.

[67] M.T. Dzindolet, S.A. Peterson, R.A. Pomranky, L.G. Pierce and H.P. Beck, "The role of trust in automation reliance". International Journal of Human–Computer Studies, 58, 697– 718, 2003.

[68] S.M. Lundberg, S.-I. Lee, "A unified approach to interpreting model predictions", Advances in Neural Information Processing Systems, pp. 4765-4774, 2017.

[69] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier", ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 1135-1144, 2016.

[70] A. Toader, "Auditability of AI Systems – Brake or Acceleration to Innovation?" (November 11, 2019). Available at SSRN: https://ssrn.com/abstract=3526222

[71] T. Miller, ''Explanation in artificial intelligence: Insights from the social sciences.'' [Online]. Available: https://arxiv.org/abs/1706.07269, 2017.

[72] SESAR, JU. "European ATM Master Plan: Digitalising Europe's Aviation Infrastructure" 2020, accessed on https://www.sesarju.eu/masterplan2020

[73] S.M. Hashemi, R.M. Botez, T.L. Grigorie, "New Reliability Studies of Data-Driven Aircraft Trajectory Prediction". Aerospace, 7, 145, 2020.

[74] N. Moller and S. O. Hansson, "Principles of engineering safety: Risk ¨and uncertainty reduction," Reliability Engineering and System Safety,vol. 93, no. 6, pp. 798–805, 2008.

M. Gharib, and A. Bondavalli, "On the Evaluation Measures for Machine Learning Algorithms for Safety-Critical Systems", 15th European Dependable Computing Conference (EDCC), 141-144, 2019.

An explainable artificial intelligence (xAI) framework for improving trust in automated ATM tools

Sanchez Hernandez, Carolina

2021-11-15

Sanchez Hernandez C, Ayo S, Panagiotakopoulos D. (2021) An explainable artificial intelligence (xAI) framework for improving trust in automated ATM tools. In: 2021 AIAA/IEEE 40th Digital Avionics Systems Conference (DASC), 3-7 October 2021, San Antonio, USA
https://doi.org/10.1109/DASC52595.2021.9594341