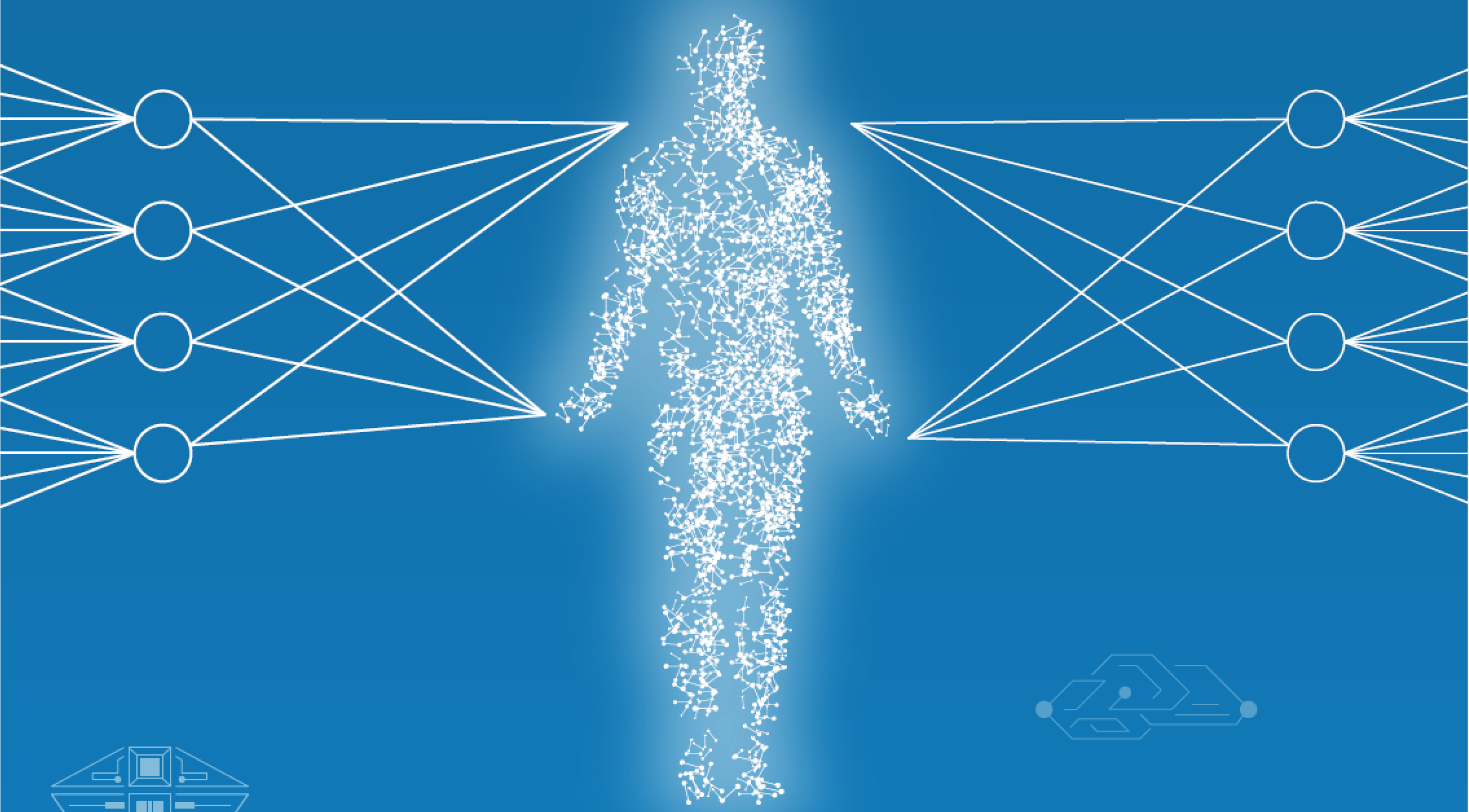
European Union Aviation Safety Agency

EASA Concept Paper: First usable guidance for Level 1 & 2 machine learning applications

A deliverable of the EASA AI Roadmap



February 2023

Proposed Issue 02

easa.europa.eu/ai

Table of Contents

Author	Guillaume Soudain — EASA AI Programme Manager
Reviewer	Francois Triboulet — EASA ATM/ANS Expert-Coordinator (SNE)
Approver	Alain Leroy — EASA Chief Engineer

A. Foreword

In line with the two first major milestones of the European Union Aviation Safety Agency (***EASA) Artificial Intelligence (AI) Roadmap 1.0 Phase I*** (‘Exploration and first guidance development’), this concept paper presents a first set of objectives for ***Level 1 Artificial Intelligence*** (‘assistance to human’) and ***Level 2 Artificial Intelligence*** (‘human-machine collaboration’), in order to anticipate future EASA guidance and requirements for ***safety-related machine learning (ML)*** applications.

It aims at guiding applicants when ***introducing AI/ML technologies*** into systems intended for use in safety-related or environment-related applications in all domains covered by the ***EASA Basic Regulation*** (Regulation (EU) 2018/1139).

It covers only an initial set of AI/ML techniques and will be enriched with more and more advanced techniques, as the EASA AI Roadmap is implemented.

This document provides a first set of usable objectives; however, it does not constitute at this stage definitive or detailed guidance. It will serve as a basis for the ***EASA AI Roadmap 1.0 Phase II*** (‘AI/ML framework consolidation’) when formal regulatory development comes into force.

On a more general note, it is furthermore important to point out to the ongoing discussions regarding the **EU Commission’s regulatory package on AI, published on 21 April 2021**[1]. While, according to that Commission proposal[2], the EASA Basic Regulation will be considered as one among various specific, sectorial frameworks, interdependencies between the final EU AI Regulation and the EASA Basic Regulation and its delegated and implementing acts can be expected. Both the ´EASA Roadmap on AI´ as well as this present guidance document will thus have to continuously take this into account and remain aligned.

After setting the scene in an introductory Chapter (Chapter B), reminding the reader of the four ***AI trustworthiness ‘building blocks’***, Chapter C develops the guidelines themselves, dealing with:

— ***trustworthiness analysis*** (Section C.0);

— ***AI assurance*** (Section C.3);

— ***human-factors for AI*** (Section C.4); and

— ***safety risk mitigation*** (Section C.5).

Chapter D introduces ***proportionality*** which is intended to allow the customisation of the objectives to the specific AI applications.

[1] EU Commission - Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM/2021/206 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206.

[2] The Commission stated that: ‘Faced with the rapid technological development of AI and a global policy context where more and more countries are investing heavily in AI, the EU must act as one to harness the many opportunities and address challenges of AI in a future-proof manner. To promote the development of AI and address the potential high risks it poses to safety and fundamental rights equally, the Commission is presenting both a proposal for a regulatory framework on AI and a revised coordinated plan on AI.’

Chapter E aims at identifying the possible ***impacts*** of the introduction of AI in the different ***implementing rules (IRs), certification specifications (CSs), acceptable means of compliance (AMC) and guidance material (GM)*** in the domains covered by the EASA Basic Regulation.

Chapter F provides the reader with a set of ***use cases*** from different aviation domains where the guidelines have been (partially) applied. These use cases serve as demonstrators to verify that the objectives defined in this guidance document are achieved.

Until IRs or AMC are available, this guidance can be used as an enabler or an all-purpose instrument facilitating the preparation of the approval or certification of products, parts and appliances introducing AI/ML technologies. In this respect, this guidance should benefit all aviation stakeholders, end users, applicants, certification or approval authorities.

B. Introduction

Following the publication in December 2021 of the EASA concept paper 'First usable guidance for Level 1 machine learning applications', this guidance document represents the next step in the implementation of the EASA AI Roadmap v1.0. It complements the first set of technical objectives and organisation provisions that EASA anticipates as necessary for the approval of both ***Level 1 AI applications*** ('assistance to human') and ***Level 2 AI applications*** ('human-machine teaming'). Where practicable, the document identifies anticipated means of compliance (MOC) and guidance material which could be used to comply with those objectives.

Note: The anticipated MOC will be completed based on the outcome of research and innovation projects, in particular the Horizon Europe 'Machine Learning application approval' (MLEAP)[3], on the discussions triggered within certification projects, as well as based on the progress of industrial standards such as the one that is under work in the joint EUROCAE/SAE WG-114/G-34 or EUROCAE/RTCA WG-72/SC-216. EASA also follows the progress of other working groups on AI, in particular ISO/IEC SC42 and CEN CENELEC JTC21.

The goal of this document is therefore twofold:

— to allow applicants proposing to use AI/ML solutions in their projects to have an early visibility on the possible expectations of EASA in view of an approval. This material may be referred to by EASA through dedicated project means (e.g. a Certification Review Item (CRI) for certification projects);

— to establish a baseline for ***Level 1 and Level 2 AI applications*** that will be further refined for ***Level 3 AI applications ('more autonomous machine')***[4].

Disclaimer: To the best of EASA´s knowledge, the information contained in these guidelines is accurate and reliable on the date of publication and reflects the state of the art in terms of approval/certification of AI/ML solutions. EASA does, however, not assume any liability whatsoever for the accuracy and completeness of these guidelines. Any information provided therein does not constitute in itself any warranty of fitness to obtain a final EASA approval. These guidelines will evolve over the next 2 years through publication of a document addressing Level 3 AI applications, while being updated based on their application to Level 1 and Level 2 AI applications. They may evolve as well depending on the research and technological development in the dynamic field of AI research.

1. Statement of issue

AI is a broad term, and its definition has evolved as technology has developed. EASA therefore chose in the EASA AI Roadmap 1.0 a wide-spectrum definition that is 'any technology that appears to emulate the performance of a human'.

The EASA AI Roadmap has defined the following taxonomy for AI:

[3] The status and reports of the MLEAP project are provided on the EASA website: https://www.easa.europa.eu/en/research-projects/machine-learning-application-approval

[4] See Section B.5 for more information on the proposed classification of AI-based systems in 3 levels.

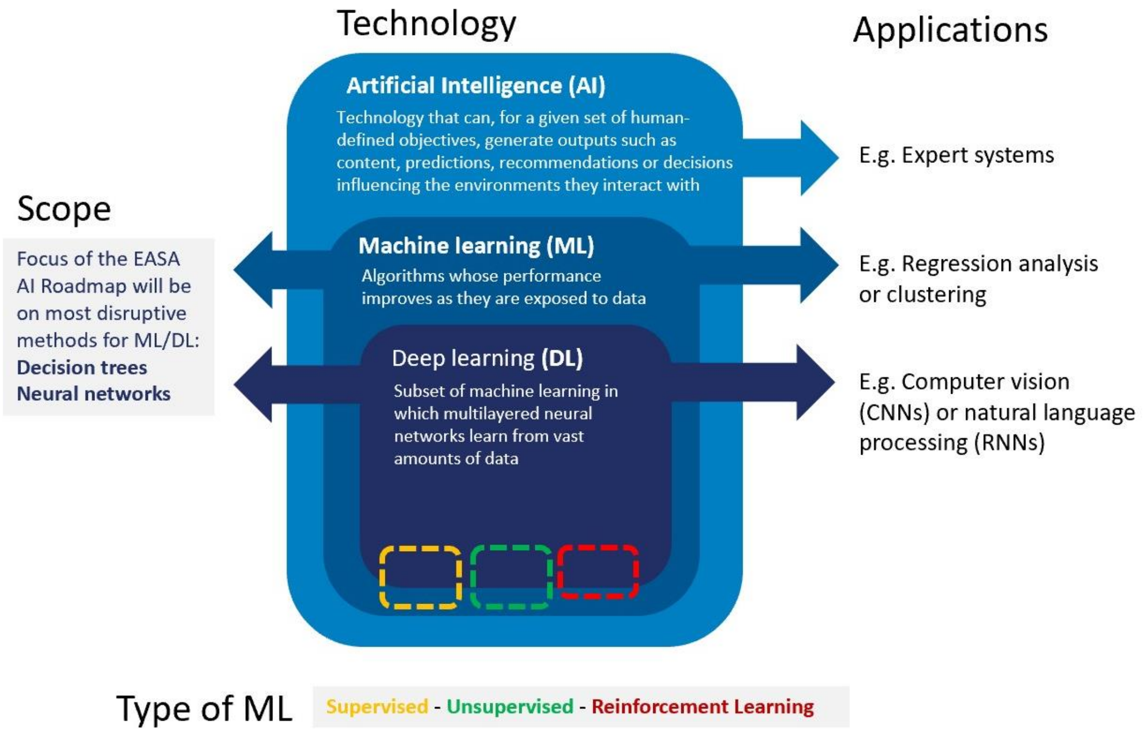


Figure 1 — AI taxonomy in EASA AI Roadmap

The current breakthrough is linked with ML, which is the ability of computer systems to improve their performance by exposure to data without the need to follow explicitly programmed instructions. ***Deep learning (DL)*** is a subset of ML that emerged with deeper neural networks (NNs), leading to large improvements in performance in the recent years. DL produced significant improvements for many problems in computer vision and natural language processing (NLP), enabling new use cases and accelerating AI adoption. This is the reason why EASA AI Roadmap 1.0 and this Level 1 & 2 AI guidance are focusing on ***data-driven AI*** approaches. The initial scope is however limited to supervised learning techniques. This limitation will be removed in the next release of this guidance document through a planned extension to unsupervised and reinforcement learning.

Data-driven learning techniques are a major opportunity for the aviation industry but come also with a significant number of challenges with respect to the trustworthiness of ML and DL solutions. Here are some of the main challenges addressed through this first set of EASA guidelines:

— Adapting assurance frameworks to cover the specificities of identified AI techniques and address development errors in AI-based systems and their constituents;

— Dealing with the particular sources of uncertainties associated with the use of AI/ML technology;

— Creating a framework for data management, to address the correctness (bias mitigation) and completeness/representativeness of data sets used for the ML items training and their verification;

— Addressing model bias and variance trade-off in the various steps of ML processes;

— Elaborating pertinent guarantees on robustness and on absence of 'unintended function' in ML/DL applications;

— Coping with limits to human comprehension of the ML application behaviour, considering their stochastic origin and ML model complexity;

— Managing shared operational authority in novel types of human-AI teaming (HAT);

— Managing the mitigation of residual risk in 'AI black box'. The expression 'black box' is a typical criticism oriented at AI/ML techniques, as the complexity and nature of ML models bring a level of opaqueness that make them look like unverifiable black boxes (unlike rule-based software); and

— Enabling trust by end users.

It is also important to note that the scope of the Roadmap will be extended to cover additional AI techniques such as symbolic (rule-based) AI techniques in an updated EASA AI Roadmap 2.0, foreseen to be published by mid-2023. This update will offer the opportunity to address approaches of hybridisation between data-driven and rule-based AI, that are anticipated to take best advantage of both techniques while mitigating their respective limitations.

2. AI trustworthiness framework overview

To address the challenges of data-driven learning approaches, EASA AI Roadmap 1.0 identifies four '***building blocks***' that are considered essential in creating a framework for ***trustworthy AI*** and for enabling readiness for use of AI/ML in aviation. Based on the novel concepts developed in this document (see Section B.6), and in anticipation of EASA AI Roadmap 2.0, two of the original building blocks require an extension in terms of scope. This is the reason why the 'Learning Assurance' block has been promoted to 'AI Assurance' and the 'AI Explainability' now covers more broadly the notion of 'Human factors for AI':

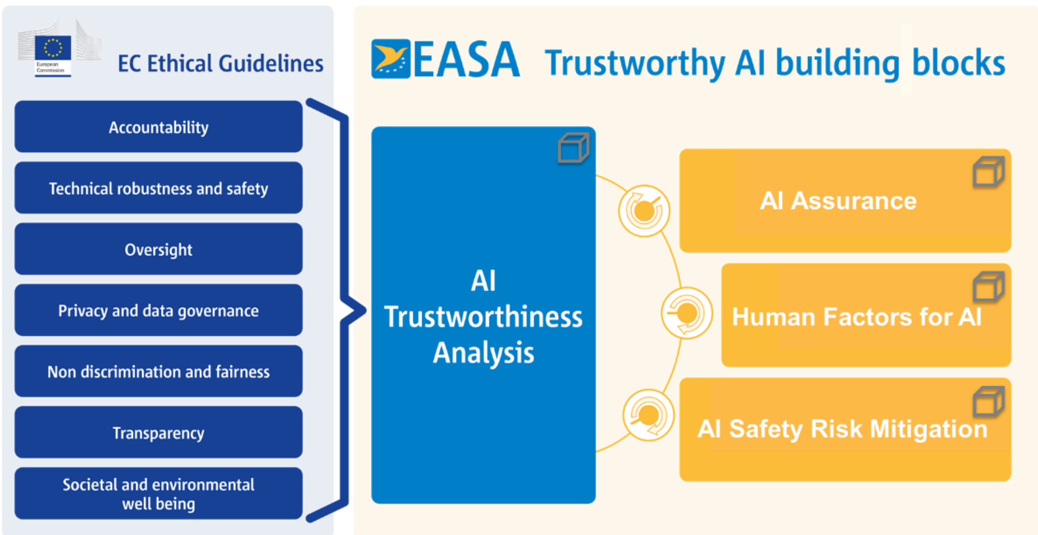


Figure 2 — EASA AI trustworthiness roadmap building blocks

The ***AI trustworthiness analysis*** building block, being one of those four building blocks, creates an interface with the EU Ethical Guidelines developed by the EU Commission (EU High-Level Expert Group on AI, 2019), and as such serves as a gate to the three other technical building blocks. The trustworthiness analysis starts with a ***characterisation of AI application***, includes an ethics-based assessment, and also encompasses the ***safety assessment*** and ***security assessment*** that are key elements of the trustworthiness analysis concept. All three ***assessments (i.e. safety, security and ethics-based)*** are important prerequisites in the development of any system developed with or embedding AI/ML, and are not only preliminary steps but also integral processes towards approval of such innovative solutions.

The ***AI assurance*** building block is intended to address the AI specific guidance pertaining to the AI-based system. It encompasses two major topics. Firstly, ***learning assurance*** covers the paradigm shift from programming to learning, as the existing development assurance methods are not adapted to cover learning processes specific to AI/ML. Secondly, ***development (respectively post-ops) explainability*** deals with the capability to provide human users with understandable, reliable and relevant information with the appropriate level of details on how an AI/ML application produces (respectively has produced) its results. This building block also includes the data recording capabilities, addressing two specific operational and post-operational purposes: on the one hand the continuous monitoring of the safety of the AI-based system and on the other hand the support to incident or accident investigation.

The ***human factors for AI*** building block introduces the necessary guidance to account for the specific human factors needs linked with the introduction of AI. Among other aspects, ***AI operational explainability*** deals with the capability to provide the human end users with understandable, reliable and relevant information with the appropriate level of details and with appropriate timing on how an AI/ML application produces its results. This block also introduced the concept of ***human-AI teaming*** to ensure adequate cooperation or collaboration between human end users and AI-based systems to achieve certain goals.

The ***AI safety risk mitigation*** building block considers that we may not always be able to open the 'AI black box' to the extent required and that the associated residual risk may need to be addressed to deal with the inherent uncertainty of AI.

All four building blocks have an importance in gaining confidence in the trustworthiness of an AI/ML application.

The detailed content of each building block is further described in the chapters as indicated in the following figure.

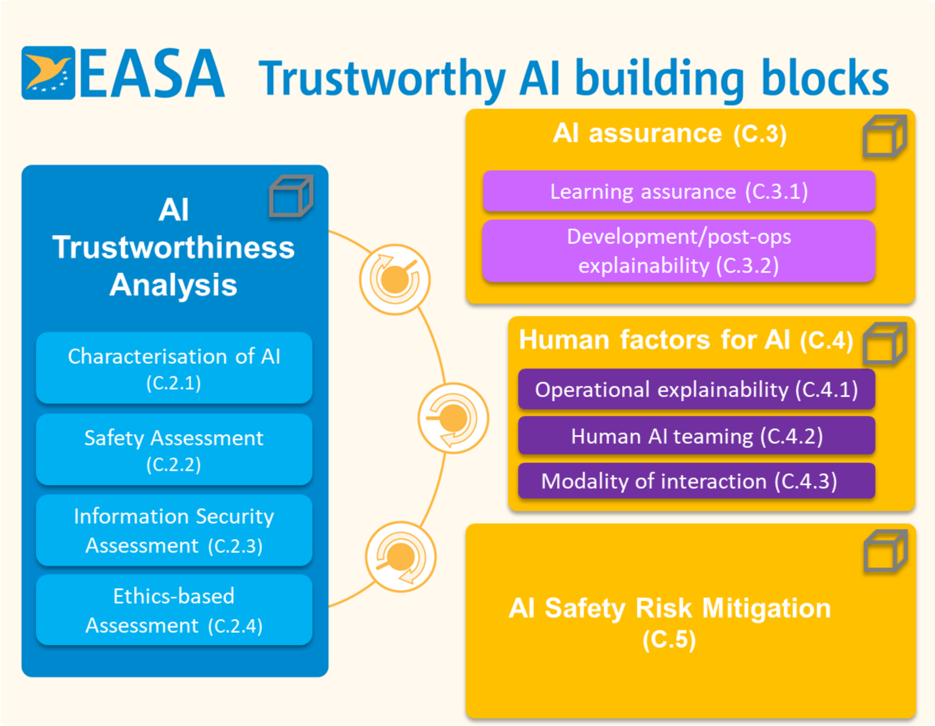


Figure 3 — EASA AI trustworthiness building blocks

The trustworthiness analysis is always necessary and should be performed in its full spectrum for any application. For the other three building blocks, the potentiometers represented in Figure 2 and Figure 3 indicate that the depth of guidance could be adapted depending on the application, as described in Chapter D.

3. Terminology and scope of the document

In EASA AI Roadmap 1.0, the initial focus has been put on ***data-driven AI*** approaches. Those can be further divided considering types of learning:

— ***Supervised learning*** — this strategy is used in cases where there is a labelled data set available to learn from. The learning algorithm processes the input data set, and a cost function measures the difference between the ML model output and the labelled data. The learning algorithm then adjusts the parameters to increase the accuracy of the ML model.

— ***Unsupervised learning*** — this strategy is used in cases where there is no labelled data set available to learn from. The learning algorithm processes the data set, and a cost function indicates whether the ML model has converged to a stable solution. The learning algorithm then adjusts the parameters to increase the accuracy of the ML model.

— ***Reinforcement learning*** — this strategy is used in cases where there is an environment available for an agent to 'practise' in. The agent(s) is(are) rewarded positively or negatively based on the effect of the actions on the environment. The ML model parameters are updated from this trial-and-error sequence to optimise the outcome.

There exist some other techniques, which have not been listed here. In particular, there are soft boundaries between some of those categories; for instance, unsupervised and supervised learning techniques could be used in conjunction with each other in a semi-supervised learning approach.

The scope of this document includes at this initial stage ***supervised learning*** approaches and will be further expanded in the next update to cover other types of learning.

Considering this scope, the following figure details the decomposition of an AI-based system and allows introducing the terminology that is used in the rest of the document when dealing with the system or portions of it.

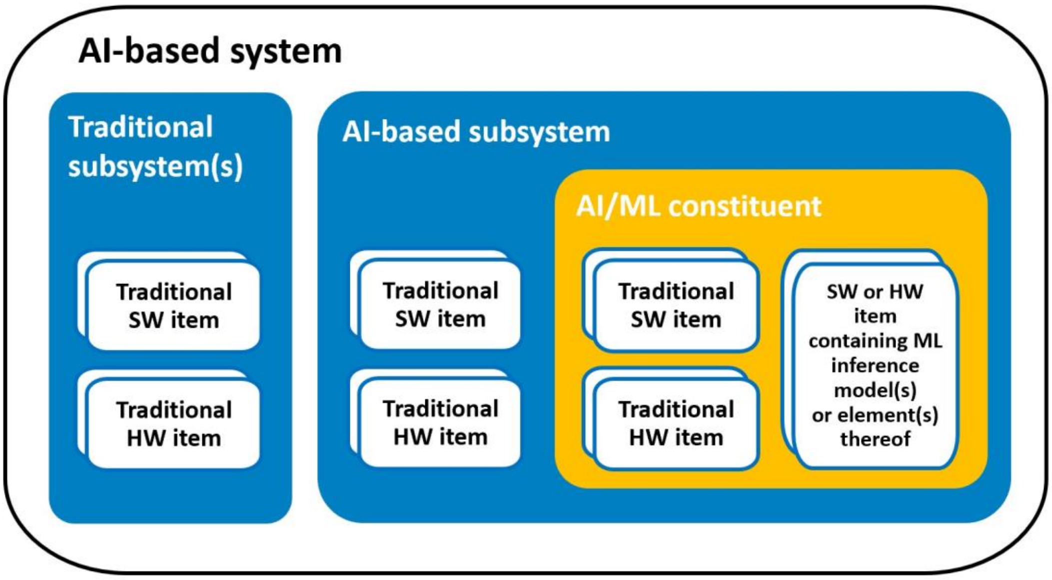


Figure 4 — Decomposition of the AI-based system

Where:

— an AI-based system is composed of several traditional subsystems, and at least one of them is an AI-based subsystem;

— an AI-based subsystem embeds at least one AI/ML constituent;

— an AI/ML constituent is a collection of hardware and/or software items (including the necessary pre- and post-processing elements), and at least one specialised hardware or software item containing one (or several) ML model(s), further referred to as ‘AI/ML item’ in this document;

— the traditional hardware and software items do not include an ML inference model.

4. Criticality of AI applications

Depending on the safety criticality of the application, and on the aviation domain, an assurance level is allocated to the AI-based (sub)system (e.g. development assurance level (DAL) for initial and continuing airworthiness or air operations, or software assurance level (SWAL) for air traffic management/air navigation services (ATM/ANS)).

A modulation of the objectives of this document based on the assurance level has been introduced in Chapter D 'Proportionality of the guidance'.

With the current state of knowledge of AI and ML technology, EASA anticipates a limitation on the validity of applications when AI/ML constituents include IDAL A or B / SWAL 1 or 2 / AL 1, 2 or 3 items. Moreover, no assurance level reduction should be performed for items within AI/ML constituents. This limitation will be revisited when experience with AI/ML techniques has been gained.

5. Classification of AI applications — overview

The EASA AI Roadmap identifies three general AI levels. This scheme has been proposed based on prognostics from industry regarding the types of use cases foreseen by AI-based systems. Indeed, these three levels can be related to the staged approach that most of the industrial stakeholders are planning for the deployment of AI applications, starting with assisting functions (Level 1 AI), then making a step towards more human-machine teaming (Level 2 AI) and at last seeking for more autonomy of the machine (Level 3 AI).

An additional split for level 2 AI has been introduced in this document, based on the human factors guidance developed in Section C.4. The resulting refinement of the three scenarios is considered in the following figure:

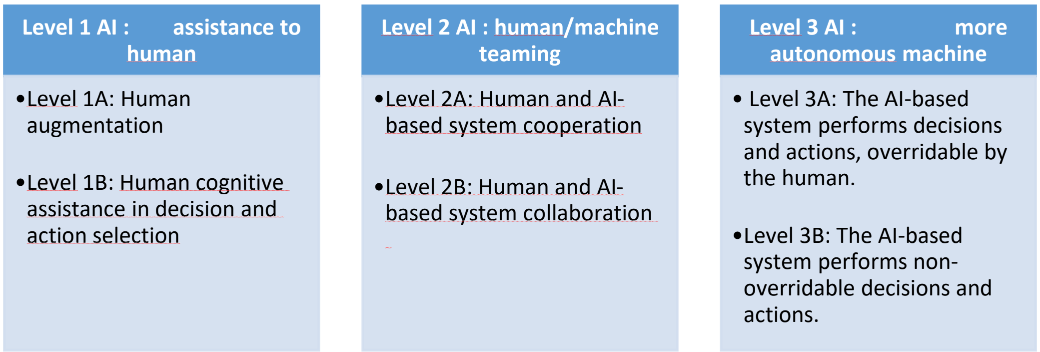


Figure 5 — Classification of AI applications

Detailed guidance on how to classify an AI-based system is provided in Section C.2.1.

Chapter D 'Proportionality of the guidance' introduces the applicability of the objectives to each AI level (i.e. Level 1A, 1B, 2A and 2B), and will be completed at a later stage with considerations for Level 3 AI.

6. Novel concepts developed for data-driven AI

The guidance contained in Chapter C of this document is building on the existing regulatory framework, while ensuring that the challenges introduced in Section B.1 are addressed through the new concepts that are highlighted in this Section B.6.

It is important to keep in mind that current AI and ML technology may not be at a level commensurate with the perspective opened by some of the objectives that are introduced in the document. The proposed guidance aims at paving the way for the future deployment of AI/ML as anticipated by the aviation industry; however, considering all necessary limitations (e.g. on level of criticality of the AI-based system) to further ensure an adequate level of safety of innovative solutions.

6.1. Learning assurance

In the current regulatory framework, the associated risk-based approach for systems, equipment and parts is mainly driven by a requirements-based 'development assurance' methodology during the development of their constituents. Although the system-level assurance still requires a requirements-based approach, it is admitted that the design-level layers that rely on learning processes cannot be addressed with 'development assurance' methods.

Intuitively, the assurance process should be shifted on the correctness and completeness/representativeness of the data (training/validation/test data sets) and on the learning and its verification. Most importantly, the main challenge lies in providing guarantees that the training performed on sampled data sets can generalise with an adequate performance on unseen operational data.

To this purpose, a new concept of 'learning assurance' is proposed to provide novel means of compliance. The objective is to gain confidence at an appropriate level that an ML application supports the intended functionality, thus opening the 'AI black box' as much as practicable.

Definition

Learning assurance: *All of those planned and systematic actions used to substantiate, at an adequate level of confidence, that errors in a data-driven learning process have been identified and corrected such that the system satisfies the applicable requirements at a specified level of performance, and provides sufficient generalisation and robustness guarantees.*

6.2. AI explainability

AI explainability — overview

Explainability is a key property that any safety-related AI-based system should possess. It was the reason for including it as a dedicated building block in the first release of the EASA AI Roadmap. The preparation of the first EASA concept paper for level 1 AI applications has allowed further refinement of the explainability concept in two views: one pertaining to the end users (operational explainability) and one pertaining to other stakeholders involved with the AI-based system at the development time or in the post-operational phase (development explainability). As mentioned previously, the development of specific objectives for level 2 AI has crystallised the need for extension of the AI explainability building block to cover a wider range of human factors guidance aspects. It has also helped to further refine the allocation of the two explainability views, bringing the development explainability closer to the learning assurance within the renamed AI assurance building block, and leaving the operational explainability as the first essential element of the extended human factors for AI.

AI explainability — definition

While industry works on developing more advanced computers which include decision-making capabilities, questions arise as to how the end user will understand and interpret the results and reasoning of AI-based systems. In other words, to use the system, the user should be in a position to understand and trust it. The development of advanced and complex AI techniques, for example, deep neural networks (DNNs), may lead to major transparency issues for the end user.

This guidance makes a clear distinction between two types of explainability driven by the profile of the user and their needs:

— The information required to make a ML model understandable; and

— Understandable information for the user on how the system came to its results.

The target audience of the explanation drives the need for explainability. In particular, the level of explainability is highly dependent on the expertise and domain of the user. Details on the intrinsic functioning of an ML model could be very useful, for example, to a developer but not understandable by an end user.

In the aviation domain, a number of stakeholders require explanations about the AI-based system behaviour: the certification authority, the safety investigator, the engineers (developer or maintainer) and the end user. Similarly, for each target audience, the qualities of the explainability will also be affected. The nature of the explanations needed are influenced by different dimensions, such as the time to get the explanation, which would depend on the stakeholders.

This guidance defines explainability as:

> Definition
>
> ***AI explainability:*** *Capability to provide the human with understandable, reliable, and relevant information with the appropriate level of details and with appropriate timing on how an AI/ML application produces its results.*

This definition might evolve over time as the AI research evolves.

Note: In this document, whereas 'explainability' refers to the capability, 'explanation' refers to the information as an instantiation of the explainability.

AI explainability — motivations

There are three groups of roles that drive the scope and need for explainability:

— Those involved in developing AI applications: systems and software engineers, data scientists, etc.;

— Those involved in working operationally with AI applications: flight crew, air traffic controllers (ATCOs), etc.;

— Those involved in analysing what an AI application has done during operations: maintenance staff, safety investigators, etc.

DEEL's white paper (DEEL Certification Workgroup, 2021) explores the need for explainability based on the categories of users/consumers.

The list of motivations shows that they are generally shared between the stakeholders involved in the development and post-operational phases. Both development and post-operational users are all interested in a very detailed level of transparency on the inner function of the AI-based system. This contrasts with the motivations of the end users who are looking for explanations that are appropriate to the operations.

The table below summarises the motivations of each group:

Development & Post-operation	Operation
▪ Develop system trustworthiness ▪ Establish causal relationships between the input and the output of the model ▪ Catch the boundaries of the model and help in its fixing ▪ Highlight undesirable bias (data sets and model bias) ▪ Allow the relevant receivers to identify errors in the model ▪ Support continuous analysis of the AI-based system behaviour ▪ **Support the safety investigation of accidents and incidents where an AI-based system was involved**	▪ Contribute to building trust for the end user ▪ Contribute to predicting AI behaviour ▪ Contribute to understanding actions/decisions

Table 1 — Needs for AI explainability

Given the above split, the remainder of this document establishes the requirement for explainability from two perspectives:

— Development & post-ops explainability (Section C.3.2);

— Operational explainability (Section C.4.1).

6.3. Operational domain (OD) and operational design domain (ODD)

As already depicted in the previous sections with the introduction of learning assurance, special attention needs to be paid to the data that will be used by the ML models, either during the training phase or when the AI-based system with its ML model infers in the operations.

In the context of ML, an OD at system level, and an ODD at AI/ML constituent level needs to be defined in order to provide constraints and requirements on the data that will be used during the learning process, the implementation, or even during inference in the operations.

Section G.1 proposes definitions of OD and ODD where the ODD at AI/ML constituent level constitutes a refinement of the operating conditions of the OD at the AI-based system level.

— Note on the definition of OD: This one has been adapted from the ODD definition used in the automotive industry which is based on the document SAE J3016 (Level of driving automation, 2021). The capture of operating conditions is already a practice in the aviation domain, which corresponds to the conditions under which a given AI-based system is specifically designed to function as intended; however, this process is not as formal as required to deal with AI-based systems. Therefore, the formalisation of this notion under the term OD.

— Note on the definition of ODD: In addition, the level of details captured at system level is not commensurate with the level of details typically needed at AI/ML constituent level to serve the purpose of the ML model design processes, in particular the data and learning management steps. This is the reason why the additional notion of AI/ML constituent ODD is introduced.

The ODD provides a framework for the selection, collection, preparation of the data during the learning phase, as well as the monitoring of the data in operations.

Special considerations are made with respect to the OD and ODD:

— Definition of the OD (Section C.2.1.2)

— Definition of the ODD (Section C.3.1.3.1).

6.4. Human-AI Teaming

The new concept of HAT refers to the cooperation and collaboration between the end user and the AI-based system to achieve goals. This HAT concept, depending on the maturity of the AI-based system, involves a shared understanding of goals, roles and processes (decision-making/problem solving) between the members. It also implies the development of trust and an effective interaction. With this evolution of the AI-based system towards Level 2 AI applications, there is a growing need for guidance on how to effectively introduce and use this concept of HAT.

To this purpose, the guidance makes a clear distinction between the notions of cooperation and collaboration to clarify the definition of the AI levelling as well as to provide novel means of compliance (C.4.2):

— ***Human-AI cooperation*** (Level 2A AI): cooperation is a process in which the AI-based system works to help the end user accomplish his or her own goal.

The AI-based system works according to a predefined task allocation pattern with informative feedback to the end user on the decisions and/or actions implementation. The cooperation process follows a directive approach. Cooperation does not imply a shared situational awareness between the end user and the AI-based system. Communication is not a paramount capability for cooperation.

— ***Human-AI collaboration*** (Level 2B AI): collaboration is a process in which the human end user and the AI-based system work together and jointly to achieve a common goal (or work individually on a defined goal) and solve a problem through co-constructive approach. Collaboration implies the capability to share situational awareness and to readjust strategies and task allocation in real time. Communication is paramount to share valuable information needed to achieve the goal, to share ideas and expectations.

The expected AI-based system capabilities for cooperation and collaboration processes are different, as they are designed to achieve different goals requiring different kind of interactions.

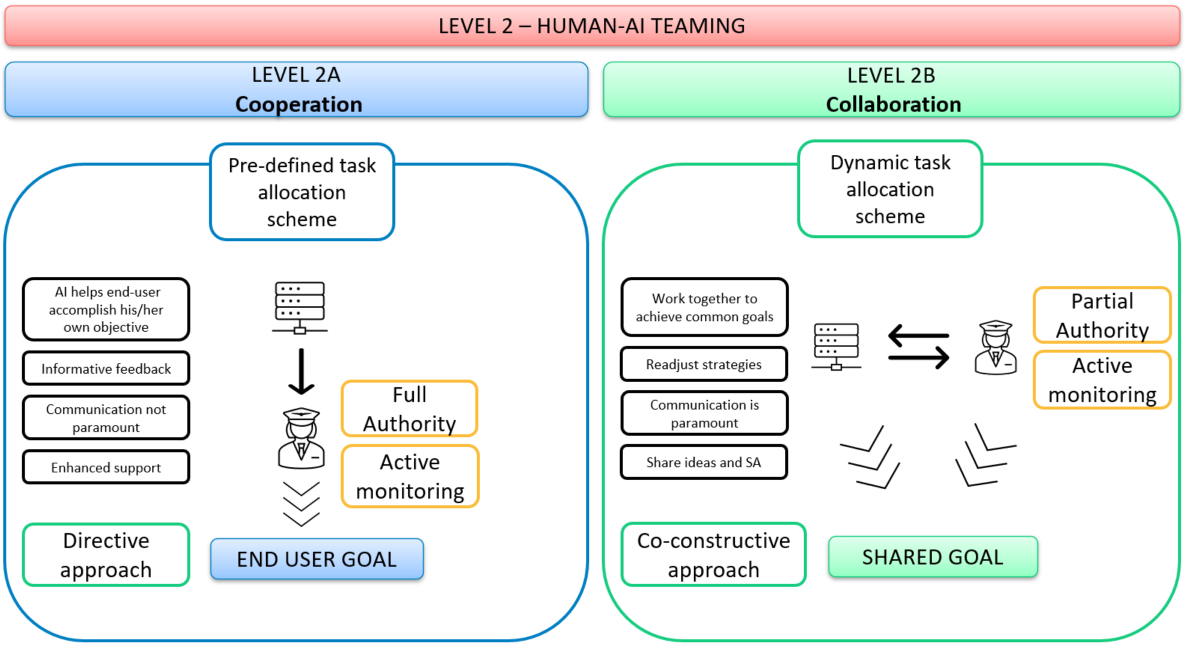


Figure 6 — HAT concept overview

In this guidance, it is anticipated that for the AI-based systems to participate effectively in the HAT, certain capabilities are needed such as the notion of communication (more specifically for Level 2B), adaptability/flexibility, situational awareness and transparency.

The human factors guidance developed in Section C.4.2 provides detailed objectives and means of compliance to the applicants to design an AI platform where the above capabilities originate.

Finally, the guidance should help reinforce that the AI-based system and its platform is designed to:

— take into account the needs and capabilities of the end user by following a human-centred design approach;

— foster cooperation, collaboration and trust between the end user and the AI-based system by ensuring clear interaction; and

— meet existing human factors/ergonomics requirements and guidance including those related to design, usability knowledge and techniques.

C. AI trustworthiness guidelines

1. Purpose and applicability

This chapter introduces a first set of objectives, in order to anticipate future EASA guidance and/or requirements to be complied with by safety-related ML applications. Where practicable, a first set of anticipated MOC has also been developed, in order to illustrate the nature and expectations behind the objectives.

The aim is to provide applicants with a first framework to orient choices in the development strategy for ML solutions. This first set of usable objectives does not however constitute either definitive or detailed means of compliance.

These guidelines apply to any system incorporating one or more ML models (further referred to as AI-based system), and are intended for use in safety-related applications or for applications related to environmental protection covered by the Basic Regulation, in particular for the following domains:

— ***Initial and continuing airworthiness***, applying to systems or equipment required for type certification or by operating rules, or whose improper functioning would reduce safety (systems or equipment contributing to failure conditions Catastrophic, Hazardous, Major or Minor);

— ***Air operations***, applying to systems, equipment or functions intended to support, complement, or replace tasks performed by aircrew or other operations personnel (examples may be information acquisition, information analysis, decision-making, action implementation and monitoring of outputs);

— ***ATM/ANS***[5], applying to equipment intended to support, complement or replace end-user tasks (examples may be information acquisition, information analysis, decision-making and action implementation) delivering ATS or non-ATS;

— ***Maintenance***, applying to systems supporting scheduling and performance of tasks intended to timely detect or prevent unsafe conditions (airworthiness limitation section (ALS) inspections, certification maintenance requirements (CMRs), safety category tasks) or tasks which could create unsafe conditions if improperly performed ('critical maintenance tasks');

— ***Training***, applying to systems used for monitoring the training efficiency or for supporting the organisational management system, in terms of both compliance and safety;

[5] For the ATM/ANS domain, according to the currently applicable Regulation (EU) 2017/373, there is no separate approval for the ATM/ANS equipment, and all the activities related to the changes to the functional system (hardware, software, procedures and personnel) are managed under the change management procedures, as part of the air navigation service provider change management process. Competent authority approval is obtained for the introduced complete change. Furthermore, in this Regulation, only the air traffic service (ATS) providers are requested to perform a safety assessment as part of the change management process whereas the non-ATS providers (e.g. CNS) are requested to perform a safety support assessment, intended to assess and demonstrate that after the introduction of the change the associated services will behave as specified and will continue to behave as specified. Finally the regulatory framework will soon be amended with the introduction of new implementing and delegated regulations identified in Opinion No 01/2023 on conformity assessment.

— ***Aerodromes***, applying to systems that automate key aspects of aerodrome operational services, such as the identification of foreign object debris, the monitoring of bird activities, and the detection of UAS around/at the aerodrome;

— ***Environmental protection***, applying to systems or equipment affecting the environmental characteristics of products. Note: While the use of AI/ML applications in such systems or equipment may not be safety-critical, the present guidance may still be relevant to establish the necessary level of confidence in the outputs of the applications.

The introduction of AI/ML in these different aviation domains may thus imply (or 'require') as well adaptations in the respective organisational rules per domain (such as for design organisation approval (DOA) holders, maintenance organisation approval (MOA) holders, continuing airworthiness management organisations (CAMOs), air navigation service providers (ANSPs), approved training organisations (ATOs), air operators, etc.). Each organisation would need to ensure compliance with EU regulations (e.g. for initial airworthiness, continuing airworthiness, air operations, ATM/ANS, occurrence reporting, etc.) as applicable to each domain. Furthermore, each organisation would need to assess the impact on its internal processes in areas such as competence management, design methodologies, change management, supplier management, occurrence reporting, information security aspects or record-keeping.

The applicability of these guidelines is limited as follows:

— covering ***Level 1 and Level 2 AI applications***, but not covering yet ***Level 3 AI applications***;

— covering ***supervised learning***, but not other types of learning such as ***unsupervised or reinforcement learning***;

— covering ***offline learning*** processes where the model is 'frozen' at the time of approval, but not ***adaptive or online learning*** processes.

2. Trustworthiness analysis

2.1. Characterisation of the AI application

2.1.1. High-level task(s) and AI-based system definition

In this section all objectives require to consider the system as a whole, as opposed to considering its subsystems or AI/ML constituents.

When characterising an AI-based system, the first step for an applicant consists in identifying the list of end users intended to interact with the AI-based system, the associated high-level tasks and the AI-based system definition.

Objective CO-01: The applicant should identify the list of end users that are intended to interact with the AI-based system, together with their roles, their responsibilities and their expected expertise (including assumptions made on the level of training, qualification and skills).

Objective CO-02: For each end user, the applicant should identify which high-level tasks are intended to be performed in interaction with the AI-based system.

Anticipated MOC CO-02: The level at which the high-level tasks are identified should be considered at the level of the interaction between the human and the AI-based system, not at the level of each single function performed by the AI-based subsystem or AI/ML constituent. The list of high-level task(s) relevant to the end user(s), in interaction with the AI-based system, should be documented.

Objective CO-03: The applicant should determine the AI-based system taking into account domain-specific definitions of 'system'.

Anticipated MOC CO-03: When relevant, the system should be decomposed into subsystems, one or several of them being an AI-based subsystem(s).

The definition of system varies between domains. For example:

— for airborne systems, ARP4761 defines a system as 'combination of inter-related items arranged to perform a specific function(s)';

— for the ATM/ANS domain (ATS and non-ATS), Regulation (EU) 2017/373 defines a functional system as 'a combination of procedures, human resources and equipment, including hardware and software, organised to perform a function within the context of ATM/ANS and other ATM network functions'.

In a second step, once the AI-based system has been determined, two separate but correlated activities should be executed:

— Definition of the concept of operations (ConOps), with a focus on the identified end users and the task allocation pattern between the end user(s) and the AI-based system (see Section C.2.1.2); and

— A functional analysis of the AI-based system (see Section C.2.1.3).

These activities will provide the necessary inputs for the classification of the AI application, for safety, security, and ethical assessment, as well as for the other building blocks of the AI trustworthiness framework.

2.1.2. Concept of operations for the AI application

To support compliance with the objectives of the AI trustworthiness guidelines, a detailed ConOps describing precisely how the system will be operated is expected to be established.

Objective CO-04: The applicant should define and document the ConOps for the AI-based system, including the task allocation pattern between the end user(s) and the AI-based system. A focus should be put on the definition of the OD and on the capture of specific operational limitations and assumptions.

Anticipated MOC-CO-04: The ConOps should be described at the level of the AI-based system, where the human is expected to achieve a set of high-level tasks.

The ConOps should consider:

— the list of potential end users identified per Objective CO-01;

— the list of high-level tasks for each end user per Objective CO-02;

— an end-user-centric description of the operational scenarios (with sufficient coverage of the high-level tasks);

— a description of the task allocation pattern between the end user(s) and the AI-based system, further dividing the high-level tasks identified under Objective CO-02 in as many sub-tasks as necessary; a scenario is: in a given context/environment, a sequence of actions in response to a triggering event that aims at fulfilling a (high-level) task;

— a description of how the end users will interact with the AI-based system, driven by the task allocation pattern;

— the definition of the OD, including the specific operating limitations and conditions appropriate to the proposed operation(s);

— some already identified risks, associated mitigations, limitations and conditions on the AI-based system.

Figure 7 shows the interrelationship between the operational scenarios for the ConOps and the operating parameters for the OD:

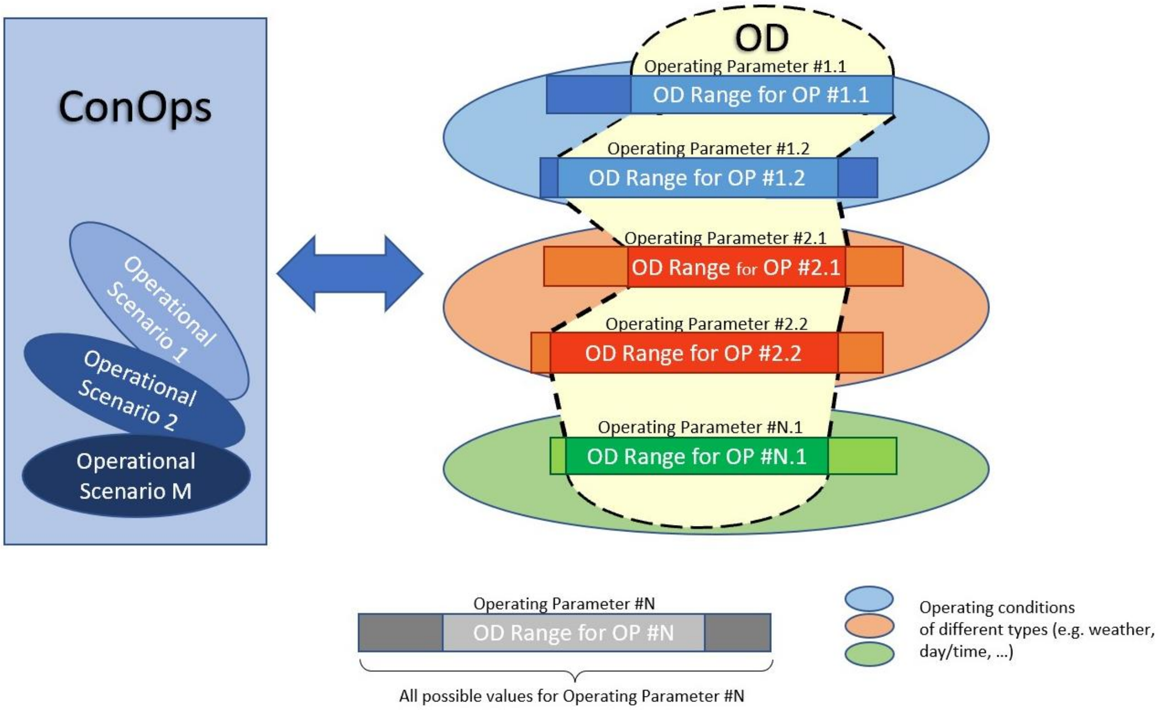


Figure 7 — Interrelationship between ConOps and OD

Notes:

— The OD is further refined during the learning process. This refinement is materialised via the definition of an ODD at AI/ML constituent level (see Section C.3.1.3.1);

— The OD also considers dependencies between operating parameters in order to define correlated ranges between some parameters when appropriate; in other words, the range(s) for one or several operating parameters could depend on the value or range of another parameter;

— ConOps limitations may be accounted for in activities related to the safety assessment or safety support assessment, as described in Sections C.2.2.2.1 and C.2.2.2.2;

— Due to the data-driven nature of ML applications, the precise definition of the ConOps is an essential element to ensure that sufficient and representative data is collected for the data sets that are used for training, validation and testing purposes.

Objective CO-05: The applicant should document how end users' inputs are collected and accounted for in the development of the AI-based system.

Anticipated MOC-CO-05: The applicant should engage the end user in planning, design, validation, verification and certification/approval of an AI based system. The end users' involvement should be documented.

2.1.3. Functional analysis of the AI-based system

Objective CO-06: The applicant should perform a functional analysis of the system.

The functional analysis consists in identifying a break-down of the high-level function(s) into sub-function(s), allocating the sub-function(s) to the subsystem(s), AI/ML constituents and items in line with the architecture choices. The delineation between AI/ML item and non-AI/ML item is performed at this stage: at least one item is allocated with AI function(s) and is thus considered an AI/ML item.

Notes:

— The functional analysis is an enabler to meet the objectives in Section C.3.1.2 'Requirements and architecture management' of the learning assurance.

— The functional analysis is a means supporting the functional hazard assessment (FHA) as per Section C.2.2.3 'Initial safety (support) assessment'.

2.1.4. Classification of the AI application

This first usable guidance document focuses on ***Level 1 AI applications***. It therefore provides classification guidelines for this level, including boundaries between Levels 1A, 1B and 2, in order to avoid confusion of the applicants on the classification of their proposed AI-based system.

To this purpose, EASA is taking advantage of the seminal 'A model for Types of Human Interaction with Automation' research paper (Parasuraman-et-al, 2000). According to the authors, the four-stage model of human information processing has its equivalent in system functions that can be automated. The authors propose that automation can be applied to four classes of functions:

— ***Information acquisition*** involves sensing and registration of input data; these operations are equivalent to the first human information processing stage, supporting human sensory processes.

— ***Information analysis*** involves cognitive functions such as working memory and inferential process.

— ***Decision-making*** involves selection from among decision alternatives.

— ***Action implementation*** refers to the actual execution of the action choice.

The research paper foresees several levels of automation (from Low to High) for each function. In early publications, the HARVIS research project (Javier Nuñez et al., 2019) made use of this scheme to develop a Level of Automation (LOAT), further splitting this scheme by distinguishing between an action performed to 'automation support' the human versus an action performed 'automatically' by the system.

To further refine this scheme, when considering the anticipated distinction between the ***Level 2 AI*** and ***Level 3 AI*** applications, a further decomposition is introduced for 'automatic' functions into '***overseen and overridable***', '***supervised***' or '***non-supervised***' by the human.

— ***Overseen and overridable***: *capability of the human end user to closely monitor the tasks allocated to the AI-based system, with the ability to intervene in every decision-making and/or action implementation of the AI-based system.*

— ***Supervised***: *capability of the human end user to supervise the operations of the AI-based system (some decisions and some action implementation), with the ability to override the authority of the AI-based system (some decisions and some action implementation) when it is necessary to ensure safety and security of the operations (e.g. upon alerting).*

— ***Non-supervised***: *no human end user is involved in the operations and therefore there is no capability to override the AI-based system's operations.*

Note: It is important to remind that the ***AI levels*** introduced in ***EASA AI Roadmap 1.0*** are not meant to be an automation scheme but a classification of AI-based systems with regard to their usage and interaction with the human end users. Therefore, the detailed levels introduced in the Raja Parasuraman paper (Low-High) or in the HARVIS deliverables (Levels A0 to D8) were not considered in this document.

The development of Level 2 guidance has determined the need for a further split into two levels, 2A and 2B, based on the notion of ***authority***. For the purpose of this document, the notion of distribution of authority between an AI-based system and an end user refers to the control and decision-making that each member has in their interactions with one another. In this context, authority can be seen as the ability to make decisions and take actions without the need for approval from the other member. In the context of the AI-based systems, authority can be divided between the human and the AI-based system in a variety of ways, ranging from full authority for the end-user (up to Level 2A AI), partial authority for the end user (Level 2B AI), up to full authority for the AI-based system (Level 3 AI).

In some cases, the authority of the AI-based system may be limited to specific tasks and may vary depending on the pre-definition of the expected automatic decision and action implementation from the AI-based system. In other cases, the AI-based system makes decisions and take actions without end-user monitoring and control. To support this classification, the following additional definitions are needed:

— ***Full authority for the end user***: *refers to a situation where the full control is kept on the end-user side. The end user performs an active monitoring with the ability to intervene and override any decisions taken and/or actions made by the AI-based system. The AI-based system does not have the capabilities to perform any oversight of the end-user activities.*

— ***Partial authority for the end user***: *refers to a situation where the end user keeps an active monitoring and some degree of control over the AI-based system. The AI-based system may therefore be designed to make decisions and take specific actions, but the human can still override these decisions/actions if needed. This authority distribution scheme allows for a more collaborative relationship between the AI-based system and the end user.*

— ***Authority for the end user upon alerting***: *refers to a situation where the full control is left to the AI-based system to make decisions and take actions, without active monitoring/oversight from the end user. However, a passive monitoring is foreseen to allow the end user to revert to 'full' or 'partial' authority depending on events occurring in the operations.*

The resulting classification scheme is as follows and provides a reference for the classification of the AI-based system. In case of doubt, the applicant should assume the higher AI level.

AI level	Function allocated to the system to contribute to the high-level task	Authority of the end user
Level 1A Human augmentation	Automation support to information acquisition	Full
	Automation support to information analysis	Full
Level 1B Human assistance	Automation support to decision-making	Full
Level 2A Human-AI cooperation	Overseen and overridable automatic decision	Full
	Overseen and overridable automatic action implementation	Full
Level 2B Human-AI collaboration	Overseen and overridable automatic decision	Partial
	Overseen and overridable automatic action implementation	Partial
Level 3A Supervised advanced automation	Supervised automatic decision	Upon alerting
	Supervised automatic action implementation	Upon alerting
Level 3B Autonomous AI	Non-supervised automatic decision	Not applicable
	Non-supervised automatic action implementation	Not applicable

Table 2 — EASA AI levels

Objective CL-01: The applicant should classify the AI-based system, based on the levels presented in Table 2, with adequate justifications.

Anticipated MOC-CL-01-1: When classifying the AI-based system, the following aspects should be considered:

— Only the AI-based system incorporating one or more ML models is to be classified following the classification scheme proposed in Table 1.

— When classifying, the applicant should consider the high-level task(s) that are allocated to the end user(s), in interaction with the AI-based system, as identified per **Objective CO-02**. It is important to avoid slicing the system into granular lower-level functions when performing the classification, as this may lead to over-classifying the AI level, on the basis of some functions that the human end user is not supposed to oversee or supervise. The classification should also exclude the tasks that are performed solely by the human, as well as the ones allocated to other (sub)systems not based on ML technology.

— When several 'AI levels' apply to the AI-based system (either because it has several constituents or is involved in several functions/tasks), the resulting 'AI level' is the highest level met by the AI-based system considering its full capability.

Note: An illustration of this classification mechanism is available in Table 6 — Classification applied to use cases, where the 'AI level' is determined by the highest AI level in the blue bounding box.

As a consequence, for a given AI-based system, the result of the classification is a static 'AI level'. This 'AI level' is an input to the development process and contributes to the modulation of the objectives in this document that apply to this system.

Note: This is the point where the 'AI level' classification scheme differs from an 'automation' scheme, as with the latter, the classification can dynamically evolve in operations, considering different phases of the operation or degraded modes for instance. On the contrary, the 'AI level' is static and reflects the highest capability offered by the AI-based system, in terms of interaction with the human end user or in terms of autonomy (when it comes to AI level 3B). The purpose of this classification is merely to provide a generic and consistent reference to all aviation domains, this classification being another important dimension to drive the modulation of AI trustworthiness objectives (see Chapter D) beyond the one linked to the criticality of the AI-based system.

Anticipated MOC-CL-01-2: The following considerations support the delineation of boundaries between 'AI levels'.

The boundary between level 1A and level 1B is based on the notion of decision-making. 1A covers the use of AI/ML for any augmentation of the information presented to the end user, ranging from organisation of incoming information according to some criteria to prediction (interpolation or extrapolation) or integration of the information for the purpose of augmenting human end-user perception and cognition. 1B addresses the step of support to decision-making, therefore the process by the human end user of selection of a course of actions among several possible alternative options. The number of alternatives could be multiple and in some cases the AI-based system could present only a subset of all possible alternatives, which would still be considered as AI level 1B. The number of alternatives could also be limited to two (e.g. validating a radio-frequency suggestion or amending the entry proposed by the AI-based system and this still consists in decision-making. On the contrary, the implementation of an action or series of actions by the human end user to perform a predefined task (such as following a predefined route or landing the aircraft) is not considered as decision-making. Finally, the notion of support implies that the decision is solely taken by the human end user and not by the AI-based system.

The boundary between level 1B and level 2A is based on the distinction between support to decision-making and automatic decision-making (e.g. proceeding with the landing when reaching decision height or going around). At level 2A it is important to remind that such automatic decisions are fully overseen and overridable by the human end user (e.g. the pilot could decide to go around despite the decision from the AI-based system to proceed with an autoland). Level 2A also addresses the automatic implementation of a course of actions by the AI-based system even when the decision is taken by the human end user (e.g. assistant supporting automatic approach configuration before landing).

While both levels 2A and 2B imply the capability of the AI-based system to undertake automatic decisions or actions implementation, the boundary between those two levels lies in the capability of level 2B AI-based systems to take over some authority on decision-making, to share situational awareness and to readjust strategies and task allocation in real time (e.g. virtual co-pilot in a single-pilot operation aircraft. The pilot and the virtual co-pilot share tasks and have a common set of goals under a collaboration scheme. The virtual co-pilot has the capability to use natural language for communication allowing an efficient bilateral communication between both members to readjust strategies and decisions).

The boundary between level 2B and level 3A lies in the level of oversight that is performed by the human end user on the operations of the AI-based system (e.g. a pilot in the cockpit). A strong prerequisite for level 2 (both for 2A and 2B) is the ability for the human end user to intervene in every decision-making and/or action implementation of the AI-based system, whereas in level 3A applications, the ability of the end user to override the authority of the AI-based system is limited to cases where it is necessary to ensure safety of the operations (e.g. an operator supervising a fleet of UAS, terminating the operation of one given UAS upon alerting).

The boundary between level 3A and 3B will be refined when developing the level 3 AI guidelines. It is for the time being solely driven by consideration of the presence or absence of capability for a human end user to override the operations of the AI-based system, therefore on the level of autonomy of the product embedding the AI-based system.

2.2. Safety assessment of ML applications

2.2.1. AI safety assessment concept

2.2.1.1. Statement of issue

The objective of a ***safety assessment*** is to demonstrate an acceptable level of safety as defined in the applicable regulations. A logical and acceptable inverse relationship must exist between the occurrence probability of a failure condition and the severity of its effect. Depending on the domain of applications in aviation, ***safety assessment*** methodologies may vary, but a common point is the consideration that only hardware components are subject to a random failure. The reliability of a given piece of software is not quantified per se. As an example, for airborne systems, it is usually considered that when recognised ***development assurance*** methodologies are used throughout the development, the risk of having an error resulting in a failure is minimised to an adequate level of confidence. Development errors are considered as a possible common source type and are mitigated by system architecture and analysed with other common mode errors and failures via dedicated techniques such as common mode analysis. The probabilistic risk assessment then usually limits the contribution of digital components to the reliability of the digital function input parameters and to the reliability of the hardware platform executing the digital code.

Due to their statistical nature and to model complexity, ML applications come with new limitations in terms of predictability and sources of uncertainties. Taking this into consideration, this guidance is intended to assist applicants in demonstrating that systems embedding AI/ML constituents operate at least as safely as traditional systems developed using existing **development *assurance*** processes and ***safety assessment*** methodologies[6]: AI technology introduction should be done at no higher risk imposed to persons, personal properties (or critical infrastructure). Furthermore, the proposed guidance is also aimed at following as closely as possible existing aviation ***safety assessment*** processes to minimise the impact on those processes.

It is acknowledged by EASA that facing uncertainty on safety-critical applications is not a challenge unique to AI/ML applications.

For embedded traditional systems, existing guidance material already recognises, for instance, that, for various reasons, component failure rate data is not precise enough to enable accurate estimates of the probabilities of failure conditions (see for example AMC 25.1309 11.e.4). This results in some degree of uncertainty. Typically, when calculating the estimated probability of a given hazard, applicable guidance, such as AMC 25.1309, requires that this uncertainty should be accounted for in a way that does not compromise safety. The need for such a conservative approach to deal with uncertainty is unchanged with AI/ML applications.

For the ATM/ANS domain, the safety assessment to be performed by ATS providers also needs to account for uncertainties during the risk evaluation step. AMC1 ATS.OR.205(b)(4) of Regulation (EU) 2017/373 requests that risk evaluation includes a comparison of the risk analysis results against the safety criteria taking the uncertainty of the risk assessment into account.

Furthermore, AI/ML applications may be able to estimate uncertainties associated with their outputs. These estimations may then feed monitoring functions which in turn contribute to the safety case or provide valuable data for the continuous safety assessment (see Section C.2.2.4).

2.2.1.2. Safety assessment concept

An adequate safety level should be achieved and maintained throughout the whole product life cycle, thanks to:

— initial safety assessment, during design phase by considering the contribution of an AI/ML constituent to system failure and by having particular architectural considerations when AI is introduced; followed by

— continuous safety assessment, with the implementation of a data-driven AI safety risk assessment based on operational data and occurrences. This 'continuous' analysis of in-service events may rely on processes already existing for domains considered in this guideline. The processes will need to be adapted to the AI introduction.

It is recognised that, depending on the domains, the necessary activities to be performed and documented in view of EASA approval vary significantly. The table below summarises per domain the expected analysis to be performed in view of the approval by EASA of a system embedding an AI/ML constituent.

[6] In the ATM/ANS domain, for non-ATS providers, the safety assessment is replaced by a safety support assessment.

Aviation domains	'Initial' safety assessment	'Continuous' safety assessment
Initial and continuing airworthiness	As per Section C.2.2.2 below	As per organisation section (DOA – Continuing airworthiness) and Section C.2.2.4'Instruction for continuous safety assessment' of this guideline
Air operations	To be defined	As per organisation section (Operators – SMS) and Section C.2.2.4 'Instruction for continuous safety assessment' of this guideline
ATM/ANS	As per Section C.2.2.2.1 for ATS providers and Section C.2.2.2.2 for non-ATS providers – see note B	As per organisation section (ANSPs – SMS) and Section C.2.2.4 'Instruction for continuous safety assessment' of this guideline – see Note G
Maintenance	None – see Notes A and D	As per organisation section (CAMO or MOA – SMS) and Section C.2.2.4 'Instruction for continuous safety assessment' of this guideline
Training	None – see Notes A and E	Managed from an organisation, operations and negative training, as per organisation section (ATO – SMS) and Section C.2.2.4 'Instruction for continuous safety assessment' of this guideline
Aerodromes	To be defined	To be defined
Environmental protection	None – see Note F	Currently not applicable

Table 3 — Safety assessment concept for the major aviation domains

Note A: For some domains, only a 'continuous' safety assessment is expected to be presented to EASA at this stage. Applicants may however use the methodology described in Section C.2.2.2 to establish their design processes.

Note B: Regulation (EU) 2017/373 that addresses ATS and non-ATS providers has introduced the need of a 'safety support assessment' for non-ATS providers rather than a 'safety assessment'. The objective of the safety support assessment is to demonstrate that, after the implementation of the change, the functional system will behave as specified and will continue to behave only as specified in the specified context. For these reasons, a dedicated Section C.2.2.2.2 has been created for non-ATS providers.

Note C: The terminology used in safety assessment/safety support assessment between the various domains varies. Footnotes have been used in the next paragraph to clarify the guideline intent with regard to domain specificities and domain-specific definitions reminded in Chapter G.

Note D: For the maintenance domain, whenever new equipment is used, it should be qualified and calibrated.

Note E: For the training domain, whenever an AI-based system is adopted, the entry into service period should foresee an overlapping time to enable validation of safe and appropriate performance.

Note F: For the environmental protection domain, the initial safety assessment is to be interpreted as the demonstration of compliance with the applicable environmental protection requirements.

Note G: For ATS and non-ATS providers, the notion of 'continuous safety assessment' should be understood as the 'Safety performance monitoring and measurement' for ATS providers, or simply the 'Performance monitoring and measurement' for non-ATS providers.

2.2.2. Impact assessment of AI introduction

In the following sections, the steps highlighted in **bold** are novel or affected by AI introduction compared to a classical safety assessment.

In Section C.2.2.2.1, safety assessment should be understood as safety assessment of the functional system when it applies to ATS providers in the ATM/ANS domain, and should be understood as a system safety assessment in the airworthiness domain. Safety support assessment of the functional system applies to non-ATS providers and is addressed in Section C.2.2.2.2.

2.2.2.1. Impact on safety assessment methodologies

The analyses below describe the typical safety assessment activities performed throughout the design phase.

— Perform functional hazard assessment in the context of the ConOps

— Safety assessment activities supporting design and validation phases

- Define safety objectives[7], proportionate with the hazard classification
- Define a preliminary system architecture to meet the safety objectives
- **Allocate assurance level (e.g. DAL or SWAL)**
- **Define AI/ML constituent performance[8] metrics**
- **Analyse and mitigate the effect of the AI-based (sub)system (respectively AI/ML constituent) exposure to input data outside of the AI-based (sub)system OD (respectively AI/ML constituent ODD)[9]**
- **Identify and classify sources of uncertainties. Analyse and mitigate their effects.**
- **Perform AI/ML item failure mode effect analysis (see note)**
- Derive safety requirements including independence requirements to meet the safety objective and support the architecture

7 In the ATM/ANS domain, for ATS providers, this activity corresponds to the definition of safety criteria.

8 The set of selected metrics should allow the estimation of the reliability of the AI/ML constituent: empirical probabilities of each failure mode relevant for the safety assessment should be obtained from selected metrics.

9 The AI-based (sub)system OD is described according to Objective CO-04. The AI/ML constituent ODD is described according to Objective DM-01.

- Define and validate assumptions

— **Verification phase**

- **Perform final safety assessment**
- **Consolidate the safety assessment to verify that the implementation satisfies the safety objectives[10].**

Note:
An AI/ML item failure mode effect analysis should be performed to identify the various failure mode and associated failure/error rates for the ML model. When a quantitative safety assessment is required, the outputs of this analysis is then fed into the safety assessment quantitative analysis (e.g. fault tree analysis, dependence diagram) to demonstrate that the quantitative safety objectives are fulfilled.

2.2.2.2. Impact on safety support assessment

The analyses below describe the typical safety support assessment activities performed during the design phase. The steps highlighted in **bold** are expected to be affected by AI introduction compared to the usual process:

— Evaluate impact on the service specification, **including service performance**

— Identify applicable service performance requirements

— Define a preliminary system architecture

— Analyse design:

- **Perform AI/ML item failure mode effect analysis**
- **Define the AI/ML constituent performance[11] metrics**
- **Analyse and mitigate the effect of the AI-based (sub)system (respectively AI/ML constituent) exposure to input data outside of the AI-based (sub)system OD (respectively AI/ML constituent ODD)[12]**
- **Identify and classify sources of uncertainties**
- **Analyse and mitigate their effects**
- **Allocate assurance level (e.g. SWAL)**

— Define safety support requirements

— **Verify that the implementation satisfies the safety support requirements**

[10] In the ATM/ANS domain, for ATS providers, these correspond to the safety criteria.

[11] The 'AI/ML Constituent performance' is a possible contributor to service performance that is defined in Regulation (EU) 2017/373: 'performance of the service refers to such properties of the service provided such as accuracy, reliability, integrity, availability, timeliness, etc.'

[12] The AI-based (sub)system OD is described according to Objective CO-04. The AI/ML constituent ODD is described according to Objective DM-01.

2.2.3. Initial safety (support) assessment

Based on the high-level impact assessment performed in C.2.2.2.1 and C.2.2.2.2, the following objective is proposed for the initial safety assessment:

Objective SA-01: The applicant should perform a safety (support) assessment for all AI-based (sub)systems, identifying and addressing specificities introduced by AI/ML usage.

The following means of compliance are proposed to address AI/ML-specific activities to be performed during the initial safety assessment:

Anticipated MOC-SA-01-1: DAL/SWAL allocation and verification**:**

The following standards and implementing rules with adaptation may be used to perform DAL/SWAL allocation

— For embedded systems:
  - ED79A/ARP4754A and ARP4761

— For ATS providers in the ATM/ANS domain, the following implementing rule requirements (and the associated AMC and GM) are applicable:
  - ATS.OR.205 Safety assessment and assurance of changes to the functional system
  - ATS.OR.210 Safety criteria

— For non-ATS providers in the ATM/ANS domain, the following implementing rule requirements (and the associated AMC and GM) are applicable:
  - ATM/ANS.OR.C.005 Safety support assessment and assurance of changes to the functional system.

The following limitations are applicable when performing the DAL/SWAL allocation :

With the current state of knowledge of AI and ML technology, EASA anticipates a limitation on the validity of applications when AI/ML constituents include IDAL A or B / SWAL 1 or 2 / AL 1, 2 or 3 items.

Moreover, no assurance level reduction should be performed for items within AI/ML constituents. This limitation will be revisited when experience with AI/ML techniques has been gained.

Anticipated MOC-SA-01-2: Metrics

The applicant should define metrics to evaluate the AI/ML constituent performance.

Depending on the application under consideration, a large variety of metrics may be selected to evaluate and optimise the performance of AI/ML constituents. The selected metrics should also provide relevant information with regard to the actual AI/ML constituent reliability so as to substantiate the safety assessment (or impact on services performance in the case of safety support assessment). Indeed, as part of the safety assessment process, AI/ML item failure modes are expected to be identified. Performance metrics should provide a conservative estimation of the probability of occurrence of the AI/ML item failures modes.

Performance evaluation is performed as part of the learning assurance per **Objectives LM-09** (for the trained model) and **IMP-06** (for the inference model). The measured performance is fed back to the safety assessment process.

When input data is within the ODD, the AI/ML constituent will make predictions with the expected level of performance as per Anticipated MOC-SA-01-2 and other performance indicators requested per the learning assurance. However, for various reasons (e.g. sensor failures, shift in OD), input data outside the AI/ML constituent ODD, or even outside the AI-based (sub)system OD may be fed to the AI/ML constituent. In such a situation, the AI-based (sub)system and/or the AI/ML constituent will need to take over the function of the model to deliver an output that will ensure the continuity of the task.

Anticipated MOC-SA-01-3: Exposure to data outside the OD or ODD

Several steps are to take place:

— Establish the monitoring capabilities to detect that the input data is outside the AI/ML constituent ODD, or the AI-based (sub)system OD;

— Put in place mechanisms for the AI/ML constituent to continue to deliver the intended function when input data is outside ODD;

— Put in place mechanisms for the AI-based (sub)system to continue to deliver the intended function when input data is outside OD.

For low-dimensional input space (e.g. categorical data, tabular data, etc.), monitoring the boundaries of the ODD or OD could be a relatively simple task. However, monitoring the limits of the ODD or OD could be much more complicated for high-dimensional input spaces (like in computer vision with images or videos, or in NLP). In such use cases, techniques like the out of distribution (OoD) discriminator could be envisaged.

To support anticipated MOC-SA-01-4 and MOC-SA-01-5, the following taxonomy for uncertainty based on Der Kiureghian and Ditlevsen (Ditlevsen, 2009) is considered in this concept paper:

— Epistemic uncertainty refers to the situation where the model has not been exposed to the relevant input domain area. In other words, the function’s parameters do not correctly fit the input data.

— Aleatory uncertainty refers to the intrinsic randomness in the data. This can derive from data collection errors, sensor noise, or noisy labels. In other words, the model has seen such data during training

Note: The main difference is that epistemic uncertainty can be reduced by adding appropriate data to the training set, while aleatory uncertainty will still be present to a certain extent. Epistemic uncertainty is addressed in this concept paper thanks to the learning assurance objectives, whereas aleatory uncertainties are addressed through the following objectives:

Anticipated MOC-SA-01-4: Identification and classification of uncertainties

Sources of uncertainties affecting the learning algorithm should be listed. Each should be classified to determine whether it is an aleatory or an epistemic source of uncertainties.

Anticipated MOC-SA-01-5: Assessment and mitigation of uncertainties

Aleatory uncertainties should be minimised to the practical extent. Effects of aleatory uncertainties should be assessed at system level. In particular, when a quantitative assessment is required, the aleatory uncertainties should be accounted for in a way that does not compromise safety.

Anticipated MOC-SA-01-6: Establishment of AI/ML item failure modes[13]:

— Establish a taxonomy of AI/ML item failures;

— Evaluate possible failure modes and associated detection means (see also MOC-SA-01-7 for considerations on the generalisation guarantees).

As an example, the following approach may be used to establish AI/ML item failure modes and the associated probability:

1. Describe precisely the desired inputs and outputs of the ML item and the pre-/post-processing steps executed by a traditional SW/HW item.

2. Identify the right metrics to evaluate the model performance and how these allow to reach the required system performance.

3. Understand and quantify generalisation guarantees either through the model complexity approach or through the validation/evaluation approach. This leads to guarantees for almost all data sets on average over all inputs.

Note: This step is done through **Objective LM-04** in the learning assurance chapter. The output of this objective may then be fed into the safety (support) assessment. There may be some iterations between **Objective LM-04** and **Objective SA-01** below in case the generalisation guarantee does not allow meeting the safety objective. In such a case, either a stronger guarantee may be achieved by constraining further the learning process or changes to the system (e.g. system architecture consideration) may be considered.

4. Identify how guarantees on average translate to performance guarantees on each input (with respect to the chosen metrics), up to a controlled failure probability.

5. Analyse the post-processing system to show how it modifies the latter guarantees/failure probabilities. Usually, the post-processing results in improved performance (with respect to the chosen metrics) and/or reduction of model failures.

[13] Based on the state of the art in AI/ML, it is acknowledged that relating the notion of probability in AI/ML with safety analyses is challenging (e.g. as discussed in Section 4.2.4.1 'Uncertainty and risk' in (DEEL Certification Workgroup, 2021)) and subject to further investigation.

6. Understand what performance guarantees (up to the chosen failure probability) follow from the sizes of the chosen data sets, the models, and their in-sample errors (with respect to the chosen metrics).

7. Study the elevated values of the error metrics for the model on the training/validation (eventually testing) data sets, and develop adequate external mitigations such as those discussed in Section C.5. This will prevent errors from exponentially accumulating over time.

Anticipated MOC SA-01-7: Link between generalisation guarantees and safety assessment

Once the generalisation gap has been evaluated (per **Objective LM-04**), the applicant should assess the impact on the safety (support) assessment.

When quantitative assessment is required to demonstrate that the safety requirements are met, AI/ML constituent failure rates may be evaluated from the 'out-of-sample error' (E_{out}). One possible approach is to define the 'in-sample error' (E_{in}) using a metric that reflects application-specific quantities commensurate with the safety hazard. Then, provided that E_{in} is defined in a meaningful and practical way, E_{out}, that reflects the safety performance in operations, can be estimated from the E_{in} and the generalisation gap. Such errors are however quantities on average, and this should be taken into account.

The refinement of this anticipated MOC SA-01-7 or additional anticipated MOC is expected to benefit from MLEAP project deliverables.

Anticipated MOC SA-01-8: Verification

Verify that the implementation satisfies the safety (support) assessment requirements including the independence requirements.

When classical architectural mitigations such as duplicating a function in independent items to improve reliability (i.e. 'spatial redundancy') are put in place, then particular care should be taken to ensure that the expected improvements are achieved (e.g. by checking that items required to be independent have uncorrelated errors).

Note: For non-AI/ML items, traditional safety assessment methodology should be used.

The following standards and implementing rules with adaptation may be used:

— For embedded systems:
  - ED79A/ARP4754A and ARP4761

— For ATS providers: the following implementing rule requirements (and the associated AMC and GM) are applicable:
  - ATS.OR.205 Safety assessment and assurance of changes to the functional system
  - ATS.OR.210 Safety criteria

Note: In Section C.5.1 the purpose of the safety risk mitigation (SRM) building block is defined. The SRM may result in architectural changes to mitigate a partial coverage of the applicable explainability and learning assurance objectives. These architectural mitigations come in addition to the architectural safety mitigations as SRM is not aimed at compensating partial coverage of objectives belonging to the AI trustworthiness analysis building block (i.e. characterisation of AI, safety assessment, information security, ethics-based assessment).

2.2.4. Instruction for continuous safety assessment

Depending on the aviation domains, different approaches exist to ensure that certified/approved systems are in a condition for safe operation, at any time in their operating life.

In the airworthiness domain, activities to ensure continuing airworthiness of the type design are required by Part 21. Such activities consist mainly in the following steps:

— Collection, investigation and analysis of data 21.A.3A(a);

— Reporting potential unsafe conditions 21.A.3A(b);

— Investigation of potential unsafe conditions 21.A.3A(c);

— Determination of an unsafe condition 21.A.3B(b);

— Determination of the required action(s) 21.A.3B(d)3;

— Determination of compliance time for the required action(s) 21.A.3B(d)4; and

— Issuance of an AD 21.A.3B(b).

In the ATM/ANS domain, requirements are set to ensure the safety performance monitoring and measurement. ATS providers shall ensure that the safety assessment comprises the specification of the monitoring criteria necessary to demonstrate that the service delivered by the changed functional system will continue to meet the safety criteria (ATS.OR.205(b)(6)). Also, non-ATS providers shall ensure that the safety support assessment comprises specification of the monitoring criteria necessary to demonstrate that the service delivered by the changed functional system will continue to behave only as specified in the specified context (ATM/ANS.OR.C.005(b)(2)). For both ATS providers and non-ATS providers, these requirements are accompanied with AMC and GM. The monitoring criteria are then used as means to monitor the safety performance in the operations (AMC2 ATM/ANS.OR.B.005(a)(3) with their associated GM like e.g. GM1 ATM/ANS.OR.B.005(a)(3)).

Note: In the rest of this section the notion of ‘continuous safety assessment’ should be understood for the ATM/ANS domain as the ‘Safety performance monitoring and measurement’ for ATS providers, or simply the ‘Performance monitoring and measurement’ for non-ATS providers.

To ensure safe operation of AI-based systems during their operating life, the following objectives are identified to address AI/ML specificities:

Objective ICSA-01: The applicant should identify which data needs to be recorded for the purpose of supporting the continuous safety assessment.

Anticipated MOC ICSA-01

Data should be collected in support of:

— the monitoring of in-service events to detect potential issues or suboptimal performance trends that might contribute to safety margin erosion, or, for non-ATS providers, to service performance degradations; and

— the guarantee that design assumptions hold. This typically covers assumptions made on ODD, (e.g. for further assessment of possible distribution shift).

Proper means should be put in place to ensure the integrity of collected data.

Objective ICSA-02: The applicant should use the collected data to perform a continuous safety assessment. This includes:

— the definition of target values, thresholds and evaluation periods to guarantee that design assumptions hold;
— the monitoring of in-service events to detect potential issues or suboptimal performance trends that might contribute to safety margin erosion, or, for non-ATS providers, to service performance degradations; and
— the resolution of identified shortcomings or issues.

An anticipated way to evaluate safety margin erosion is to update the analysis made during the initial safety assessment with in-service data to ensure that the safety objectives are still met throughout the product life.

More generally, it is expected that best practices and techniques will emerge from in-service experience of continuous safety assessment of AI-based systems. These will enable additional objectives or anticipated MOC to be developed.

2.3. Information security considerations for ML applications

When dealing with ML applications, regardless of the domain considered, the data-driven learning process triggers specific considerations from an ***information security*** perspective.

Focusing on the initial and continuing airworthiness domains, with Decision 2020/006/R, EASA has amended the Certification Specifications (CSs) for large aircraft and rotorcraft, as well as the relevant AMC and GM introducing objectives aimed at ***assessing and controlling safety risks posed by information security threats***. Such threats could be the consequences of ***intentional unauthorised electronic interaction (IUEI)*** with systems on the ground and on board of the aircraft.

For systems and equipment based on AI/ML applications, the above-mentioned modifications to the products certification regulation will serve as a basis to orient the specific guidelines for information security. To this extent, key aspects are:

— the identification of security risks and vulnerabilities through a product information security risk assessment (PISRA) or, more in general, an information security risk assessment;

— the implementation of the necessary mitigations to reduce the risks to an acceptable level (acceptability is defined in the relevant CS for the product); and finally

— the verification of effectiveness of the implemented mitigations. Effectiveness verification should entail a combination of analysis, security-oriented robustness testing and reviews.

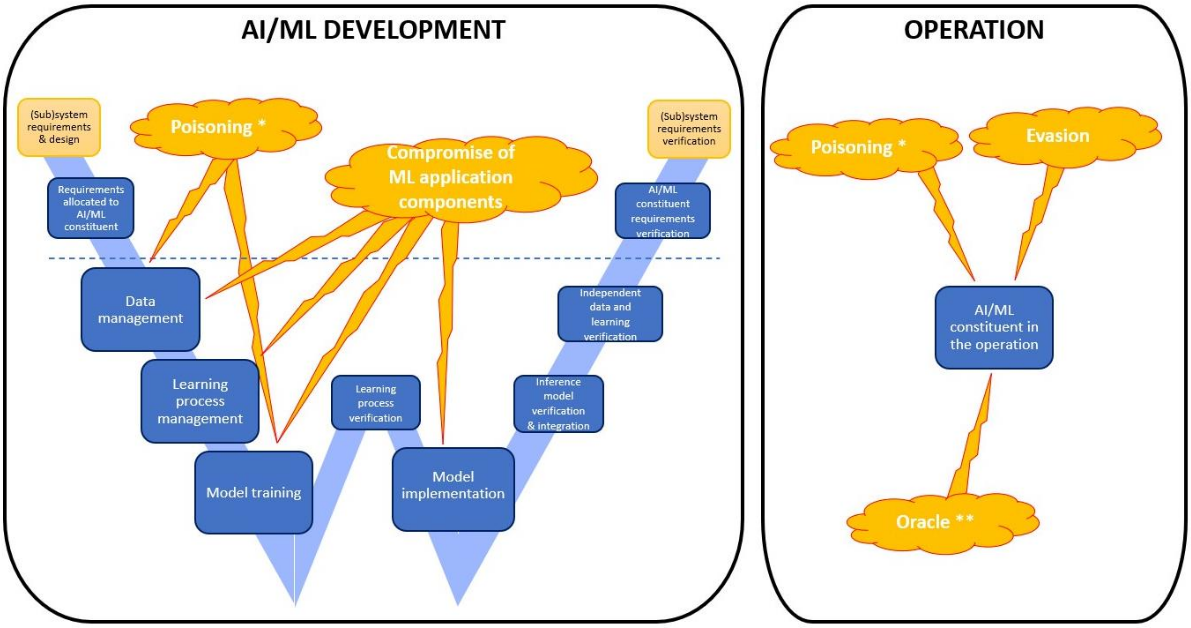
For the initial and continuing airworthiness of airborne systems embedding AI/ML applications, the guidance from ***AMC 20-42 'Airworthiness information security risk assessment'*** is applicable, although contextualised to take into account the peculiarities of the AI/ML techniques.

For other domains, as already stated in Section C.2.2.1.2 for the safety risk assessment, the necessary activities to be performed and documented in view of EASA approval may be different. However, the aforementioned key aspects remain applicable and before dedicated AMC are defined for those other domains, the principles of AMC 20-42 could be used to deal with AI/ML applications information security risk assessment and mitigation.

Moreover, Commission Delegated Regulation (EU) 2022/1645 (applicable as of 16 October 2025) and Commission Implementing Regulation (EU) 2023/203 that have introduced a set of information security requirements for approved organisations, should be also taken into account (see further considerations in Section C.6).

Since security aspects of AI/ML applications are still an object of study, there are no commonly recognised protection measures that have been proved to be effective in all cases. Therefore, we have to consider that the initial level of protection of an AI/ML application may degrade more rapidly if compared to a standard aviation technology. In light of this, systems embedding an AI/ML constituent should be designed with the objective of being resilient and capable of failing safely and securely if attacked by ***unforeseen and novel information security threats***.

Figure 8 — Threats during the life cycle of the AI/ML constituent refers to a set of high-level threats which are harmful to AI-based applications and positions them in the life cycle of the AI/ML constituent. These threats are aligned with the taxonomy and definitions published with the ENISA report (ENISA, December 2021) on SECURING MACHINE LEARNING ALGORITHMS and possible threats identified in Table 3. As depicted on the figure, these attacks can be preliminary steps to more complex attacks, like model extraction. This set is subject to change depending on application specificities and threats evolutions.



* Can be in support of backdoor attacks
** Can be a preliminary attack before model extraction

Figure 8 — Threats during the life cycle of the AI/ML constituent

2.3.1. Proposed objectives for the information security risks management

Based on the high-level considerations made in the previous section, and recognising that the management of identified risks is an iterative process that requires assessment and implementation of mitigation means until the residual risk is acceptable (acceptability criteria depend on the context that is considered for the certification of the affected product or part), the following objectives are considered in the guidelines:

Objective IS-01: For each AI-based (sub)system and its data sets, the applicant should identify those information security risks with an impact on safety, identifying and addressing specific threats introduced by AI/ML usage.

Anticipated MOC IS-01: In performing the system information security risk assessment and risk treatment, while taking advantage of the ENISA report (ENISA, December 2021) on SECURING MACHINE LEARNING ALGORITHMS and possible threats identified in Table 3, the applicant could address the following aspects:

— Consider ‘evasion’ attacks, in which the attacker works on the learning algorithm’s inputs to find small perturbations leading to large modification of its outputs (e.g. decision errors).

— Consider the ‘oracle’ type of attack in which the attacker explores a model by providing a series of carefully crafted inputs and observing outputs. These attacks can be predecessors to more harmful types, evasion, poisoning, or even model extraction.

Objective IS-02: The applicant should document a mitigation approach to address the identified AI/ML-specific information security risk.

Anticipated MOC IS-02: Based on the identified threats, the applicant should apply security controls that are specific to applications using ML. Some are listed in Table 5 -section 'SPECIFIC ML' of the ENISA report (ENISA, December 2021) and appear to be in line with some of the learning assurance objectives (see Section C.3).

Objective IS-03: The applicant should validate and verify the effectiveness of the security controls introduced to mitigate the identified AI/ML-specific information security risks to an acceptable level.

2.4. Ethics-based assessment

As already mentioned above, the EU Commission's AI High-Level Expert Group (HLEG), in 2019, elaborated that, deriving from a fundamental-rights-based and domain-overarching list of ***4 ethical imperatives*** (i.e. respect to human autonomy, prevention of harm, fairness and explainability), the ***trustworthiness of an AI-based system*** is built upon ***3 main pillars*** or expectations, i.e. lawfulness[14], adherence to ethical principles, and technical robustness. The HLEG further refined these expectations by means of a set of ***7 gears*** and sub-gears (i.e. human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; accountability). To ease self-evaluation and provide orientation to applicants, the HLEG, in 2020, underpinned this set of gears by the so-called ***Assessment List for Trustworthy AI (ALTAI)***[15], containing several questions and explanation.

Building on this 2019/2020 Commission approach, the present EASA guidelines further clarify and tailor the (sub-)gears of the HLEG to the EASA remit and to the needs of the aviation sector and its stakeholders. This is reflected in a slightly adapted wording of the ALTAI, as reflected in the mapping tables provided in Annex 5 — Full list of questions from the ALTAI adapted to aviation.

Objective ET-01: The applicant should perform an ethics-based trustworthiness assessment for any AI-based system developed using ML techniques or incorporating ML models.

Anticipated MOC ET-01: When performing this assessment, it is suggested to take into account the seven gears from the Assessment List for Trustworthy AI (ALTAI), while considering the clarifications and specific objectives developed by EASA in the following sections (one section per gear).

14 Note: With regard to the 'lawfulness' component, the HLEG-Ethics guidelines state (p. 6): 'The Guidelines do not explicitly deal with the first component of Trustworthy AI (lawful AI), but instead aim to offer guidance on fostering and securing the second and third components (ethical and robust AI). While the two latter are to a certain extent often already reflected in existing laws, their full realisation may go beyond existing legal obligations.'

15 https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

2.4.1. Gear #1 — Human agency and oversight

Most of the questions related to 'Human agency and autonomy' and 'Human oversight' are considered by EASA to be addressed through compliance with the objectives of the AI trustworthiness framework contained in Chapter C of this document.

The table for gear #1 contained in Annex 5 — Full list of questions from the ALTAI adapted to aviation is intended to clarify the precise links to the EASA guidelines.

The following objective should be addressed in the ethics-based assessment that is requested through **Objective ET-01**.

Objective ET-02: The applicant should ensure that the AI-based system bears no risk of creating human attachment, stimulating addictive behaviour, or manipulating the end user's behaviour.

Anticipated MOC ET-02: AI-based systems with the potential of creating human attachment, stimulating addictive behaviour, or manipulating user behaviour are not considered acceptable for the aviation domain. In the frame of these guidelines, the understanding of item G1.f requires some precision on the definition of the terms 'overreliance' and 'attachment' which have been added in Annex 3 - Definitions. A notable difference is that attachment is related to an emotional link whereas overreliance is more pragmatically related to trust and dependence on support. Risks related to 'overreliance' are considered to be addressed through the guidance on operational explainability as reflected for item G1.c. The organisation processes and procedures shall ensure that the risks associated with this item G1.f and its associated sub-items are strictly avoided. In addition, it is important to clarify the differences between the terms 'overreliance' and 'reliance' in order to better delineate the border between what is suitable (reliance) and what is not acceptable (overreliance).

2.4.2. Gear #2 — Technical robustness and safety

Most questions related to 'Technical robustness and safety' are considered by EASA to be addressed through compliance with the objectives of the AI trustworthiness framework contained in Chapter C of this document.

The table for gear #2 contained in the Annex 5 — Full list of questions from the ALTAI adapted to aviation is intended to clarify the precise links to the EASA guidelines.

The following objective should be addressed in the ethics-based assessment that is requested through **Objective ET-01**.

Objective ET-03: The applicant should ensure that the AI-based system presents no capability of adaptive learning.

Note: Adaptive learning (also known as continual or online learning) is not addressed in the current guidelines; therefore, such applications will not be accepted by EASA at this stage.

2.4.3. Gear #3 — Privacy and data governance

All ALTAI questions related to 'Privacy and data governance' in terms of personal data are considered to be addressed through compliance with the EU and national data protection regulations, including, as applicable, involvement of the national DPO, consultation with the National Data Protection Authority, etc.

The table for gear #3 contained in the Annex 5 — Full list of questions from the ALTAI adapted to aviation is intended to clarify the precise links to the EASA guidelines.

The following objective should be addressed in the ethics-based assessment that is requested through **Objective ET-01**.

Objective ET-04: The applicant should comply with national and EU data protection regulations (e.g. GDPR), i.e. involve their Data Protection Officer (DPO), consult with their National Data Protection Authority, etc.

Anticipated MOC ET-04: The applicant should thus ensure and provide a confirmation that a 'data protection'-compliant approach was taken, e.g. through a record or a data protection impact assessment (DPIA).

2.4.4. Gear #4 — Transparency

All questions related to 'Transparency' are considered to be addressed through compliance with the objectives of the AI trustworthiness framework contained in Chapter C of this document.

The table for gear #4 contained in the Annex 5 — Full list of questions from the ALTAI adapted to aviation is intended to clarify the precise links to the EASA guidelines.

2.4.5. Gear #5 — Diversity, non-discrimination and fairness

This gear may not be applicable to all aviation use cases. Therefore, EASA encourages applicants in a first analysis, to check whether the AI-based system could have any impact on diversity, non-discrimination and fairness. Diversity, non-discrimination and fairness, in the context of Gear #5, have to be interpreted as applying to individual persons or groups of people, not to data sources (which are addressed through the Learning Assurance guidance).

If no impact exists, the outcome of this analysis should be recorded in the ethics-based assessment documentation.

In case of an impact, please consider the questions from the ALTAI related to Gear #5. The table for gear #5 contained in the Annex 5 — Full list of questions from the ALTAI adapted to aviation is intended to clarify the precise links to the EASA guidelines.

The following objective should be addressed in the ethics-based assessment that is requested through **Objective ET-01**.

Objective ET-05: The applicant should ensure that procedures are in place to avoid creating or reinforcing unfair bias in the AI-based system, regarding both the data sets and the trained models.

Anticipated MOC ET-05: The applicant should establish means (e.g. an ethics-based policy, procedures, guidance or controls) to raise the awareness of all people involved in the development of the AI-based system in order to avoid the creation or reinforcement of unfair bias in the AI-based system (regarding both input data and ML model design), as far as such unfair bias could have a negative impact on safety.

2.4.6. Gear #6 — Societal and environmental well-being

Environmental well-being

The following objectives should be addressed in the ethics-based assessment that is requested through **Objective ET-01**.

Objective ET-06: The applicant should perform an environmental impact analysis, identifying and assessing potential negative impacts of the AI-based system on the environment and human health throughout its life cycle (development, deployment, use, end of life).

Anticipated MOC ET-06: The environmental impact analysis should address at least the following questions:

— Does the AI-based system require additional energy and/or generates additional carbon emissions throughout its life cycle compared to other (non-AI-based) systems?

  - While there is no agreed international guidance on how to assess the environmental impact of software, various research initiatives have tried to identify criteria and indicators that could be taken into account for such an assessment. In particular, the German Environment Agency (UBA) has developed a list of software sustainability criteria covering the domains of resource efficiency (system requirements, hardware utilisation, energy efficiency), the potential useful life of hardware (backward compatibility, platform independence and portability, hardware sufficiency), and user autonomy[16].

— Does the AI-based system have adverse effects on the regulated aircraft/engine noise and emissions or aircraft fuel venting?

— Does the AI-based system have adverse effects on the product’s environmental performance in operation?

  - If relevant, the applicant should consider at least adverse effects on aircraft fuel consumption (CO_2 emissions) and aircraft noise around airports.

— Could the use of the AI-based system have rebound effects, e.g. lead to an in increase in traffic, which in turn could become harmful for the environment or human health?

[16] Kern et al., Sustainable software products – Towards assessment criteria for resource and energy efficiency, Elsevier B.V., 2018.

— Could the use of the AI-based system have direct effects on the human health, including the right to physical, mental and moral integrity?

Objective ET-07: The applicant should define measures to reduce or mitigate the impacts identified under **Objective ET-06.**

Anticipated MOC ET-07: The applicant should follow standard practices in environmental management as documented in the European Union's Eco-Management and Audit Scheme (EMAS) or ISO 14001. In particular, the applicant should implement procedures in line with the principles of the Plan-Do-Check-Act (PDCA) cycle.

Impact on work and skills and on society at large or democracy

Except for topics related to **Objective ET-08** and **Objective ET-09**, this sub-gear may not be applicable to all aviation use cases. Therefore, EASA encourages applicants in a first analysis to check whether the AI-based system could have any impact on work and skills.

If no impact exists, the outcome of this analysis should be recorded in the ethics-based assessment documentation.

In case of an impact, please consider the questions from the ALTAI related to Gear #6 'Work and skills' and 'Impact on society at large or democracy'. The original questions can be found in Annex 5 — Full list of questions from the ALTAI adapted to aviation. The assessment of the answers to these questions does not fall under the remit of EASA and would be performed by a competent authority, at European level or at national level as applicable.

The following objectives should be addressed in the ethics-based assessment that is requested through **Objective ET-01**.

Objective ET-08: The applicant should identify the need for new skills for users and end users to interact with and operate the AI-based system, and mitigate possible training gaps (link to **Provision ORG-06, Provision ORG-07**).

Objective ET-09: The applicant should perform an assessment of the risk of de-skilling of the users and end users and mitigate the identified risk through a training needs analysis and a consequent training activity (link to **Provision ORG-06, Provision ORG-07**).

2.4.7. Gear #7 — Accountability

Some of the 'Accountability' gear items may not be applicable to all aviation use cases. Therefore, EASA encourages applicants in a first analysis to check whether the AI-based system could have any impact on the monitoring of ethical concerns from an organisation's perspective.

If no impact exists, the outcome of this analysis should be recorded in the ethics-based assessment documentation (per **Objective ET-01**).

In case of an impact, please consider the respective questions from the ALTAI related to Gear #7 'Accountability' (see Annex 5).

2.4.8. Link to the AI trustworthiness building blocks

The following figure provides an overview of the distribution of the ethical gears over the AI trustworthiness building blocks:

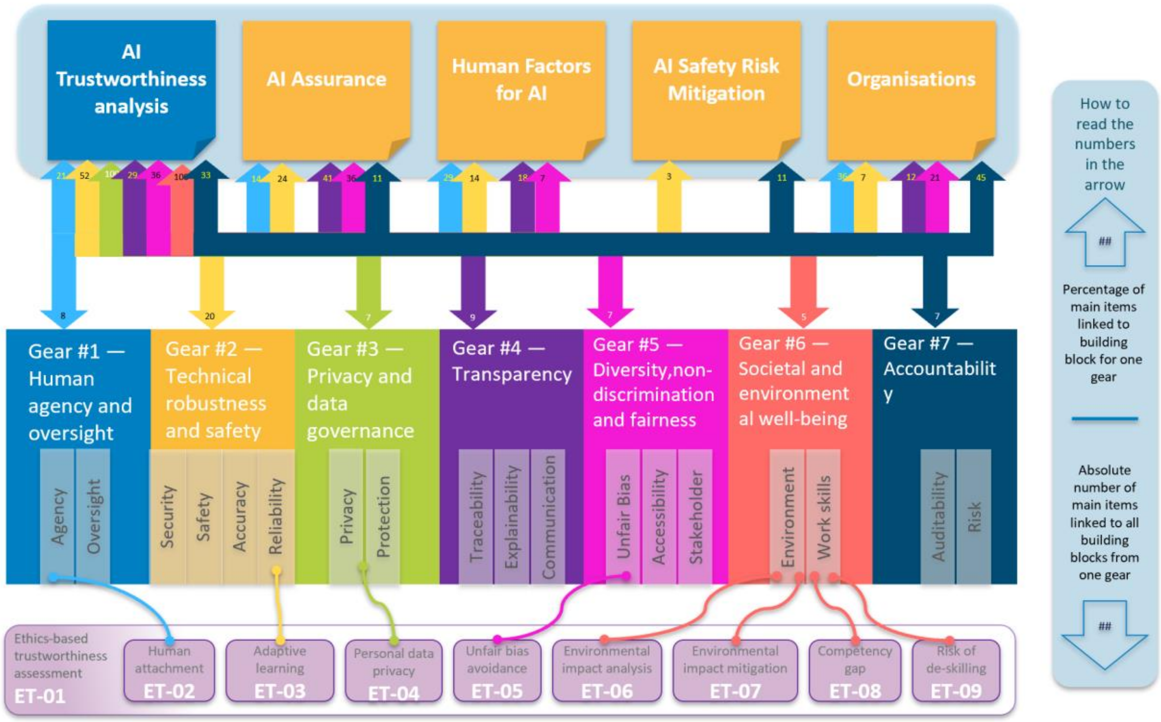


Figure 9 — Mapping of the 7 gears to the AI trustworthiness bullding-blocks

The following figure provides an overview of the ALTAI items requiring additional oversight from authorities other than EASA:

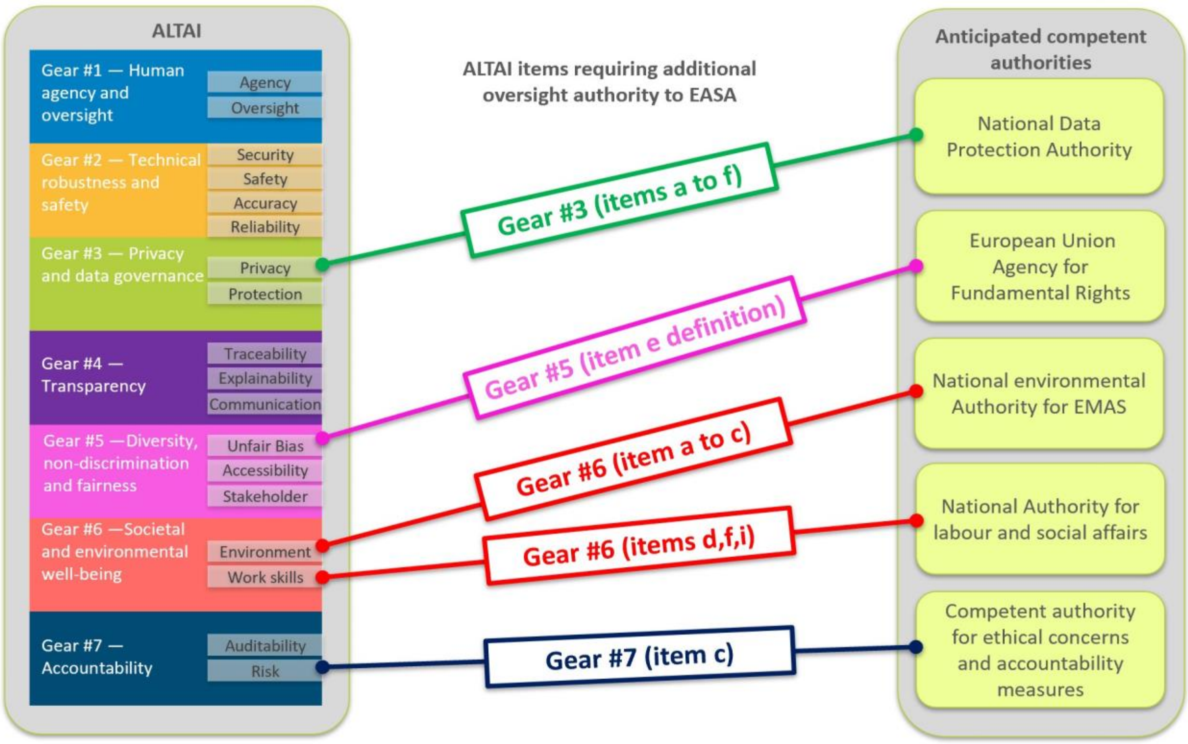


Figure 10 — Mapping of the 7 gears to the AI trustworthiness building-blocks

3. AI assurance

The ***AI assurance*** building block proposes system-centric guidance to address the development of the AI-based system. This system-centric view is then complemented with an end-user centric approach which will put some focus on ***human factors aspects of AI*** (see Section C.4).

The ***AI assurance*** defines objectives to be fulfilled by the AI-based system, considering the novelties inherent to ML techniques, as depicted in Section B.6.1.

Recognising the limitations of traditional development assurance for data-driven approaches, the ***learning assurance*** concept is defined in Section C.3.1, and then associated objectives are developed, with an emphasis on data management aspects and learning processes.

Another set of objectives address the perceived concerns regarding lack of transparency of the ML models under the ***development explainability*** Section C.3.2.

Finally, the ***AI assurance*** continues during the operations of the AI-based system and with a set of ***data-recording*** objectives in Section C.3.2.5 which will serve as an entry for many different aspects to be addressed by the guidance. The data-recording capabilities of the AI-based system will indeed feed the continuous safety assessment, the monitoring by the applicant of the performance of the system during its actual operations, as well as the investigations by the safety investigators in case of an incident or accident.

3.1. Learning assurance

The learning assurance concept aims at providing assurance on the intended function of the AI-based system at an appropriate level of performance, and at ensuring that the resulting trained models possess sufficient guarantees of generalisation and robustness.

To illustrate the anticipated learning assurance process steps, EASA proposes the following W-shaped process outline.

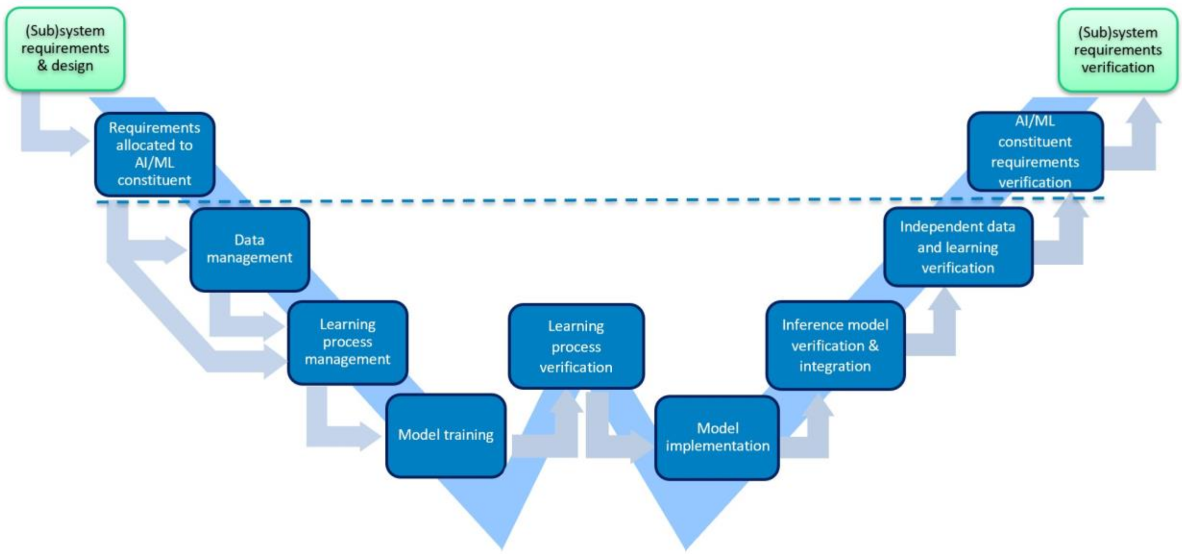


Figure 11 — Learning assurance W-shaped process

This cycle adapts the typical development assurance V-cycle to ML concepts and allows to structure the learning assurance guidance.

The dotted line is here to make a distinction between the use of traditional development assurance processes (above) and the need for processes adapted to the data-driven learning approaches (below).

Note: The pure learning assurance processes start below the dotted line. It is however important to note that this dotted line is not meant to split specific assurance domains (e.g. system / software).

This W-shaped process is concurrent with the traditional V-cycle that is required for development assurance of non-AI/ML constituents.

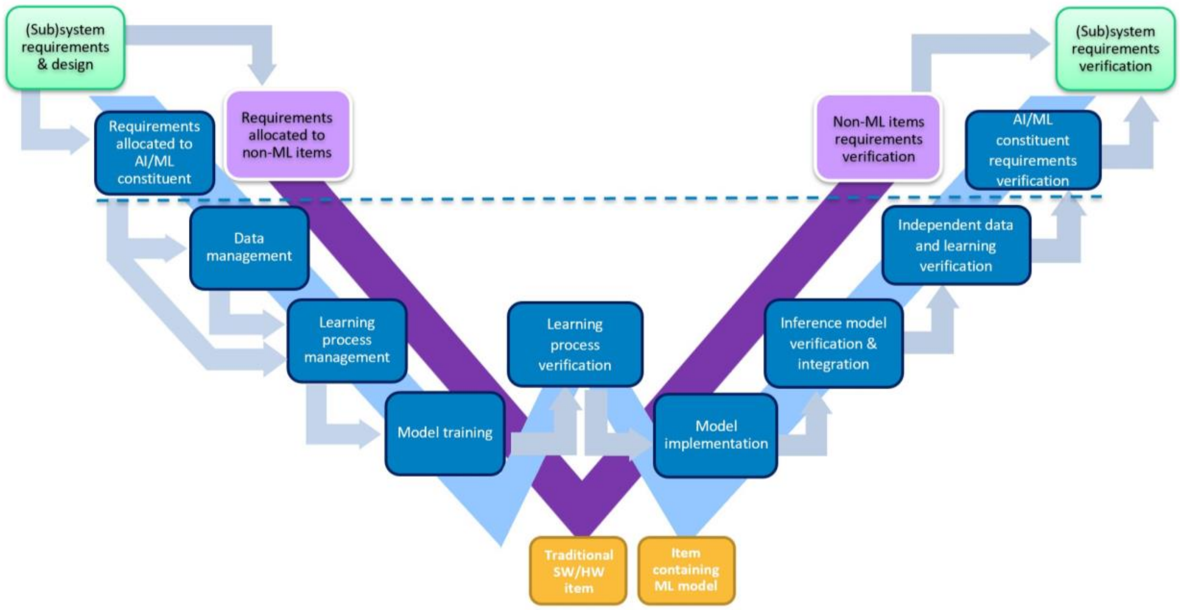


Figure 12 — Global view of learning assurance W-shaped process, non-AI/ML constituent V-cycle process

This new learning assurance approach will have to account for the specific phases of learning processes, as well as to account for the highly iterative nature of certain phases of the process (orange and green arrows).

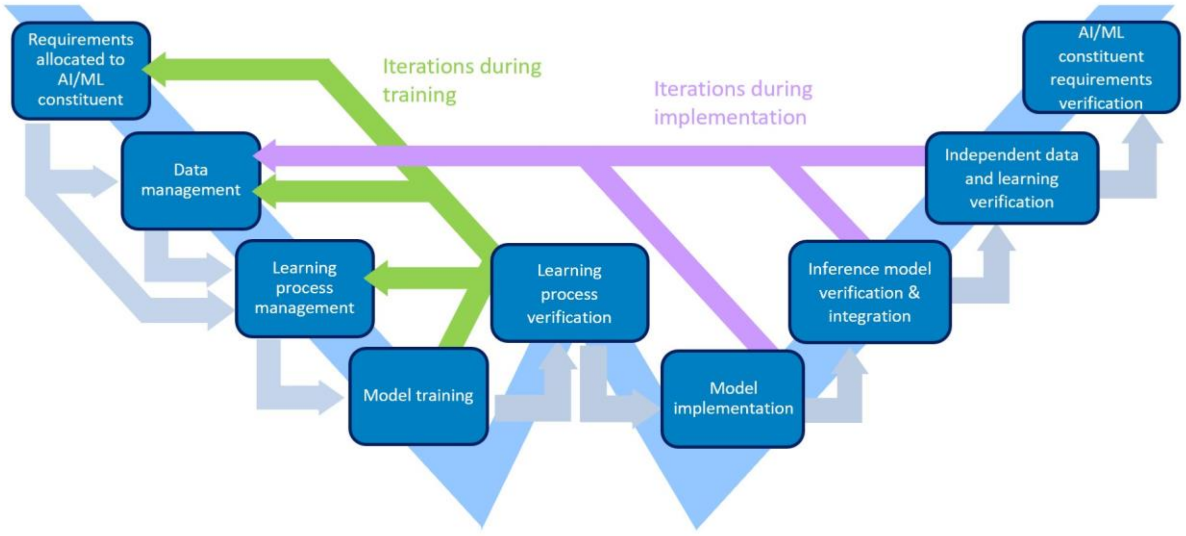


Figure 13 — Iterative nature of the learning assurance process

As with traditional software development where DevOps methodologies or frameworks could be deployed at the applicant organisational level, some applicants could implement MLOps principles and frameworks for the development of the AI/ML constituent of an AI-based system.

MLOps is a set of practices and tools that helps organisations to improve the speed, reliability, and security of the ML process. It aims to bridge the gap between data scientists, who are responsible for developing ML models, and software engineers, who are responsible for deploying and maintaining those models in production.

In the context of safety-related systems, MLOps is particularly important because it helps organisations to ensure that their ML models continuously deliver accurate and reliable results. This is especially important in the aviation domain.

Some key components of MLOps for safety-related systems include:

— Version control: ML models should be treated like any other software, with strict version control to ensure that only tested and approved models are deployed to production;

— Continuous integration and delivery:

  - Automated data pipelines: data pipelines can be automated using tools that can handle tasks from the data management process[17];

  - Automated pipelines can be used to build and test the ML models quickly and reliably; and

— Model drift detection on recorded data: over time, the data that an ML model is trained on may change, leading to a phenomenon known as 'model drift'. MLOps practices should include methods for detecting and addressing model drift to ensure that the model remains accurate and reliable.

3.1.1. Learning assurance process planning

Objective DA-01: The applicant should describe the proposed learning assurance process, taking into account each of the steps described in Sections C.3.1.2 to C.3.1.12, as well as the interface and compatibility with development assurance processes.

Anticipated MOC DA-01-1: The set of plans should include a plan for learning assurance (e.g. plan for learning aspects of certification), addressing all objectives from Section C.3 and detailing the proposed MOC.

3.1.2. Requirements and architecture management

The ***requirements management*** process covers the preparation of a complete set of requirements for the design of the ***AI/ML constituent***. This step may be divided in several successive refinement steps and is preceded by a traditional flow-down of requirements (e.g. from aircraft to system for the ***initial and continuing airworthiness*** or ***air operations*** domains).

[17] While it is not possible to completely automate all the process steps (e.g. feature engineering) there are ways to make it more efficient (e.g. automating the feature selection by ranking and scoring the features).

This step is further divided in:

— requirements capture;

— AI-based (sub)system architecture development[18];

— requirements validation.

3.1.2.1. Capture of (sub)system requirements allocated to the AI/ML constituent

Based on the definition of the ConOps and OD (**Objective CO-04**), ***requirements capture*** consists in the capture and unique identification of all requirements allocated to the AI/ML constituent, which are necessary to design and implement the AI/ML constituent.

Objective DA-02: Documents should be prepared to encompass the capture of the following minimum requirements:

— safety requirements allocated to the AI/ML constituent;

— information security requirements allocated to the AI/ML constituent;

— functional requirements allocated to the AI/ML constituent;

— operational requirements allocated to the AI/ML constituent, including ODD and AI/ML constituent performance monitoring (to support related objectives in Section C.4.1.4.2), detection of OoD input data and data-recording requirements (to support objectives in Section C.3.2.5);

— non-functional requirements allocated to the AI/ML constituent (e.g. performance, scalability, reliability, resilience, etc.); and

— interface requirements.

3.1.2.2. AI-based (sub)system architecture development

AI-based (sub)system and constituents architecture development is not a novel step compared to traditional systems development approaches; it is however an essential step in detailing the AI-based system, subsystem (if applicable) and AI/ML constituents architecture.

Objective DA-03: The applicant should describe the system and subsystem architecture, to serve as reference for related safety (support) assessment and learning assurance objectives.

3.1.2.3. Requirements validation

Requirements validation is considered to be covered by traditional system development methods. (e.g. ED-79A/ARP-4754A for product certification).

Objective DA-04: Each of the captured requirements should be validated**.**

3.1.2.4. Management of derived requirements

The ***data management*** process, the ***learning process management***, and the ***trained model implementation*** process described in the following sections of the document produce requirements

[18] This step is different from the model architecture described in Section C.3.1.4.

which may not be always traceable to the higher-level requirements discussed above. These ***derived requirements*** are a subpart of the requirements produced via **Objective DM-01**, **Objective DM-02**, **Objective DM-03**, **Objective LM-01**, **Objective LM-02**, **Objective LM-04**, and **Objective IMP-01**, and need special attention by the applicant.

Objective DA-05: The applicant should document evidence that all derived requirements generated through the learning assurance processes have been provided to the (sub)system processes, including the safety (support) assessment.

Note: In order to determine the effects of derived requirements on both the (sub)system requirements and the safety (support) assessment, all derived requirements should be made available to the (sub)system processes including the safety (support) assessment.

Objective DA-06: The applicant should document evidence of the validation of the derived requirements, and of the determination of any impact on the safety (support) assessment and (sub)system requirements.

Notes:

— Derived requirements should be validated as other higher-level requirements produced at (sub)system level.

— Derived requirements should be reviewed from a safety (support) perspective. They should be examined to determine which function they support so that the appropriate Failure Condition classification can be assigned to the requirements validated.

3.1.3. Data management

The ***data management*** process is the first step of the data life cycle management. Figure 14 below depicts the main activities covered.

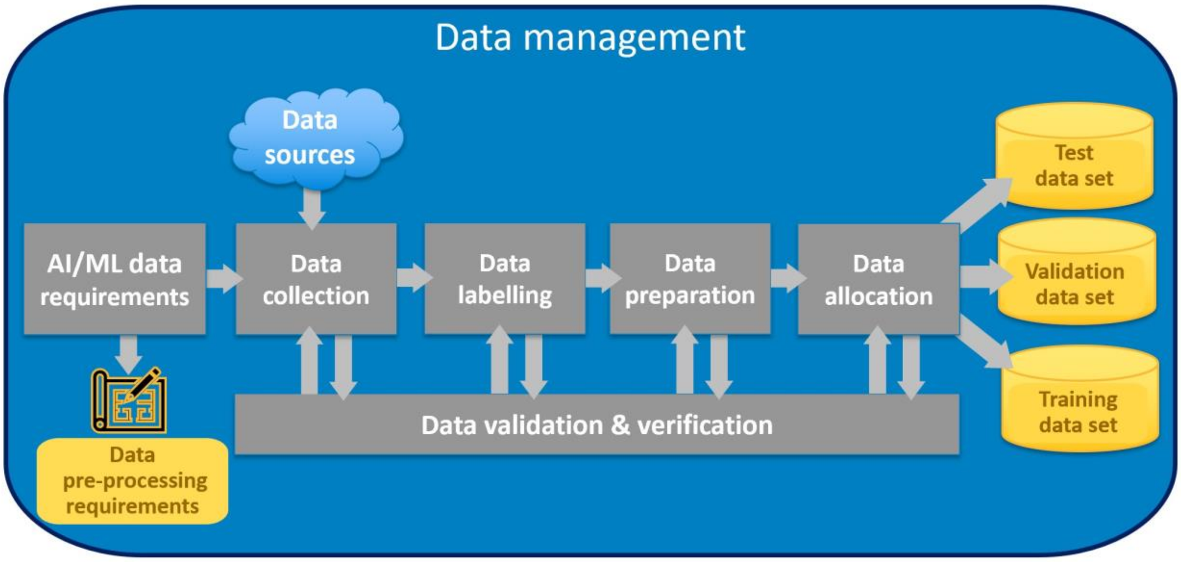


Figure 14 — Data management process

The ***data management*** process covers:

— data management requirements capture;

— data collection;

— data labelling;

— data preparation (pre-processing, data transformation and feature engineering);

— identification of the various data sets used in the learning phase (typically the training, validation and test data sets);

— data sets validation and verification (including accuracy, completeness and representativeness, with respect to the ML requirements and the AI/ML constituent ODD);

— independence requirements between data sets;

— identification and elimination of unwanted bias inherent to the data sets.

The data generated by the ***data management*** process is verified at each step of the process against the subset of data quality requirements (DQRs) pertaining to this step.

3.1.3.1. Data management requirements

The ***data management*** process will encompass its own requirements.

The requirements capture will benefit from a precise definition of the AI/ML constituent ODD which consists in a refinement of the defined OD (see **Objective CO-04**).

Objective DM-01: The applicant should define the set of parameters pertaining to the AI/ML constituent ODD.

Figure 15 — AI/ML constituent ODD shows the refinement of the OD into the AI/ML constituent ODD.

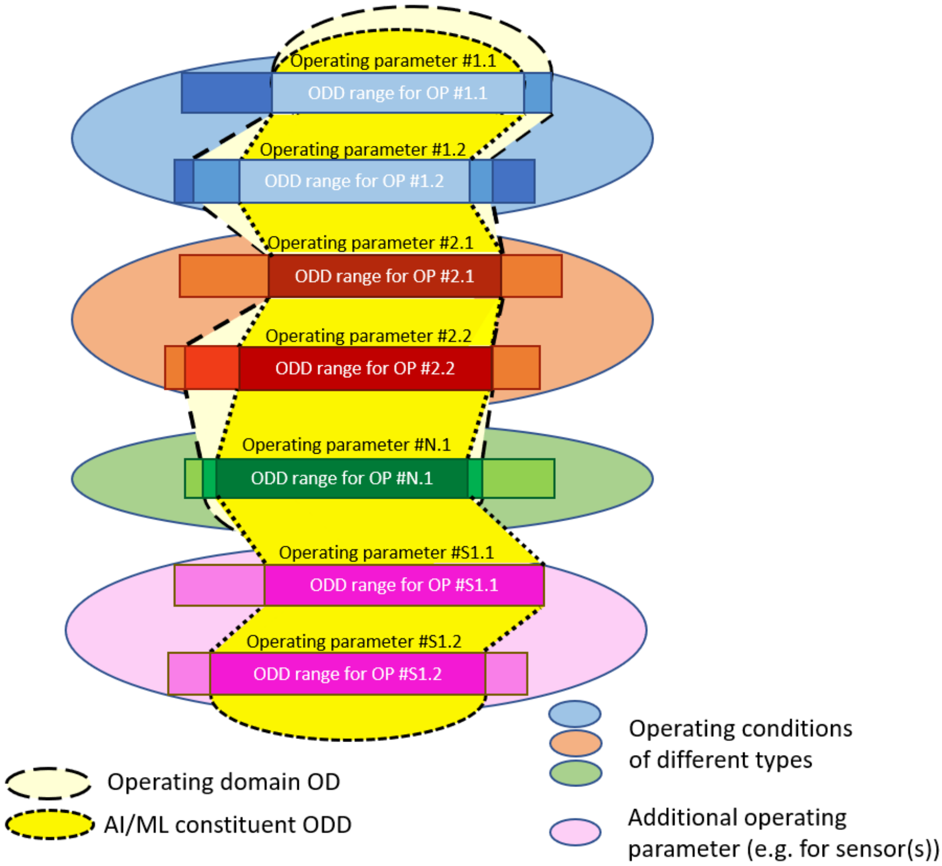


Figure 15 — AI/ML constituent ODD

Notes:

— Additional parameters can be identified and defined for the AI/ML constituent ODD (e.g. parameters linked to the sensors used for the input data of the ML model like brightness, contrast characteristics of a camera, level of blur coming from vibrations at the level of a camera, or characteristics like sensitivity, directionality of a microphone, etc.);

— Ranges for the parameters in the AI/ML constituent ODD can be a subset or superset of the ranges at the level of the operation domain (OD) (see Figure 16 — Definition of ODD ranges versus OD ranges below);

— Exceptionally, one or few ranges for the parameters in the AI/ML constituent ODD can be a superset of the ranges for the corresponding parameters at the level of the OD (in order to improve the performance of the model for these parameters) (see Figure 16 — Definition of ODD ranges versus OD ranges below)

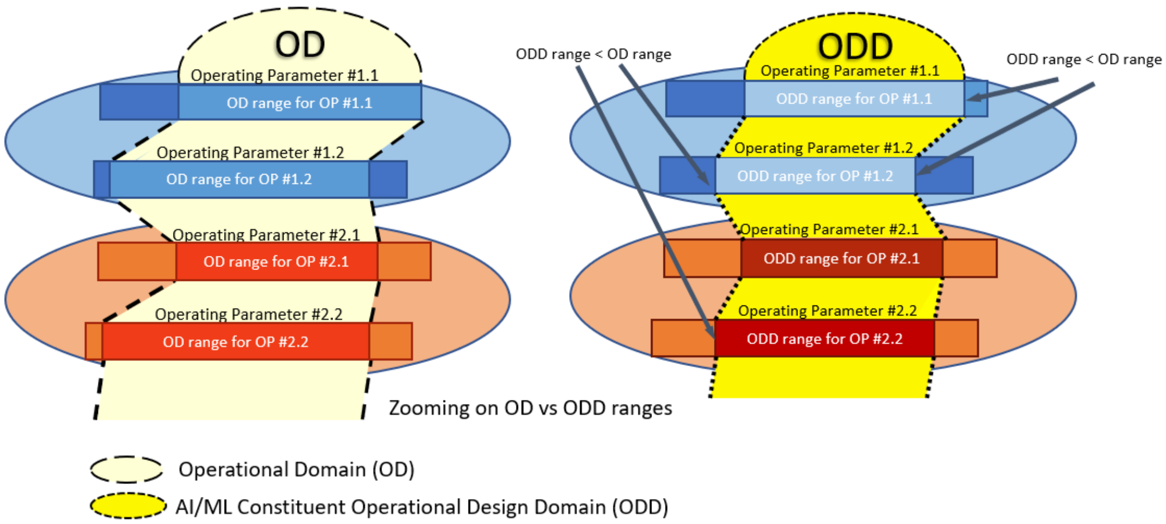


Figure 16 — Definition of ODD ranges versus OD ranges

— As for the OD, the range(s) for one or several operating parameters could depend on the value or range of another parameter as depicted in Figure 17 — Relations between operating parameters in ODD:

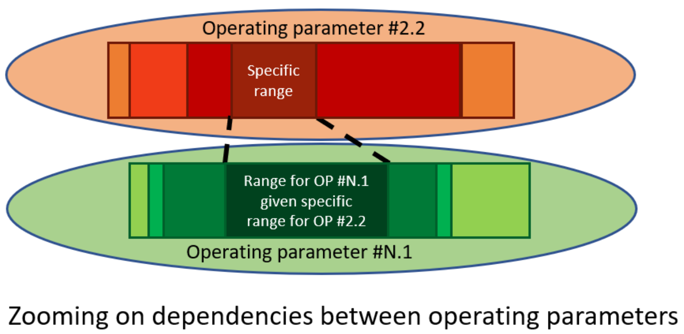


Figure 17 — Relations between operating parameters in ODD

Anticipated MOC DM-01-1: The definition of the parameters pertaining to the AI/ML constituent ODD should be the outcome of the 'ML Constituent Operational Design Domain Process' depicted in Section 6.2 of ED-XXX/AS6983.

During the different iterations which will happen during the learning phase, particular attention should be paid to:

— the definition of nominal data;

— the identification of edge cases, corner cases data in preparation of stability of the model;

— the definition of infeasible corner cases data;

— the detection and removal of inliers;

— the detection and management of novelties;

— the definition of outliers for their detection and management.

In parallel with the definition of the AI/ML constituent ODD, a subset of these requirements will deal with DQRs.

Objective DM-02: The applicant should capture the DQRs for all data pertaining to the data management process, including but not limited to:

— the data needed to support the intended use;

— the ability to determine the origin of the data;

— the requirements related to the annotation process;

— the format, accuracy and resolution of the data;

— the traceability of the data from their origin to their final operation through the whole pipeline of operations;

— the mechanisms ensuring that the data will not be corrupted while stored or processed,

— the completeness and representativeness of the data sets; and

— the level of independence between the training, validation and test data sets.

Anticipated MOC DM-02-1: Starting from ED-76A Section 2.3.2 and accounting for specificities of data-driven learning processes, the DQRs should characterise, for each type of data representing an operating parameter of the AI/ML constituent ODD:

— the accuracy of the data;

— the resolution of the data;

— the quality of the annotated data;

— the confidence that the data has not been corrupted while stored, processed or transmitted over a communication network (e.g. during data collection);

— the ability to determine the origin of the data (traceability);

— the level of confidence that the data is applicable to the period of intended use (timeliness);

— the data needed to support the intended use (completeness and representativeness); and

— the format of the data, when needed.

Note: Where relevant for the AI/ML based system, the requirement on representativeness of the data sets will consider the diversity among the end users (e.g. accents in NLP applications, etc.).

The MOC will need refinements based on progress in the standardisation (e.g. EUROCAE/SAE WG-114/G-34) and other best practices (e.g. reference: (DEEL Certification Workgroup, 2021)).

The ***data management*** process will also capture the requirements to be transferred to the implementation, regarding the pre-processing and feature engineering to be performed on the inference model.

Objective DM-03: The applicant should capture the requirements on data to be pre-processed and engineered for the inference model in development and for the operations.

3.1.3.2. ML constituent ODD validation

Objective DM-04: The applicant should ensure the validation to an adequate level of the correctness and completeness of the ML constituent ODD.

Anticipated MOC DM-04: The correctness and completeness of the operating parameters of the ML constituent ODD, as well as their ranges and interdependencies should be reviewed by representatives from the business/system and the data science.

3.1.3.3. Data management requirements validation

Objective DM-05: The applicant should ensure the validation of the correctness and completeness of requirements on data to be pre-processed and engineered for the trained and inference model, as well as of the DQRs on data.

These validated requirements are then used during the data management process and also transferred to the implementation phase.

3.1.3.4. Data collection

The collection of data can be of different nature depending on the project (i.e. database, text, image, video, audio records); the applicant should always take into account that the data collected might fall under the category of *personal data* or affect *privacy.* In this case, there is a need to take into account Gear #3 of this Guidance since personal data requires special protection.

The ***data collection*** should identify the different sources of data of relevance to the training.

Objective DM-06: The applicant should identify data sources and collect data in accordance with the defined ODD, while ensuring satisfaction of the defined DQRs, in order to drive the selection of the training, validation and test data sets.

The sources of data are inherent to the AI/ML project. The sources can be internal or external to the applicant. External sources can be open-source or sourced via a contract to be established between the applicant and the data provider (e.g. weather data from a MET office, or databases shared between aeronautical organisations).

Depending on data sources, data sampling could be applied (simple random sampling, clustered sampling, stratified sampling, systematic sampling, multiphase sampling (reference: (DEEL Certification Workgroup, 2021)). The applicant should ensure completeness and representativeness of the sampling.

In order to address a lack of data completeness or representativeness, additional data may need to be collected via data augmentation techniques (e.g. image rotation, flipping, cropping in computer vision), or the existing data may be complemented with synthetic data (e.g. coming from models, digital twins, virtual sensors).

3.1.3.5. Data labelling

In the context of supervised learning techniques, the data set will need to be annotated or labelled.

Objective DM-07: Once data sources are collected and labelled, the applicant should ensure the high quality of the annotated or labelled data in the data set.

All data items are annotated according to a specific set of annotation requirements, created, refined and reviewed by the applicant. Annotation can be a manual or automated process. Depending on the project, the annotation step can be effort-consuming (e.g. image annotations for detection purposes), and the applicant could decide to keep the annotation step insourced or outsourced, depending on its capabilities. In the case of outsourcing of the activity, the applicant should decide on the DQRs to be achieved by the supplier.

3.1.3.6. Data preparation

The ***data preparation*** is paramount as it will be a key success factor for the ability of the AI/ML constituent to generalise. The ***data preparation*** is a multi-step process which involves a very significant part of the effort needed to implement an AI/ML constituent.

All operations on the data during ***data preparation*** should be performed in a way that ensures that the requirements on data are addressed properly, in line with the defined ODD.

Objective DM-08: The applicant should define the data preparation operations to properly address the captured requirements (including DQRs).

The main steps of the ***data preparation*** consist of:

— the pre-processing of the data, which is the act of cleaning and preparing the data for training;

— the feature engineering, aiming at defining the most effective input parameters from the data set to enable the training; and

— the data normalisation.

Note: Feature engineering does not apply to all ML techniques.

Data pre-processing

The ***data pre-processing*** should consist in a set of basic operations on the data, preparing them for the ***feature engineering*** or the ***learning process***.

Objective DM-09: The applicant should define and document pre-processing operations on the collected data in preparation of the training.

Anticipated MOC DM-09: Depending on data sets, different aspects should be considered for cleaning and formatting the data:

— fixing up formats, typically harmonising units for timestamp information, distances and temperatures;

— binning data (e.g. in computer vision, combining a cluster of pixels into one single pixel);

— filling in missing values (e.g. some radar plot missing between different points on a trajectory); different strategies can apply in this case, either removing the corresponding row in the data set, or filling missing data (in general by inputting the mean value for the data in the data set);

— correcting erroneous values or standardising values (e.g. spelling mistakes, or language differences in textual data, cropping to remove irrelevant information from an image);

— identification and management of outliers (e.g. keeping or capping outliers, or sometimes removing them depending on their impact on the DQRs).

For all the above steps, a mechanism should be put in place to ensure sustained compliance with the DQRs after any data transformation.

Feature engineering

Feature engineering is a discipline consisting in transforming the pre-processed data so that it better represents the underlying structure of the data to be an input to the model training.

It is to be noted that ***feature engineering*** does not apply to all ML techniques. For example, many applications in computer vision use the feature learning/extraction capabilities of a convolutional neural network, and do not apply any ***feature engineering*** step.

When ***feature engineering*** is applied, it should identify the relevant functional and operational parameters from the input space that are necessary to support the ML model training.

Objective DM-10: When applicable, the applicant should define and document the transformations to the pre-processed data from the specified input space into features which are effective for the performance of the selected learning algorithm.

Considering the objective, and depending on the data in the input space, different techniques could apply including:

— breaking data into multiple parts (e.g. date in the year decomposed in week number and day of the week);

— consolidating data into features that better represent some patterns for the ML model (e.g. transforming positions and time into speed, or representing geospatial latitudes and longitudes in 3 dimensions in order to facilitate normalisation).

Anticipated MOC DM-10-1: In relation with the objective, the applicant should manage the number of input variables, applying a dimensionality reduction step on the candidate features. This step aims at limiting the dimension of the feature space.

Anticipated MOC DM-10-2: In relation with the objective, the applicant should aim at removing multicollinearity between candidate features.

Data normalisation

It is to be noted that ***data normalisation*** does not apply to all ML techniques. In particular, ***data normalisation*** is not needed if the learning algorithm used for training the model is not sensitive to the scale of the input data (e.g. learning algorithms such as decision trees and random forests are not sensitive to the scale of the input data and do not require normalisation). Also, depending on the distribution of the data in the ODD, normalisation may distort the data and make it harder for the model to learn.

Objective DM-11: If the learning algorithm is sensitive to the scale of the input data, the applicant should ensure that the data is effective for the stability of the learning process.

Anticipated MOC DM-11-1: Data normalisation is one possible means of compliance with this objective. Depending on the data and the characteristics of the ODD, data normalisation could be achieved via different techniques such as:

— Min-Max normalisation: $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$

— Mean normalisation (X_{min} is replaced by the mean)

— Z normalisation (Standardisation): $X' = \frac{X - \mu}{\sigma}$

where:

$Xmin$ and $Xmax$ are the minimum and maximum values of the candidate feature respectively

μ is the mean of the candidate feature values and σ is the standard deviation of the candidate feature values

3.1.3.7. Data allocation

The ***data allocation*** corresponds to the last step of the ***data management*** process.

Objective DM-12: The applicant should distribute the data into three separate and independent data sets which will meet the specified DQRs:

— the training data set and validation data set, used during the model training;

— the test data set used during the learning process verification, and the inference model verification.

Particular attention should be paid to the independence of the data sets, in particular to that of the test data set. Particular attention should also be paid to the completeness and representativeness of each of the three data sets (as per **Objectives DM-02 and DM-13**).

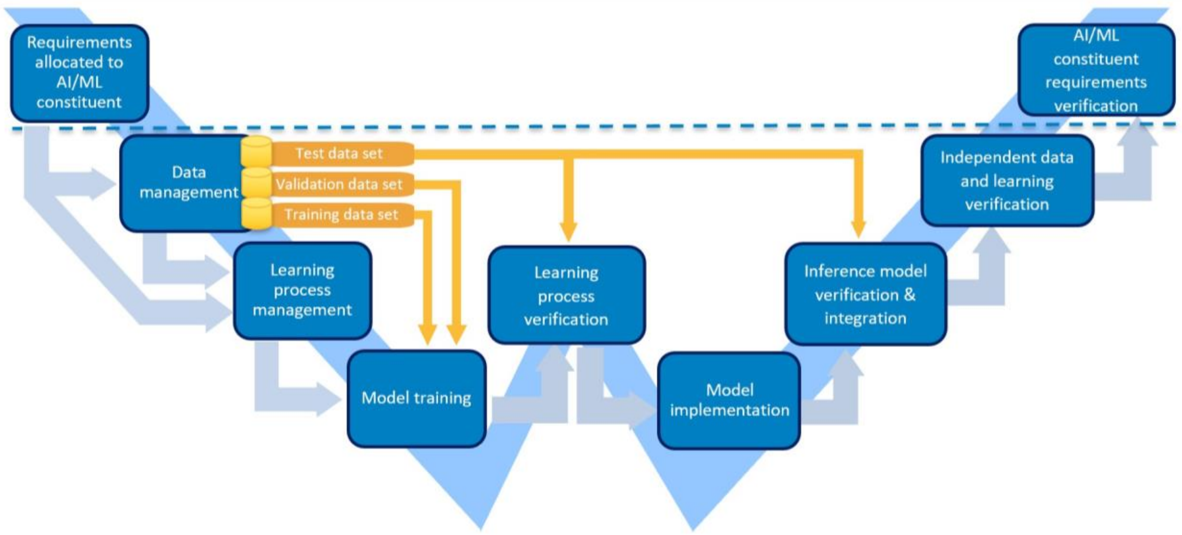


Figure 18 — Training, validation and test data sets usage in W-shaped cycle

3.1.3.8. Data validation and verification

The ***data validation*** should be ensured all along the ***data management*** process, in order to provide the training phase with data aligned with the DQRs and the other data management requirements.

Objective DM-13: The applicant should ensure validation and verification of the data, as appropriate, all along the data management process so that the data management requirements (including the DQRs) are addressed.

Focusing on the DQRs, the following represents a non-exhaustive list of anticipated MOC for a set of quality attributes which are expected for the data in the data set:

Completeness and representativeness of the data sets are prerequisites to ensure performance on unseen data and to derive generalisation guarantees for the trained model.

Anticipated MOC DM-13-1: Data completeness

The data sets should be reviewed to evaluate their completeness with respect to the set of requirements and the defined ODD.

Various methods exist to assess the completeness of the data sets (training, validation or test). For example, the input space can be subdivided into a union of hyper-cubes whose dimensions are defined by the set of operating parameters, and the number of subdivisions for each dimension, by the granularity required for the associated operating parameter.

The completeness can be analysed through the number of points contained in the hypercubes.

The scalability of such an approach may be an issue and alternatives can be considered.

It is expected that the MLEAP Horizon Europe research project will deliver additional means of compliance on completeness of the data set(s).

Anticipated MOC DM-13-2: Data representativeness

Representativeness of the data sets consists in the verification that the data they contain has been uniformly (according to the right distribution) and independently sampled from the input space. There exist multiple methods to verify the representativeness of data sets according to a known or unknown distribution, stemming from the fields of statistics and ML.

To avoid the pitfalls of a posteriori justification or confirmation bias, it is important to first determine requirements to select and verify the chosen technique(s).

For parameters derived from operating parameters (e.g. altitude, time of day) or low-dimensional features from the data (e.g. image brightness), different statistical methods (e.g. Z-test, Chi-square test, Kolmogorov-Smirnov test) may apply to assess the goodness of fit of distributions.

However, considering only such parameters for high-dimensional spaces such as images might be too shallow, and techniques applying on images or other high-dimensional data might be necessary. For example, it is impossible to codify all possible sets of backgrounds on images.

There exist multiple methods adapted to high-dimensional data, sometimes by reducing to low-dimensional spaces. One of them is the distribution discriminator framework discussed in (Daedalean, 2020). A generic representativeness/completeness verification method is viewed as function D taking as input data sets, and returning a probability of them being in-distribution. Two opposite requirements must then hold:

(1) The probability of D evaluated on in-distribution data sets is high.

(2) The probability of D evaluated on out-of-distribution data sets is low.

The exact verification setting is to be determined depending on the required statistical significance and use case, but the framework remains method- and data-agnostic. Moreover, it is meant to allow easy verification as only in- or out-of-distribution (unannotated) data is required.

It is expected that the MLEAP Horizon Europe research project will deliver additional means of compliance on representativeness of the data set(s).

Anticipated MOC DM-13-3: Data accuracy, correctness

In order to achieve correctness of the data, different types of errors and bias should be identified before unwanted bias in data sets is eliminated, and variance of data is controlled.

Errors and bias include:

— errors already present in the sourced data (e.g. data collected from databases or data lakes with residual errors or missing data);

— errors introduced by sensors (e.g. bias introduced by different cameras for the design and operational phases in the case of image recognition);

— errors introduced by collecting data from a single source;

— errors introduced by any sampling which could be applied during data collection from the data source;

— errors introduced by the human or tools when performing data cleaning or removal of presupposed outliers;

— annotation errors, especially when such an activity is performed manually by an annotation team.

Anticipated MOC DM-13-4: Data traceability

The applicant should establish an unambiguous traceability from the data sets to the source data, including intermediate data. Each operation should be shown to be reproducible.

Note: Traceability is of particular importance when data is cleaned during ***data pre-processing*** or is transformed as per the ***feature engineering*** activities.

Anticipated MOC DM-13-5: Data sets independence

The applicant should ensure that the training, validation and test data sets are verified against the independence requirements set in the DQRs.

Depending on the criticality of the AI application, more stringent requirements should be allocated to the independence of the test data set.

For highly critical applications, the applicant should ensure that the test data set is allocated independently from the training and validation data sets. That is to say, the test data set should have no common data point with the training and validation data in the corresponding data sets. The test data set should be ideally collected from real data, complemented by synthetic data where appropriate (e.g. data at the limit or beyond flight envelope).

3.1.4. Learning process management

The ***learning process management*** considers the preparatory step of the formal training phase.

Objective LM-01: The applicant should describe the AI/ML constituents and the model architecture.

Anticipated MOC LM-01-1: The applicant should describe AI/ML constituents and model (computational graph) architecture in the planning documentation, including the activation functions.

Objective LM-02: The applicant should capture the requirements pertaining to the learning management and training processes, including but not limited to:

— model family and model selection;

— learning algorithm(s) selection;

— cost/loss function selection describing the link to the performance metrics;

— model bias and variance metrics and acceptable levels;

— model robustness and stability metrics and acceptable levels;

— training environment (hardware and software) identification;

— model parameters initialisation strategy;

— hyper-parameters and parameters identification and setting;

— expected performance with training, validation and test data sets.

Anticipated MOC LM-02-1: The applicant should describe the selection and validation of the requirements for the learning management and training processes in the planning documentation. The acceptable levels for the various metrics are to be defined and documented by the applicant and generally depend on the use case. In particular for the model stability metrics, the level of the perturbation should be representative of the ODD.

Objective LM-03: The applicant should document the credit sought from the training environment and qualify the environment accordingly.

Objective LM-04: The applicant should provide quantifiable generalisation guarantees.

Anticipated MOC LM-04-1: The field of statistical learning theory (SLT) offers means to provide guarantees on the capability of generalisation of ML models. As introduced in the CoDANN report (Daedalean, 2020) Section 5.3.3, ensuring guarantees on the performance of a model on unseen data is one of the key goals of the field statistical learning theory. This is often related to obtaining 'generalisation bounds' or 'measuring the generalisation gap', that is the difference between the performance observed during development and the one that can be guaranteed during operations. The seminal work of Vapnik and Chervonenkis (On the Uniform Convergence of Relative Frequencies of Events to their Probabilities, 1971) established a relation of the generalisation capability of a learning algorithm with its hypothesis space complexity. Various forms of such bounds have been derived since then.

A good generalisation guarantee means that the 'in-sample errors' (i.e. the errors computed during the design phase) should be a good approximation of the 'out-of-sample errors' (i.e. the errors computed during the operations of the AI-based system). The generalisation gap of a model $\hat{f}$ with respect to an error metric m and a data set D_{train} can be defined as:

$$G(\hat{f}, D_{train}) = \left|E_{out}(\hat{f}, m) - E_{in}(\hat{f}, D_{train}, m)\right|$$

where:

$\chi\ is\ the\ input\ space,$
$E_{in}\ is\ the\ in-sample\ error\ (training\ error\ of\ the\ model)\ ,$
$E_{out}\ is\ the\ out-of-sample\ error\ (expected\ operational\ error),$
$D_{train}\ is\ the\ training\ dataset\ sampled\ from\ \chi,$
$\hat{f}\ is\ the\ model\ trained\ using\ D_{train},$
$m\ is\ the\ error\ metric\ used\ to\ compute\ the\ errors.$

A generalisation bound is by nature a probabilistic statement with the probability taken over possible data sets of a fixed size drawn from the input space χ. Because of this, such bounds usually output a probability tolerance δ for some given generalisation gap tolerance Ɛ:

$$P_{D_{train}\sim\chi}\left(G\left(\hat{f}, D_{train}\right) < \varepsilon\right) > 1 - \delta$$

where:

$$\varepsilon\ is\ the\ generalisation\ gap\ tolerance,$$
$$\delta\ is\ the\ probability\ tolerance.$$

Such bounds can be dependent on the properties of the data or on the learning algorithm, with the amount of data typically tightening the generalisation gap and the model complexity loosening it. For example, VC dimension-based bounds give (in the above setting):

$$G\left(\hat{\hat{f}}, D_{train}\right) < \sqrt{\frac{d_{vc} \cdot \log\left(\frac{2|D_{train}|}{d_{vc}}\right) + \log\left(\frac{1}{\delta}\right)}{|D_{train}|}}$$

where:

$$d_{vc}\ is\ the\ VC - dimension\ of\ the\ model\ family$$

Other techniques like model compression can be used to reduce model complexity and also can help in obtaining stronger generalisation bounds (refer to (Stronger generalization bounds for deep nets via a compression approach., 2018)).

Based on the CoDANN report (Daedalean, 2020), it appears that, in the current state of knowledge, the values of the generalisation upper bounds obtained for large models (such as neural networks) are often too large without an unreasonable amount of training data. It is however not excluded that applicants could rely on such approaches with sharper bounds in a foreseeable future.

In the meantime, generalisation bounds not depending on model complexity can be obtained during the testing phase (refer to (Kenji Kawaguchi, 2018)). The drawback is that this requires the applicant to have a large test data set in addition to the training data set.

The refinement of this anticipated MOC is expected to benefit from the MLEAP project deliverables.

3.1.5. Training and learning process validation

The ***training*** consists primarily in applying the learning algorithm in the conditions defined in the previous step (typically an optimisation process for the weights of a defined architecture), using the training data set originating from the ***data management*** process step. Once trained, the model performance is evaluated, using the validation data set. Depending on the resulting performance, new training iteration with a different set of model hyperparameters or even a different model type is considered, as necessary. The ***training phase and its validation*** can be repeated iteratively until the trained model reaches the expected performance.

Objective LM-05: The applicant should document the result of the model training.

Anticipated MOC LM-05-1: The records should include the training curves for the cost/loss functions and for the error metrics.

The model performance with the validation data sets should also be recorded, linking this evaluation to the metrics defined under **Objective SA-01**.

Objective LM-06: The applicant should document any model optimisation that may affect the model behaviour (e.g. pruning, quantisation) and assess their impact on the model behaviour or performance.

Anticipated MOC LM-06-1: This step may need to be performed to anticipate the inference model implementation step (e.g. embedded hardware limitations). Any optimisation that can impact the behaviour of the model is to be addressed as part of the model training and validation step. This objective only applies to optimisations performed after the model training is finished.

Objective LM-07: The applicant should account for the bias-variance trade-off in the model family selection and should provide evidence of the reproducibility of the training process.

Anticipated MOC LM-07-1: The model family bias and variance should be evaluated. The selection should aim for a model family whose complexity is high enough to minimise the bias, but not too high to avoid high variance, in order to ensure reproducibility.

The applicant should identify methods to provide the best possible estimates of the bias and variance of the selected model family; for instance, using random resampling methods (e.g. 'Bootstrapping' or 'Jack-knife').

Regularisation is a typical method to avoid overfitting (high variance) with complex models like neural networks.

Objective LM-08: The applicant should ensure that the estimated bias and variance of the selected model meet the associated learning process management requirements.

Anticipated MOC LM-08-1: For the selected model, bias is measured as the mean of the 'in sample error' (E_{in}), and variance is measured by the statistical variance of the 'in sample error' (E_{in}).

The applicant should analyse the errors on the training data set to identify and mitigate systematic errors.

3.1.6. Learning process verification

The ***learning process verification*** consists then in the evaluation of the trained model performance using the test data set. Any shortcoming in the model quality can lead to the need to iterate again on the data management process step or learning process management step, e.g. by correcting or augmenting the data set, or updating learning process settings. It is important to note that such an iteration may invalidate the test data set and lead to the need to create a new independent test data set.

Objective LM-09: The applicant should perform an evaluation of the performance of the trained model based on the test data set and document the result of the model verification.

Anticipated MOC LM-09-1: The final performance with the test data set should be measured and fed back to the safety assessment process, linking this evaluation to the metrics defined under the **Objective SA-01** and explaining any divergence in the metrics compared to the ones used to fulfil **Objective LM-04**.

Objective LM-10: The applicant should perform a requirements-based verification of the trained model behaviour and document the coverage of the AI/ML constituent requirements by verification methods.

Anticipated MOC LM-10-1: Requirements-based testing methods are recommended to reach this objective, focusing on the learning management process requirements (per **Objective LM-02**) and the subset of requirements allocated to the AI/ML constituent (per **Objective DA-02**) which can be verified at the level of the trained model. In addition, an analysis should be conducted to confirm the coverage of all requirements by test cases.

Objective LM-11: The applicant should provide an analysis on the stability of the learning algorithms.

Anticipated MOC LM-11: As outlined in (Daedalean, 2020) Section 6.4.1, perturbations in the design phase due to fluctuations in the training data set (e.g. replacement of data points, additive noise or labelling errors) could be a source of instability. Other sources may also be considered such as random initialisation of the model, optimisation methods or hyperparameter tuning. Managing the effects of such perturbations will support the demonstration of the learning algorithm stability and of the learning process repeatability.

Objective LM-12: The applicant should perform and document the verification of the stability of the trained model.

Anticipated MOC LM-12-1: The notion of trained model stability is covered through verification cases addressing anticipated perturbations in the operational phase due to fluctuations in the data input (e.g. noise on sensors) and having a possible effect on the trained model output.

This activity should address the verification of the trained model stability throughout the ML constituent ODD, including:

— nominal cases (for all equivalence classes);

— boundary cases (for all singular points).

Objective LM-13: The applicant should perform and document the verification of the robustness of the trained model in adverse conditions.

Anticipated MOC LM-13-1: The activity should be supported by test cases, including edge or corner cases within the ODD (e.g. weather conditions like snow, fog for computer vision).
In addition, two additional sets of test cases should be considered:

— OoD test cases;

— 'adversarial' test cases consisting in defining cases that are not based on the requirements but that may affect the AI/ML constituent expected behaviour.

The use of formal methods is anticipated to be a promising means of compliance with this objective, although in the current state of research those methods appear to be limited to local evaluations.

Formal methods could, for example, be used for identifying 'adversarial' test cases. Recent tools that are based on optimisation algorithms (e.g. MILP) could be used to mimic an adversary searching for an input attacking the ML model. Once identified, these 'adversarial' inputs could be added to the collected data set, so that the ML model is retrained on an augmented data set to increase its robustness.

The refinement of this anticipated MOC is expected to benefit from the MLEAP project deliverables.

Objective LM-14: The applicant should verify the anticipated generalisation bounds using the test data set.

Anticipated MOC LM-14-1: Evidence of the validity of the anticipated generalisation guarantees proposed to fulfil **Objective LM-04** should be recorded.

The refinement of this anticipated MOC is expected to benefit from the MLEAP project deliverables.

3.1.7. Model implementation

The implementation phase starts with the ***requirements capture***.

Objective IMP-01: The applicant should capture the requirements pertaining to the implementation process.

Anticipated MOC IMP-01: Those requirements include but are not limited to:

— AI/ML constituents requirements pertaining to the implementation process (C.3.1.2.1);

— requirements originating from the learning requirements capture (C.3.1.4), such as the expected performance of the inference model with the test data set;

— data processing requirements originating from the data management process (C.3.1.3.1);

— requirements pertaining to the conversion of the model to be compatible with the target platform;

— requirements pertaining to the optimisation of the model to adapt to the target platform resources;

— requirements pertaining to the development of the inference model into software and/or hardware items, such as processing power, parallelisation, latency, WCET.

The ***implementation*** then consists in transforming the trained model into an executable model that can run on certain target platform (including the compilation or synthesis/PAR steps). This implementation follows different steps:

— Model conversion

— Model optimisation

— Inference model development

Objective IMP-02: Any post-training model transformation (conversion, optimisation) should be identified and validated for its impact on the model behaviour and performance, and the environment (i.e. software tools and hardware) necessary to perform model transformation should be identified.

3.1.7.1. Trained model conversion

One of the first activities after the learning process is the freezing of the model. The trained model is represented in formats specific to the framework on which it is trained. This conversion needs to be applied to the trained model in order to obtain a representation that is compatible with the target platform. This step is the procedure of removing graph components that are not required during inference, as well as making changes that reduce the graph size and complexity without impacting the model behaviour and performance.

For example, since weights will not be updated any longer after training, gradients can be safely removed, the weight variables turned into constants and any other metadata that is relevant for training deleted. The result is a subset of the original training graph, where only the graph components that are required by the inference environment are kept, as captured in the set of requirements pertaining to implementation allocated to the AI/ML constituent.

Another conversion activity is the conversion of the model into an open format. The format in which frozen models are saved and restored is likely to be different between the learning and inference environment essentially due to the difference of framework.

Anticipated MOC IMP-02-1: Identification of the different conversion steps and confirmation that no impact on the model behaviour is foreseen. In addition, the applicant should describe the environment for each transformation step, and any associated assumptions or limitations should be captured and validated.

3.1.7.2. Trained model optimisation

In the scope of the implementation, allowable optimisations are the ones that do not affect the behaviour or performance of the trained model. Alternatively, those optimisations affecting the behaviour or performance of the trained model, shall be fed back to the learning management process (refer to **Objective LM-06**) to ensure that it is addressed through the learning process verification.

A list of possible optimisations allowable during the implementation phase includes:

— Choice of the arithmetic (e.g. fixed point format)

— Winograd algorithms for convolution: these algorithms are targeting high-performance inference. Their efficiency comes from the reduction of the number of multiplication operations due to linear and Fourier transforms.

Anticipated MOC IMP-02-2: Identification of the different optimisation steps performed during implementation and confirmation that no impact on the model behaviour is foreseen. In addition, the applicant should describe the environment for each transformation step, and any associated assumptions or limitations should be captured and validated.

3.1.7.3. Inference model development

Once confirmed that the transformations of the trained model had no impact, the last step that could impact its behaviour or performance is the implementation of the inference model into software and/or hardware items.

Objective IMP-03: The applicant should plan and execute appropriate development assurance processes to develop the inference model into software and/or hardware items.

Anticipated MOC IMP-03-1:

— For software aspects, it is anticipated that the provisions of applicable software development assurance guidance (e.g. AMC 20-115D for product certification projects) would provide the necessary means to confirm that **Objective IMP-03** is fulfilled. This guidance may need to be complemented to address specific issues linked to the implementation of an ML model into software, such as memory management issues.

— For hardware aspects, it is anticipated that the provisions of applicable hardware development assurance guidance (e.g. AMC 20-152A for product certification projects) would provide the necessary means to confirm that **Objective IMP-03** is fulfilled. FPGAs, ASICs and COTS architectures are covered by the existing guidance; however, other ML architectures, such as graphics processing units (GPUs), have specificities that are not accounted for in the existing guidance (e.g. very complex interference mechanisms or non-deterministic pipelining).

— For multicore processor (MCP) aspects, it is anticipated that the provisions of applicable MCP development assurance guidance (e.g. AMC 20-193 for product certification projects) would provide the necessary means to confirm that **Objective IMP-03** is fulfilled.

3.1.8. Inference model verification and integration

The ***inference model verification*** aims at verifying that the inference model behaves adequately compared to the trained model, in evaluating the model performance with the test data set, explaining any difference in the evaluation metric compared to the one used in the ***training phase verification*** (e.g. execution time metrics). This process step should also foresee verification that the model properties have been preserved (e.g. based on the implementation analysis or through the use of formal methods).

The ***inference model integration*** within the associated AI/ML constituent and (sub)system implies several steps of integration, as many as considered necessary to support adequate verification; an important one being the integration of the AI/ML constituent with the target platform, together with the other AI-based subsystem items (in particular with the sensors).

3.1.8.1. Verification of inference model properties preservation

Objective IMP-04: The applicant should verify that any transformation (conversion, optimisation, inference model development) performed during the trained model implementation step has not adversely altered the defined model properties.

Anticipated MOC IMP-04-1: First a set of model properties that are expected to be preserved should be captured. In addition, the performance metrics for this verification and the associated acceptable bounds of variation should be documented. The use of specific verification methods (e.g. formal methods) is expected to be necessary to comply with this objective.

3.1.8.2. Platform verification

Objective IMP-05: The differences between the software and hardware of the platform used for training and those used for the inference model verification should be identified and assessed for their possible impact on the inference model behaviour and performance.

Anticipated MOC IMP-05-1: The analysis of the differences, such as the ones induced by the choice of mathematical libraries or ML framework, is an important means to reach this objective. This objective does not apply when the complete verification of the ML model properties is performed with the inference model on the target platform.

3.1.8.3. Inference model verification

Objective IMP-06: The applicant should perform an evaluation of the performance of the inference model based on the test data set and document the result of the model verification.

Anticipated MOC IMP-06-1: The final performance with the test data set should be measured and fed back to the safety assessment process, linking this evaluation to the metrics defined under the **Objective SA-01** and explaining any divergence in the metrics compared to the ones used to fulfil **Objective LM-09**.

Objective IMP-07: The applicant should perform and document the verification of the stability of the inference model.

Anticipated MOC IMP-07-1: The notion of inference model stability is covered through verification cases addressing anticipated perturbations in the operational phase due to fluctuations in the data input (e.g. noise on sensors) and having a possible effect on the inference model output.

This activity should address the verification of the inference model stability throughout the ML constituent ODD, including:

— nominal cases (for all equivalence classes);

— boundary cases (for all singular points).

Objective IMP-08: The applicant should perform and document the verification of the robustness of the inference model in adverse conditions.

Anticipated MOC IMP-08-1: The activity should be supported by test cases, including edge or corner cases within the ODD (e.g. weather conditions like snow, fog for computer vision) and OoD test cases.

The refinement of this anticipated MOC is expected to benefit from the MLEAP project deliverables.

3.1.8.4. Inference model integration into the AI/ML constituent

Objective IMP-09: The applicant should perform a requirements-based verification of the inference model behaviour when integrated into the AI/ML constituent and document the coverage of the AI/ML constituent requirements by verification methods.

Anticipated MOC IMP-09-1: Requirements-based testing methods are necessary to reach this objective, focusing on the requirements pertaining to the implementation (per **Objective IMP-01**) as well as all requirements allocated to the AI/ML constituent (per **Objective DA-02**). In addition, an analysis should be conducted to confirm the coverage of all requirements by verification cases.

The test environment should at least foresee:

— the AI/ML constituent integrated on the target platform (environment #1),

— the AI/ML constituent integrated in its subsystem, with representative interfaces to the other subsystems, including to the directly interfacing sensors (environment #2).

3.1.9. Data and learning verification

The ***data verification*** step is meant to close the data management life cycle, by verifying with independence that data sets were adequately managed, considering that the verification of the data sets can be achieved only once the inference model has been satisfactorily verified on the target platform. It is important to mention however that this does not imply waiting for the end of the process to initiate this step, considering the highly iterative nature of learning processes.

Objective DM-14: The applicant should perform a data and learning verification step to confirm that the appropriate data sets have been used for the training, validation and verification of the model and that the expected guarantees (generalisation, robustness) on the model have been reached.

Anticipated MOC DM-14: The associated activities include:

— independent verification that the data sets (training, validation, test) comply with the data management requirements;

— independent verification of the correct identification of the input space, including a reassessment of the defined ODD;

— independent verification that the data sets (training, validation, test) are complete and representative of the input space of the application;

— independent verification that the expected guarantees (generalisation, robustness) on the model have been reached.

Note 1: The level of independence should be commensurate with the safety criticality of the application.

Note 2: This independent verification step may be requested only for higher-criticality levels.

3.1.10.Verification of (sub)system requirements allocated to the AI/ML constituent

The ***requirements verification*** is addressing the verification of the AI/ML constituent fully integrated in the overall system. It is considered to be covered by traditional assurance methodologies (e.g. ED-79A/ARP4754A).

Objective DA-07: Each of the captured (sub)system requirements allocated to the AI/ML constituent should be verified**.**

3.1.11.Configuration management

The ***configuration management*** is an integral process to the development of an AI/ML constituent.

Objective CM-01: The applicant should apply all configuration management principles to the AI/ML constituent life-cycle data, including but not limited to:

— identification of configuration items;

— versioning;

— baselining;
— change control;
— reproducibility;
— problem reporting;
— archiving and retrieval, and retention period.

Anticipated MOC CM-01-1: The collected data, the training, validation, and test data sets used for the frozen model, as well as all the tooling used during the transformation of the data are to be managed as configuration items.

3.1.12. Quality and process assurance

Quality and process assurance is an integral process that aims at ensuring that the life-cycle process objectives are met, and the activities have been completed as outlined in plans (as per **Objective DA-01**) or that deviations have been addressed.

Objective QA-01: The applicant should ensure that quality/process assurance principles are applied to the development of the AI-based system, with the required independence level.

3.2. Development & post-ops AI explainability

Development & post-ops AI explainability is driven by the needs of stakeholders involved in the development cycle and the post-operational phase. The figure below shows the scope of development & post-ops AI explainability.

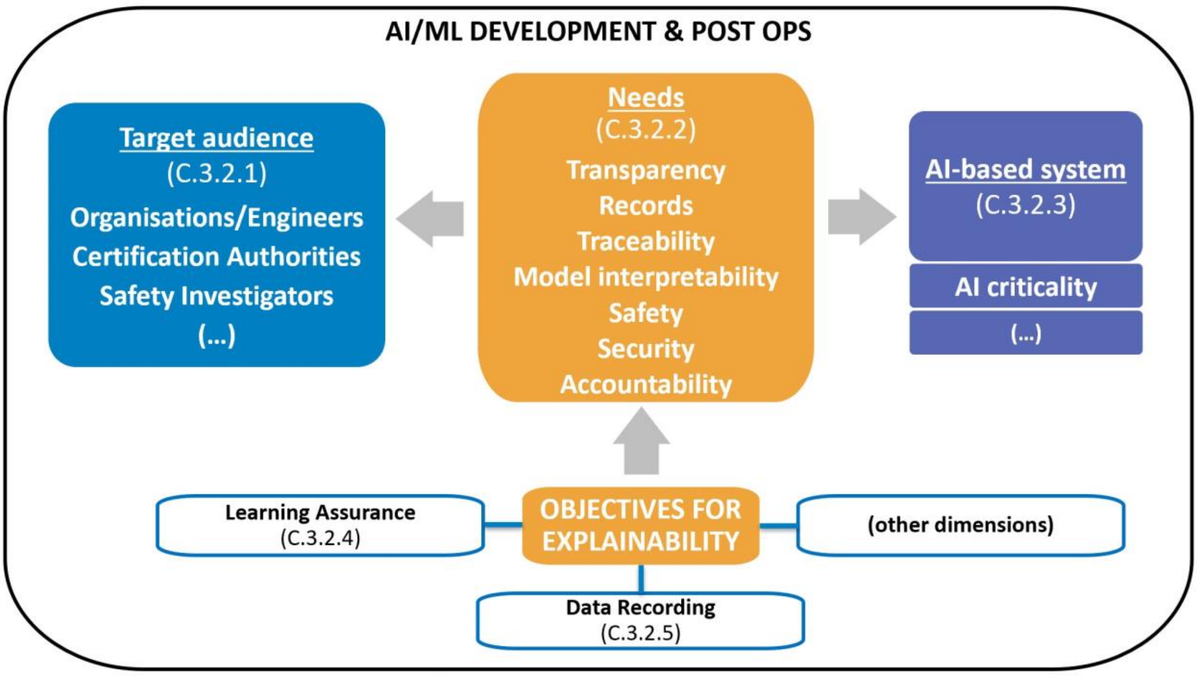


Figure 19 — Development & post-ops explainability view

3.2.1. Target audience for development & post-ops AI explainability

The need for a deep insight into AI-based system explainability concerns a wide range of stakeholders. These include at least the engineers (e.g. applicant, system designer, end-developers, users, etc.), the certification authorities, and the safety investigators.

3.2.2. Need for development & post-ops AI explainability

In addition to the needs already addressed via the learning assurance or the trustworthiness analysis (e.g. safety assessment), these stakeholders typically express needs for a deeper level of insight in the design details of the AI-based system.

3.2.3. Anticipated development & post-ops AI explainability modulation

As anticipated in the introduction of this document (Chapter B), the proportionality of guidance can be influenced from at least two different angles:

— the AI level as an outcome of **Objective CL-01** with the classification of the AI-based system, based on the levels presented in Table 2; and

— the criticality allocated to the AI-based system.

The development & post-ops explainability guidance is anticipated to be necessary for all AI levels (1 to 3); therefore, the modulation of objectives in Section C.3.2 is expected to be driven mainly by criticality.

3.2.4. Objectives for development & post-ops AI explainability

This section proposes a series of objectives related to AI explainability.

Learning assurance is a prerequisite to ensure confidence in the performance and intended function of ML-based systems. Without this confidence, AI explainability is impractical. Learning assurance is therefore considered as one of the fundamental elements for developing explainability.

The set of objectives developed in this section intend to clarify the link between learning assurance and development/post-ops explainability, by providing a framework for reaching an adequate level of transparency on the ML model. The associated explainability methods will support the objectives of learning assurance from Section C.3, and the objectives of the operational explainability developed in Section C.4.1 below.

It is acknowledged, however, that the learning assurance W-shaped process may not necessarily provide sufficient level of transparency on the inner design of the ML model (in particular for complex models such as NNs).

Identification of relevant stakeholders

Objective EXP-01: The applicant should identify the list of stakeholders, other than end users, that need explainability of the AI-based system at any stage of its life cycle, together with their roles, their responsibilities and their expected expertise (including assumptions made on the level of training, qualification and skills).

Note: This objective focuses on the list of stakeholders other than the end users, as these have been identified already as per **Objective CO-01**.

Identification of need for explainability

Objective EXP-02: For each of these stakeholders (or groups of stakeholders), the applicant should characterise the need for explainability to be provided, which is necessary to support the development and learning assurance processes.

Anticipated MOC EXP-02: The need for explainability should at least support the following goals:

— Strengthening the input-output link;

— Detection of residual biases in the trained and/or inference model; and

— Absence of unintended functions.

Object of the explanation

When dealing with development & post-ops explainability, the object of the explanation could be either:

— the ML item itself (a priori/global explanation);

— an output of the ML item (post hoc/a posteriori/local explanation).

It must be made clear which item is being referred to and what the requirements of explainability are for each of them. Explanations at ML item level will be focused on the stakeholders involved during development & post operations, whereas explanations on the output of an ML item could be useful for all stakeholders, including end users in the operations. Output-level explanations can be simpler/more transparent and therefore accessible to non-AI/ML experts like end user communities.

The AI explainability methods necessary to fulfil the development explainability requirements can be further grouped in two different objectives:

— item-level; and

— output-level explanations.

At this stage, this split is used to distinguish two anticipated MOC for item-level and output-level explanations.

Objective EXP-03: The applicant should identify and document the methods at AI/ML item and/or output level satisfying the specified AI explainability needs.

Anticipated methods both for the item level and output level explainability can be found in the Innovation Partnership Contract CODANN2 (Daedalean, 2021). Item-level explainability methods for CNNs include filters visualisations, generative methods and maximally activating inputs. For output-level explanations, methods include local approximation, activations visualisation and saliency maps. This material is illustrative at this point in time, as it applies particularly to computer vision types of applications using CNNs. These will evolve with the progress of research and standardisation efforts.

Note: The methods pertaining to this **Objective EXP-03** may be used also to support the objectives related to operational explainability as developed in Section C.4.1.

Explainability at item level or output level is a key area for current research. It is therefore expected that best practices and techniques will emerge, which will enable additional objectives or anticipated MOC to be developed.

3.2.5. Specific objectives for AI data recording capability

To support the general development and post-ops explainability objectives, specific objectives related to the collection of data are defined in this section.

With regard to the recording of data for the purpose of development and post-operation assessment, at least two distinct types of use should be addressed:

— Data recording for the purpose of monitoring the safety of AI-based system operations (as part of safety management and/or continued operation approval)

  - This monitoring consists in recording and processing data from day-to-day operation to detect and evaluate deviations from the expected behaviour of the AI-based system, as well as issues affecting interactions with human users or other systems.
  - This monitoring is usually performed by (or on behalf of) the organisation using the AI-based system.
  - The purpose of this monitoring is to support the continuous or frequent assessment of the safety of the operations in which the AI-based system is used and to assess whether mitigation actions are effective.
  - This monitoring is meant to be part of the safety management system (SMS) of the organisation using the AI-based system.
  - This monitoring may also serve the purpose of continued operation approval, by providing the designer team of the AI-based system with data to monitor the in-service performance of the system.

— Data recording for the purpose of accident or incident investigation in line with ICAO Annex 13 and Regulation (EU) 996/2010

  - This recording is meant for analysing an accident or incident for which the operation of the AI-based system could have been a contributing factor.
  - There are many kinds of accident or incident investigations (internal investigation, judicial investigation, assurance investigation, etc.) but in this document, only the official safety investigation (such as defined in ICAO Annex 13 and Regulation (EU) 996/2010) is considered. An official safety investigation aims at preventing future incidents and accidents, not at establishing responsibilities of individuals.
  - The recorded data is used, together with other recordings, to accurately reconstruct the sequence of events that resulted in the accident or serious incident.

Notes:

- It is not forbidden to address these two types of use with a single data recording solution.
- The recording of data does not need to be a capability of the AI-based system. It is often preferable that the relevant data is output for recording to a dedicated recording system.

Objective EXP-04: The applicant should provide the means to record operational data that is necessary to explain, post operations, the behaviour of the AI-based system and its interactions with the end user.

3.2.5.1. Start and stop logic for the data recording (applicable to both types of use)

Anticipated MOC EXP-04-1: The recording should automatically start before or when the AI-based system is operating, and it should continue while the AI-based system is operating. The recording should automatically stop when or after the AI-based system is no longer operating.

3.2.5.2. Data recording for the purpose of monitoring the safety of AI-based system operations

Anticipated MOC EXP-04-2: The recorded data should contain sufficient information to detect deviations from the expected behaviour of the AI-based system, whether it operated alone or interacting with an end user. In addition, this information should be sufficient:

(a) to accurately determine the nature of each individual deviation, its time and the amplitude/severity of that individual deviation (when applicable);

(b) to reconstruct the chronological sequence of inputs to and outputs from the AI-based system before and during the deviation;

(c) for monitoring trends regarding deviations over longer periods of time.

Anticipated MOC EXP-04-3: The recorded data should be made available to those entitled to access and use it in a way so that they can perform an effective monitoring of the safety of AI-based system operations. This includes:

(a) timely and complete access to the data needed for that purpose;

(b) access to the tools and documentation necessary to convert the recorded data in a format that is understandable and appropriate for human analysis;

(c) possibility to gather data over longer periods of time for trend analyses and statistical studies. In any case, the data should be retained for a minimum of 30 days.

3.2.5.3. Data recording for the purpose of accident or incident investigation

Anticipated MOC EXP-04-4: The recorded data should contain sufficient information to accurately reconstruct the operation of the AI-based system and its interactions with the end user before an accident or incident. In particular, this information should be sufficient to:

(a) accurately reconstruct the chronological sequence of inputs to and outputs from the AI-based system;

(b) identify when communication or cooperation/collaboration between the AI-based system and the end user was degraded. This may require recording additional communications of the end user with other team members or with other organisations (including voice communications), or recording additional actions performed by the end user at their workstation (for instance, by means of images), as necessary;

(c) identify any unexpected behaviour of the AI-based system that is relevant for explaining the accident or incident.

Anticipated MOC EXP-04-5: The data should be recorded in a way so that it can be retrieved and used after an accident or an incident. This includes:

(a) crashworthiness of the memory media if they could be exposed to severe environmental conditions resulting from an accident;

(b) recording technology that is reliable and capable of retaining data for long periods of time without electrical power supply;

(c) means to facilitate the retrieval of the data after an accident (e.g. means to locate the accident scene and the memory media, tools to retrieve data from damaged memory media) or an incident;

(d) provision of tools and documentation necessary to convert the recorded data in a format that is understandable and appropriate for human analysis.

4. Human factors for AI

The objectives developed in this section provide initial human factors guidance to applicants in order to design an AI-based system and equipment for use by the end users.

Note on the status of human-factors-related guidance:

— For Level 1A, existing guidelines and requirements for interface design should be used.

— For Level 1B, an initial set of design principles are proposed for the concept of operational explainability.

For Level 2A and Level 2B, new objectives have been developed and others from existing human factors certification requirements and associated guidance have been adapted to account for the specific end-user needs linked to the introduction of AI-based systems.

Background on the existing human-factors-related regulatory framework and guidance for flight deck design

CS 25 has contained certification specifications for flight deck design for large aeroplanes since Amendment 3. CS 25.1302 requires applicants to design the flight deck considering a comprehensive set of design principles that are very close to what is described in the literature under the concept of usability. The ultimate intent of designing a usable flight deck is to prevent, as much as possible, the occurrence of flight crew errors while operating the aircraft. It aims at preventing any kind of design-related human performance issue.

On top of it, CS 25.1302 also requires that the operational environment (flight deck design, procedures and training) allows efficient management of human errors, should they occur despite the compliance of the flight deck with the usability principles. CS 25.1302 (a), (b) and (c) intend to reduce design contribution to human error by improving general flight deck usability while CS 25.1302 (d) focuses on the need to support human error management through design to avoid safety consequences. The same requirement exists for rotorcrafts (CS 27 / 29.1302) and as a Special Condition for gas airships (SC GAS) and for VTOL aircraft (SC VTOL).

AMC 25.1302 provides recommendations including design guidance and principles as well as human factors methods to design flight deck for future certification. The requirements and guidance for flight deck design were developed for aircraft equipped initially with automation systems. The design guidance proposed in AMC 25.1302 (5) is a set of best practices agreed between EASA and industry. This part includes four main topics: Controls (means of interactions) / Presentation of information (Visual, tactile, auditory) / System Behaviour (conditions to provide information on what the system is doing) / Flight Crew Error Management (impossible to predict the probabilities of error).

CS 25.1302 and its associated AMC are considered by EASA to be a valid initial framework for the implementation of Level 1 AI-based system applications and can be used as the basis on which further human factors requirements for AI could be set for Level 2 AI-based systems.

Background on the existing human-factors-related regulatory framework and guidance for design in the ATM domain

Regulation (EU) 2017/373 lays down the common requirements for air traffic management and air navigation services. Yet, there are no requirements that specify the incorporation of human factors within the scope of equipment design or the introduction of new technology. The Regulation does contain requirements to address human factors subjects such as psychoactive substances, fatigue, stress, rostering, but these are largely outside the consideration of AI systems and cannot be used as the basis for the development of human factors AI requirements.

Further to point (1)(i) of point ATS.0R.205 'Safety assessment and assurance of changes to the functional system' of Regulation (EU 2017/373, the scope of the safety assessment for a system change as includes the 'equipment, procedural and human elements being changed'. By definition, therefore, any change impacting the functional ATM system should include an assessment of the impact on the human, but from a safety perspective, not necessarily from a human factors perspective. There are therefore currently no existing requirements that cover the entire ATM domain to which human factors requirements for AI could be attached.

In the absence of regulatory requirements on human factors in ATM/ANS, existing material should be referred to, which includes but should not be limited to, Human Performance Assessment Process (HPAP), SESAR and/or Eurocontrol - human factors case version 2.

For all of these domains, elements from the existing human factors requirements and guidance are applicable for AI-based installed systems and equipment for use by the end users. However, this guidance needs to be complemented and/or adapted to account for the specific needs linked with the introduction of AI.

Section C.4 covers the following themes through dedicated objectives:

— AI operational explainability

— Human-AI teaming

— Modality of interaction and style of interface

— Error management

— Workload management

— Failure management and alerting system

— Integration

— Customisation of human-AI interface

4.1. AI operational explainability

A clear distinction is made in this document between the explainability needed to make ML models understandable (development & post-ops AI explainability) and the need to provide end users with 'understandable' information on how the AI-based system came to its results (operational explainability).

Explainability is a concept, which, to be measurable or practically assessed, has to be operationalised. An initial set of attributes in the frame of future development and certification are proposed: understandability, relevance, level of abstraction, timeliness, and reliability. All these attributes are further developed in the objectives and anticipated MOC for the operational explainability in Section C.4.1.4.

Note: The term 'operational' in this section refers to any type of activity for which an AI-based system is used, and is not limited to air operations.

The figure below illustrates the scope proposed for operational explainability.

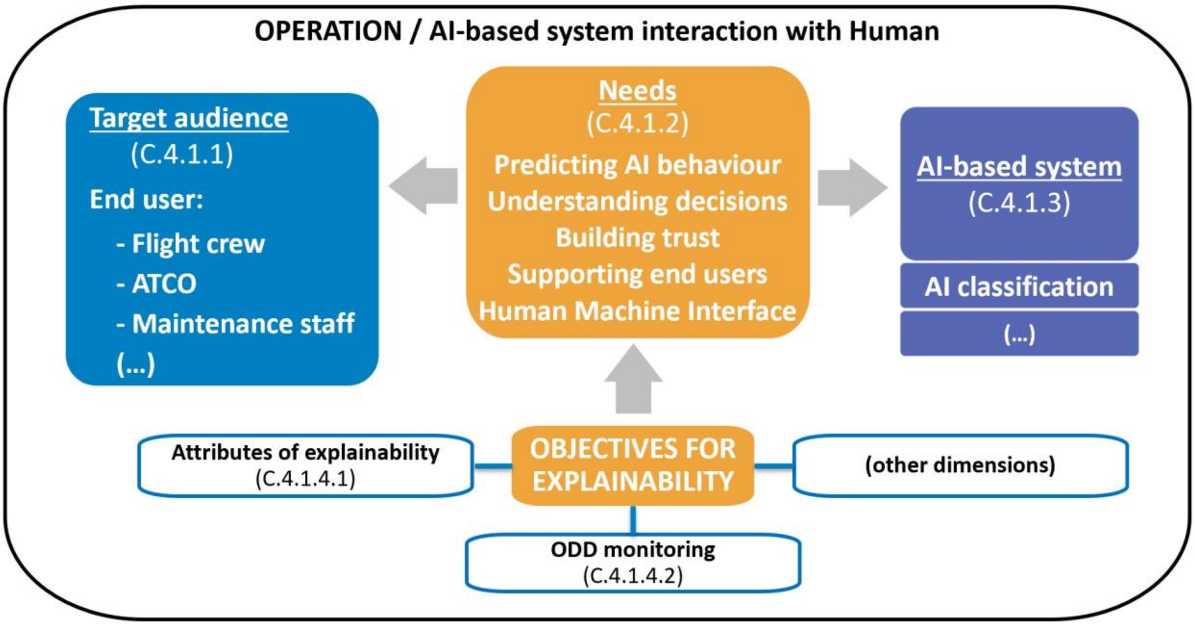


Figure 20 — Operational explainability view

4.1.1. Target audience for operational explainability

The expected target audience for operational explainability includes, but is not limited to, the crew members for airborne operations, the ATCO and the room supervisor (RSUP) for the ATM domain, and the maintenance engineer for the maintenance domain. These stakeholders are expected to have dedicated needs for explainability in order to be able to use the AI-based system, interact with it, and influence their level of trust.

4.1.2. Need for operational explainability

In operation, the introduction of AI is expected to modify the paradigm of interaction between the end user and the system. Specifically, it will affect the function allocation distribution[1] by progressively giving more authority to the AI-based systems. This allocation will be made already at the system development phase. This will lead to a reduction of end-user awareness of the logic behind the automatic decisions or actions taken by the AI-based system. This decreasing awareness may limit the efficiency of the interaction and lead to a failure in establishing trust or a potential reduction of

[1] Function allocation distribution refers to strategies for distributing system functions and tasks across people and technology. (Function Allocation Considerations in the Era of Human Autonomy Teaming, December 2019).

trust from the end user. In order to ensure an adequate efficiency of the interactions, the AI-based system will need to provide explanations with regard to its automatic decisions and actions.

Note on explainability and trust

Preliminary work examining the relationship between trust and explainability is made available below. The main consideration is that explainability is one amongst a number of contributors that build or increase the trust that the end user has in the system. It is actually a contributor to the perception people have on the trustworthiness of the AI-based system.

Indeed, explanations given through explainability could be considered as one variable among others. It is also clear that not all explanations will serve this purpose. As an example, if the explanation is warning the end user about the malfunction of the AI based system, the explanation will not positively influence the end user's trust in the system. The efficiency of an explanation in eliciting trust and improving the end user's perception that a system is trustworthy depends highly on factors such as the context, the situation, and the end user's experience and training.

The following list illustrates other possible factors that may influence the trust of the end user:

— End user's general experience, belief, mindset, and prior exposure to the system

— The maturity of the system

— The end user's experience with the AI-based system, whether the experience is positive and there is a repetition of a positive outcome

— The AI-based system knowledge on the end user's positive experience regarding a specific situation

— The predictability of the AI-based system decision and whether the result expected is the correct one

— The reinforcement of the reliability of the system through assurance processes

— The fidelity and reliability of the interaction:

- interaction will participate in end user's positive belief over the AI-based system's trustworthiness;
- weak interaction capabilities, reliability, and experience can have a strong negative impact on the belief an end user may have in the trustworthiness of the whole system. It can even force him or her to turn off the system.

4.1.3. Anticipated operational explainability modulation

It is also important to consider the AI Level of the AI-based system. The need for explainability is significantly dependent on the pattern of authority and functional allocation distribution between the end user and the AI-based system. For example, the operation of a Level 1A AI-based system will not be fundamentally different from the operation of existing systems. Therefore, there is no need to develop specific explainability mechanisms on top of the existing human factors requirements and/or guidance that are already in use (e.g. CS/AMC 25.1302 for flight deck design).

However, from Level 1B and above, there is a need to identify and characterise the importance of explainability as well as its attributes.

	OVERALL IMPACT ASSESSMENT	**HAII** Expected level of evolution in the human-AI interaction (HAII) compared to existing interactions	**EXPLAINABILITY** Expected level of explainability needed during operation	**GUIDANCE** Need for specific human factors certification guidance linked with the introduction of AI-based systems
Level 1A Human augmentation	The implementation of an AI-based system is not expected to have an impact on the current operation of the end user. e.g. Enhanced visual traffic detection/indication system in flight-deck. e.g. The analysis of aircraft climb profiles by an AI-enhanced conflict probe when checking the intermediate levels of an aircraft climb instruction.	No change compared to existing systems.	No change compared to existing systems as the implementation of an AI-based system at Level 1A is impacting neither the operation, nor the interaction that the end user has with the systems.	No need for dedicated guidance. Existing guidelines and requirements for interface design should be used. e.g. CS/AMC 25.1302
Level 1B Human assistance	The implementation of an AI-based system is expected to impact the current operation of the end user with the introduction of, for example, a cognitive assistant. e.g. Cognitive assistant that provides the optimised diversion option or optimised route selection. e.g. An enhanced final approach sequence within an AMAN	Medium change: There is a need for explainability so that the end user is in a position to use the AI outcomes to take decisions/actions.	Explainability is there to support and facilitate end-user decisions. At this level, decision still requires human judgement or some agreement on the solution method.	Specific guidance needed. Need for operationalising the explainability concept in the frame of future design and certification. → Definition of attributes of explainability with design principles.
Level 2A Human-AI teaming: Cooperation	Level 2A corresponds to the implementation of an AI-based system capable of teaming with an end user. The operation is expected to change by moving from human-human teams to human-AI-based system teams . More specifically, Level 2A is introducing the notion of cooperation as a process in which the AI-based system works to help the end user accomplish their own objective and goal. The operation evolves by taking into account the work from the AI-based system based on a predefined task allocation pattern. e.g. AI advanced assistant supporting landing phases (automatic approach configuration) e.g. conflict detection and resolution in ATM.	Medium change: Communication is not a paramount capability for cooperation. However, informative feedback on the decision and/or action implementation taken by the AI-based system is expected. HAII evolution is foreseen to account for the introduction of the cooperation process.	With the expected introduction of new ways of working with an AI-based system, the end user will require explanations in order to cooperate to help the end user accomplish their own goal. A trade-off is expected at design level between the operational needs, the level of detail given in an explanation and the end-user cognitive cost to process the information received.	Specific guidance needed Existing human factors certification requirement and associated guidance will have to be adapted for the specific needs linked with the introduction of AI. → Development of future design criteria for novel modality of interaction and style of interface as well as criteria for HAT, and criteria to define roles and tasks allocation at design level.

Level 2B HAT; Collaboration	Level 2B corresponds to the implementation of an AI-based system capable of collaboration. On top of the evolution linked to the notion of HAT, the collaboration will make the operation evolve towards a more flexible approach where the human and the AI-based system will both communicate and share strategies/ideas to achieve a common goal. e.g.: Virtual co-pilot in single-pilot operations	High change: Existing human factors certification requirements and associated guidance are adapted to the specific needs linked with the introduction of AI. → Development of design criteria for novel modality of interaction and style of interface as well as criteria for HAT, and criteria to define roles and tasks allocation at design level.	With the expected introduction of new ways of working with an AI-based system, the end user will require explanations in order to collaborate, negotiate or argument towards a common goals. A trade-off is expected at design level between the operational needs, the level of detail given in an explanation and the end-user cognitive cost to process the information received.	Specific guidance needed Existing human factors certification requirements and associated guidance will have to be adapted to the specific needs linked with the introduction of AI. → Development of future design criteria for novel modality of interaction and style of interface, criteria for HAT, and criteria to define roles and tasks allocation at design level.
Level 3A More autonomous AI	The AI-based system is operating independently with the possibility from the end user to override an action/decision only when needed. No permanent oversight from the end user. A significant modification in the current operation is expected. e.g. UAS ground end user managing several aircraft	Very high change: Expected change in the job design with evolution in HAII to support the end user being in a position to override the decision and action of the AI-based system when needed.	In order for the end user to override the AI/ML systems' decision, the appropriate level of explanation or information is going to be needed for the good operation of the system.	Specific guidance needed. On top of the specific guidance needed for Level 2, EASA anticipates additional guidance development.
Level 3B Fully autonomous AI	There is no more end user. The AI-based system is fully autonomous. e.g. Fully autonomous flights e.g. Fully autonomous sector control.	N/A: The end user is effectively removed from the process. There is no requirement for end-user interaction.	There is no need for explainability at the level of the end user. There is no end user.	N/A in operation.

Table 4 — Anticipated human factors guidance modulation

4.1.4. Objectives for operational AI explainability

4.1.4.1. Objectives related to the attributes of AI operational explainability

Given the importance that EASA attributes to AI explainability, the following objectives and anticipated MOC can be used as design principles for operational explainability.

Note: The explainability methods used to meet **Objective EXP-03** from the development/post-ops explainability may be used to meet some of the objectives below.

Objective EXP-05: For each output of the AI-based system relevant to task(s) (per **Objective CO-02**), the applicant should characterise the need for explainability.

Understandable and relevant explainability

Objective EXP-06: The applicant should present explanations to the end user in a clear and unambiguous form.

Anticipated MOC EXP-06: The explanation provided should be presented in a way that is perceived correctly, can be comprehended in the context of the end user's task and supports the end user's ability to carry out the action intended to perform the tasks.

Objective EXP-07: The applicant should define relevant explainability so that the receiver of the information can use the explanation to assess the appropriateness of the decision / action as expected.

Anticipated MOC EXP-07: The explanation should be relevant so that the receiver of the information can use it to assess the appropriateness of the decision / action as expected.

As an example, a first set of arguments that could be contained in an explanation might be:

— *Information about the goals:* The underlying goal of an action or a decision taken by an AI-based system should be contained in the explanation to the receiver. This increases the usability and the utility of the explanation.

— *Historical perspectives:* To understand the relevance of the AI-based system proposal, it is important for the receiver to get a clear overview on the assumptions and context used for training of the AI-based system.

— *Information on the 'usual' way of reasoning:* This argument corresponds to the information on the inference made by the AI-based system in a specific case, either by giving the logic behind the reasoning (e.g. causal relationship) or by providing the information on the steps and on the weight given to each factor used to build decisions.

— *Information about contextual elements:* It might be important for the end user to get precise information on what contextual elements were selected and analysed by the AI-based system when making decisions/ implementing actions. The knowledge of relevant contextual elements will allow the end user to complement their understanding and form an opinion on the decision.

— *Information on strategic aspects:* The AI-based system might be performing a potential trade-off between operational needs / economical needs / risk analysis. These strategies could be part of the explanation when needed.

— *Sources used by the AI-based system for decision-making:* This element is understood as the type of explanation given regarding the source of the data used by the AI-based system to build its decision. For example, the need in a multi-crew aeroplane for one pilot to understand which source the other pilot used in order to assess the weather information as data can come from different sources (ops/data/radar/etc.). As the values and their reliability may vary, it is fundamental that both pilots are aligned using the same sources of data.

Level of abstraction

Objective EXP-08: The applicant should define the level of abstraction of the explanations, taking into account the characteristics of the task, the situation, the level of expertise of the end user and the general trust given to the system.

Anticipated MOC EXP-08: The level of abstraction corresponds to the degree of details provided within the explanation. As mentioned before, there are different possible arguments to substantiate the explainability (ref. relevant explainability). The level of detail of these arguments and the number of arguments provided in an explanation may vary depending on several factors.

— *The level of expertise of the end user:* An experienced end user will not have the same needs in terms of rationale and details provided by the AI-based system to understand how the system came to its results, as a novice end user who might need advice or/and detailed information to be able to follow a proposition coming from the AI-based system.

— *The characteristics of the situation:* In a very time-critical situation, the end user might not have the cognitive capacity to understand and follow explanations. Indeed, a lengthy explanation will lose its efficiency in case the end user is not able to absorb it. During a non-critical situation, with a low level of workload on the side of the end user, the explanation can be enriched.

— *The general trust given to the system:* There is a link between the trust afforded to the system and the need for detailed explanation. If the end user trusts the system, they might accept an explanation with fewer details; however, an end user with low trust might request additional information to reinforce or build trust in the AI-based system and accept the decision/action.

There are advantages and disadvantages in delivering a detailed explanation. On one side, it may ensure an optimal level of understanding of the end user. However, it may generate a significant cognitive cost due to the high amount of information to process. Additionally, it may reduce the interaction efficiency in the context of a critical situation. On the other side, a laconic explanation may lead to a lack of understanding from the end user, resulting as well in a reduction of the interaction efficiency. Therefore, a trade-off between the level of details given in an explanation and the cognitive cost seems to be essential to maintain an efficient HAII.

Objective EXP-09: Where a customisation capability is available, the end user should be able to customise the level of details provided by the system as part of the explainability.

Anticipated MOC EXP-09: The level of abstraction has an impact on the collaboration between the AI-based system and the end users. In order to enhance this collaboration during operation, there is a possible need to customise the level of details provided for the explanation. This can be tackled in three ways:

— Firstly, the designer could set by default the level of abstraction depending on factors identified during the development phase of the AI.

— Secondly, the end users could customise the level of abstraction. If the level is not tailored to their needs or level of experience, the explainability can go against its objective.

— Thirdly, the level of abstraction could come from an adaptive explainability thanks to context-sensitive mechanisms. The AI-based system will have the capabilities to adapt to its environment by design or by learning (adaptive explainability).

Timeliness of explainability

Objective EXP-10: The applicant should define the timing when the explainability will be available to the end user taking into account the time criticality of the situation, the needs of the end user, and the operational impact.

Objective EXP-11: The applicant should design the AI-based system so as to enable the end user to get upon request explanation or additional details on the explanation when needed.

Anticipated MOC EXP-10 & EXP-11: The notion of timeliness depends on the end user's need and is imposed by the situation. This notion covers both the appropriate timing and the appropriate sequencing of explanations. This guidance defines two temporalities: before the operation and during the operation.

Before operation, or latent explainability

— It should be considered that the knowledge gained by the end user during training about the way an AI-based system is working will contribute to the end user's ability to decrypt the AI-based system's actions and decisions during operations. This can be considered as a latent explainability. The end users retrieve this knowledge to build their awareness and compute their own explanation and to interpret, on behalf of the AI-based system, the reason behind the system's decision and/or action/behaviour. In addition, information concerning the AI-based system customisation made by the operators/airlines to answer specific operational needs could also be provided to the end users before operation.

***During operation* —** The following trade-offs should be considered by the applicant:

— **Before the decision/action taken by the AI-based system:** Information should be provided before the decision or action in case the outcome of the decision/action has an impact on the conduct of the operation. As an example for airborne operations, if an AI-based system has the capability to lower the undercarriage, it would be necessary to provide the information to the crew before the action is performed, as it will have an impact on the aircraft performance. Another general reason could be to avoid any startle effect and provide the end user with sufficient anticipation to react accordingly to the decision/action.

— **During the decision/action:** Explanation provided during the decision and action should include information on strategic and tactical decisions. Strategic information with a long-term impact on the operation should be provided to the end user during the decision/action.

Note: The more information relates to short-term tactical approach, the more it should be provided before the decision/action. The end user will need to be aware of the steps performed by the AI-based system that will have a short-term impact on the operation.

— **After the decision/action**

Here are four different examples for explainability to be provided after the decision/action was identified:

- When there is a time-critical situation, there will be no need or benefit for the end user to get an explanation in real time.
- The explanation could come a posteriori as programmed by the applicant for any justified reason.
- The explanation is requested on-demand by the end user, either to complement their understanding, or because the end user put the AI on hold voluntarily prior to the decision/action.
- The AI-based system by design is providing the explanation after the decision/action in order to reinforce trust and update the situational awareness (SA) of the end users.

Figure 21 provides an illustration of the notion of timeliness that should be assessed when designing explainability.

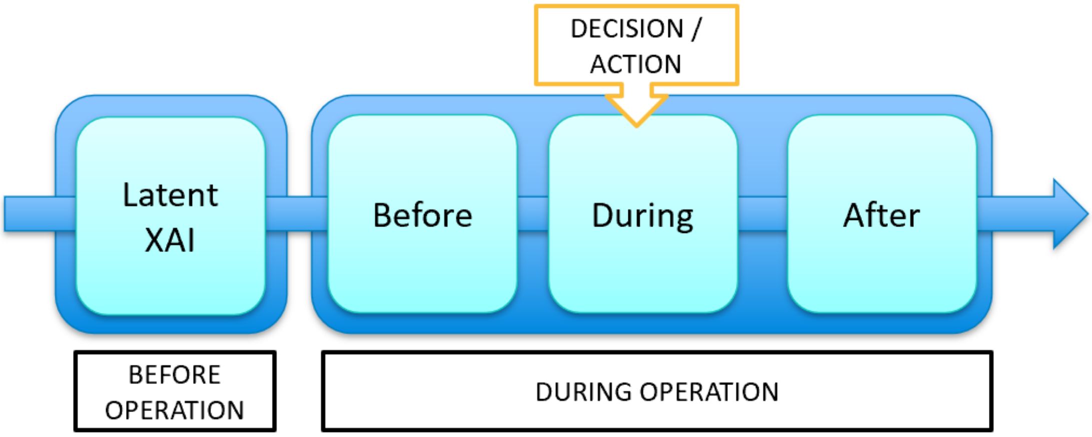


Figure 21 — Timeliness of the explainability

Reliability of the information

Objective EXP-12: For each output relevant to the task(s), the applicant should ensure the validity of the specified explanation, based on actual measurements (e.g. monitoring) or on a quantification of the level of uncertainty.

Objective EXP-13: The AI-based system should be able to deliver an indication of the degree of reliability of its output as part of the explanation based on actual measurements (e.g. monitoring) or on a quantification of the level of uncertainty.

Anticipated MOC EXP-13: Assuming that the decisions, actions, or diagnoses provided by an AI-based system may not always be fully reliable, the AI-based system should compute a degree of reliability of its outputs. Such an indication should be part of the elements provided within the explanations when needed.

4.1.4.2. Objectives related to the ODD and performance monitoring in operations

As mentioned in Section C.3, learning assurance guarantees are given in the frame of the defined ODD and at a given level of performance. One important objective is therefore to monitor whether or not the operational conditions remain within acceptable boundaries and the performance is aligned with the expected level.

The feedback of this monitoring is another contributor to the operational AI explainability guidelines.

The following objectives are anticipated:

Objective EXP-14: The AI-based system inputs should be monitored to be within the operational boundaries (both in terms of input parameter range and distribution) in which the AI/ML constituent performance is guaranteed, and deviations should be indicated to the relevant users and end users.

Objective EXP-15: The AI-based system outputs should be monitored to be within the specified operational performance boundaries, and deviations should be indicated to the relevant users and end users.

Objective EXP-16: The training and instructions available for the human end user should include procedures for handling possible outputs of the ODD and performance monitoring.

Objective EXP-17: Information concerning unsafe AI-based system operating conditions should be provided to the human end user to enable them to take appropriate corrective action in a timely manner.

4.2. Human-AI teaming

Initially, AI-based systems were developed to improve team performance but, with the technological advances in this area, AI-based systems will soon become teammates.

While moving from human-human teams to human-AI-based system teams (HAT), complexity is arising at a level where we are not able to fully comprehend as of today. The concept of HAT encompasses in this paper the notion of cooperation and collaboration.

Cooperation is a process in which the AI-based system works to help the end user accomplish their own objective and goal. The AI-based system will work according to a predefined task allocation pattern with informative feedback on the decision and/or action implementation. Cooperation does not imply a shared vision between the end user and the AI-based system. Communication is not a paramount capability for cooperation.

Collaboration is a process in which the human and the AI-based system work together and jointly to achieve a common goal (or work individually on a defined goal) and solve a problem through co-constructive approach. Collaboration implies the capability to share situational awareness and to readjust strategies and task allocation in real time. Communication is paramount to share valuable information needed to achieve the goal, to share ideas and expectations.

The following design considerations are anticipated, focusing on the different capabilities that the AI-based system should have to perform an efficient collaboration:

— Sharing of elements of situational awareness;

— Identification of abnormal situation and performance of diagnostics;

— Evaluation of the relevance of the solution proposed by the end user;

— Negotiation/argumentation;

— Adaptiveness.

Objective HF-01: The applicant should design the AI-based system with the ability to build its own individual situational awareness.

Objective HF-02: The applicant should design the AI-based system with the ability to allow the end user to ask questions and to answer questions from the end user, in order to reinforce the end-user individual situational awareness.

Objective HF-03: The applicant should design the AI-based system with the ability to modify its individual situational awareness on end-user request.

Objective HF-04: If a decision is taken by the AI-based system, the applicant should design the AI-based system with the ability to request from the end user a cross-check validation.

Corollary objective HF-04: The applicant should design the AI-based system with the ability to cross-check and validate a decision made by the end user automatically or on request.

Objective HF-05: For complex situations under normal operations, the applicant should design the AI-based system with the ability to identify a suboptimal strategy and propose through argumentation an optimised solution.

Corollary objective HF-05: The applicant should design the AI-based system with the ability to accept a proposal rejection, upon request by the end user.

Objective HF-06: For complex situations under abnormal operations, the applicant should design the AI-based system with the ability to identify the problem, share the diagnosis including the root cause, the resolution strategy and the anticipated operational consequences.

Corollary objective HF-06: The applicant should design the AI-based system with the ability to consider the arguments shared by the end user.

Anticipated MOC for HF-01 to HF-06: [to be further developed]

The AI-based system should have the capability to gather information from its environment, process it, and take decisions or actions based on that information without human intervention.

The AI-based system in order to build this SA will need to have the ability to sense, comprehend the surroundings, navigate through complex environments and adapt to changing conditions.

The level of initiative will differ depending on the task and functions allocated to the AI-based system:

— The AI-based system could present the relevant information and request from the end user to take the decision.

— The AI-based system could propose decisions and leave the choice to the end user.

— The AI-based system could pre-decide and request validation by the end user.

Objective HF-07: The applicant should design the AI-based system with the ability to detect poor decision-making by the end user in a time-critical situation.

Objective HF-08: The applicant should design the AI-based system with the ability to take the appropriate action outside of a collaboration scheme, in case of detection of poor decision-making by the end user in a time-critical situation.

Objective HF-09: The applicant should design the AI-based system with the ability to negotiate, argue, and support its positions.

Anticipated MOC HF-09: Several dimensions for collaboration have been identified:

— Human-AI work together on an agreement to achieve goals and solve a problem.

- Negotiation: There is a need to design an AI-based system that can exchange in case of inconsistent goals to propose alternative solutions.

— Human-AI work individually on a defined goal and, when ready, share their respective solution for agreement.

- Argumentation: In order for the AI to be able to negotiate, it should be designed to support its positions: what can be argued and how to argue.

Objective HF-10: The applicant should design the AI-based system with the ability to accept the modification of task allocation / task adjustments (instantaneous/short-term).

Anticipated MOC HF-10

As an example, role/task allocation between a pilot and co-pilot is defined by operation and airlines policy through CRM under pilot flying / pilot monitoring roles. With the introduction of collaborative capabilities, the definition of the roles and tasks will be performed at development level. Development of requirements on task sharing, adaptability and need for collaboration are foreseen.

The following should be considered:

— Context-sensitive task allocation: ML models should be tailored to answer the needs imposed or directed by the situation (time-critical / diversion selection).

— Customisation: capability given to the operators to tailor the AI-based system to answer operational needs.

— Live task adjustments/distribution: capability of the AI-based system to adjust the task allocation in real time to answer operational needs. The end user and the AI-based system need to stay in the loop of decisions to be able to react to any adjustment in real time. Both parties will need to have a mutual recognition and knowledge about the level of SA of each other. These adjustments could be anticipated at different levels:

  - Macro adjustment: e.g. The pilot could tell the AI-based system to take control of the communication task for the rest of the flight.

  - Micro adjustment: e.g. The pilot could request the AI-based system to perform a check to lower its workload as he or she is busy performing the radio communication.

4.3. Modality of interaction and style of interface

The introduction of AI-based systems is changing the paradigm of end user/machine interactions. The rise of AI is leading to a new mode of interaction through voice, gesture, or other natural interactions by bringing emerging technologies to a level that allows the machine to better communicate with the human and vice versa. The upcoming future flexible platforms (flight decks, controller working positions, etc.) open the way to less restrictive means of communication.

The following objectives focus on the emergence of languages, including natural spoken language and procedural spoken language, where voice is used as a new interface for communication. The exploration of other methods has broadened the field to the use of gesture recognition where movements and gestures are also used as a language to exchange.

4.3.1.1. Design criteria for communication to address spoken natural language

'Human-like' natural language could be defined as the result of a voice and speech recognition system, allowing the machine to understand human language and use the same language processes as the ones used for human-human conversation.

Spoken natural language conversation can provide smooth communication with the AI-based system and contribute to build trust, if the AI-based system outcome is relevant, efficient, non-ambiguous and timely. In addition, the end user will not have to learn a specific syntax to interact properly with the system. Spoken natural language is bringing flexibility in the interaction allowing clarifications on request.

On the other hand, spoken natural language dialogue implies a bidirectional conversation. It increases the chance of misleading comprehension or misinterpretation between the two parties and can lead to a reduction in efficiency. In particular, it can create errors that could lead to operational consequences. In addition, misinterpretation can increase the workload and create frustration, especially if no other interface is available to share the expected information.

Objective HF-11: The applicant should design the AI-based system with the ability to understand through the end-user responses or his or her action that there was a misinterpretation from the end user.

Objective HF-12: The applicant should design the AI-based system with the ability to notify the end user that he or she misunderstood the information provided through spoken natural language.

Objective HF-13: In case of misinterpretation, the applicant should design the AI-based system with the ability to resolve the misunderstanding through repetition, modification of the modality of interaction and/or with the provision of additional rationale on the initial information.

Note: In case of degradation of the interaction performance linked with the use of spoken natural language, the end user may have to use other modalities (see Section C.4.3.1.4).

Objective HF-14: The applicant should design the AI-based system with the ability to make room for dialogue turns by other participants, keeping silent when needed and not hindering the user (while the user is engaged in a complex activity), sticking to information that answers a given question by other participants, etc.)

Anticipated MOC HF-14

The natural language capabilities to be addressed include:

— Conversation
— Questions/answers
— Argumentation/negotiation
— Follow-up questions
— Acknowledgements

Objective HF-15: If spoken natural language is used, the applicant should design the AI-based system with the ability to provide information regarding the associated AI-based system capabilities and limitations.

Anticipated MOC HF-15

The end user might tend to have an erroneous expectation regarding the capabilities of the AI-based system due to the nature of the 'human-like' interaction: erroneous optimistic confidence and premature enthusiasm can be observed.

4.3.1.2. Design criteria for communication to address spoken procedural language (SPL)

Moving away from natural language to a procedural language requires a significant restriction on the lexicon available to the AI and end user. This style of language limits the use of vocabulary and imposes a strict syntax on communication. Examples include the issuing of instructions and requests between ground and air in radio telephone (RT) communication. Implementing a spoken procedural interface provides the end user with a constant and homogeneous outcome.

Using spoken 'procedure or programming style' language presents the message sender and receiver with a fixed syntax by which they communicate. This fixed syntax format is similar to that which currently exists in flight deck through the crew resource management (CRM) methods and on the ground through team resource management (TRM). The use of fixed syntax language provides a structure to a communication so that it is clear:

— which parameter are being discussed;

— the value to be associated with the parameter;

— a qualifier, if required, for the value;

— a clear path for acknowledgment of the reception of the communication.

SPL provides the opportunity to reduce error in communications as they are less subject to interpretation and the expectation of the fixed grammar ensures that potential errors can be more easily identified.

The fixed syntax associated with procedural language does however lack the flexibility of natural language and may affect the understanding of communication that is based on context. In addition, a fixed syntax prevents smooth and natural conversation between the AI-based system and the end user. While procedural languages are associated with reduced errors, they can be also associated with increased cognitive costs due to the necessity of remembering the way to interact as well as the syntax and totality of commands and qualifiers available. The end user will therefore be required to continuously access to knowledge and memory.

Objective HF-16: The applicant should design the syntax of the spoken procedural language so that it can be learned easily by the end user.

Objective HF-17: The applicant should design the AI-based system with the ability to transition from verbal natural language to verbal procedural language depending on its own perception of the performance of the dialogue, the context of the situation and the characteristics of the task.

4.3.1.3. Design criteria for gesture non-verbal language

Gesture language is considered in this paper as a non-verbal, unidirectional communication tool where the end user would have their body movements tracked and processed through dedicated technology. Gesture language can also be combined with spoken languages as a resource to reinforce the efficiency of the bidirectional communication.

Objective HF-18: The applicant should design the gesture language syntax so that it is intuitively associated with the command that it is supposed to trigger.

Objective HF-19: The applicant should design the AI-based system with the ability to filter the intentional gesture used for language from non-intentional gesture, such as spontaneous gestures that are made to complement verbal spoken language.

Objective HF-20: If gesture non-verbal language is used, the applicant should design the AI-based system with the ability to recognise the end-user intention.

Objective HF-21: If non-verbal language is used, the applicant should design the AI-based system with the ability to acknowledge the end-user intention with appropriate feedback

4.3.1.4. Design criteria for management of multi-modal interaction

A combination of several interaction modalities such as voice (speech recognition), visual (e.g. keyboard, mouse, display) and gesture can be foreseen.

As an example, a combination could be performed by the AI-based system to increase:

— usability;

— the understanding by confirmation;

— accessibility by providing the end user with back up/additional interface to compensate for senses (sight, hearing, touch, vision) availabilities;

— efficiency by performing two distinct actions through two different means.

Objective HF-22: If spoken natural language is used, the applicant should design the AI-based system so that it can be deactivated for the benefit of other modalities in case of degradation of the interaction performance.

Objective HF-23: The applicant should design the AI-based system with the ability to combine or adapt the interaction modalities depending on the characteristics of the task, the operational event, and/or the operational environment.

Objective HF-24: The applicant should design the AI-based system with the ability to automatically adapt the modality of interactions to the end-user states, the situation, the context and/or the perceived end user's preferences.

Anticipated MOC HF-22, HF-23 and HF-24

Adaptive interaction modality is the AI capacity to adapt the modality of interaction to external/internal factors with the objective of optimising the HAII. The following attributes have been identified: context-sensitive criteria (e.g. if pilot is speaking with ATC, the AI will communicate using other interfaces than natural language), task-sensitive criteria, pilot-state-sensitive criteria.

As an example, during ATC-pilot communication, the AI-based system should avoid using natural language to interact with the pilot; it should instead display the information collected.

By inference, the following design possibilities have been identified:

— The modality of interaction can be predefined by the applicant in adaptation with the characteristics of the task or the flight event.

— The modality of interaction should adapt to pilot's state:

  - from permanent state (crew personal setting);
  - from instantaneous state (workload, stress, cognitive resources).

— The modality of interaction should automatically adapt to the situation, the task, and/or the context.

— The AI should propose modality of interaction according to the perception of pilot's preferences and expectations. As an example, during ATC-pilot communication, AI system should avoid using natural language to interact with the pilot; it should instead display the information collected.

4.4. Error management

4.4.1. Contribution of AI-based systems to a new typology of human errors

4.4.1.1. Design-related human errors

CS 25.1302 states that 'to the extent practicable, installed equipment must enable the flight crew to manage errors resulting from the kinds of flight crew interactions with the equipment that can be reasonably expected in service, assuming the flight crew is acting in good faith. This sub-paragraph (d) does not apply to skill-related errors associated with manual control of the aeroplane.' The requirement stipulates that equipment shall be designed to be tolerant to human error.

The emergence of AI-based systems is likely to introduce new types of errors. One may expect that these new types of errors will result from the end user or from the HAT. The errors resulting directly from the AI-based system should also be considered.

Objective HF-25: The applicant should design the AI-based system to minimise the likelihood of design-related errors made by the end user.

Note: The minimisation of the likelihood of errors made by the AI-based system is addressed through the AI assurance Section C.3.

Objective HF-26: The applicant should design the AI-based system to minimise the likelihood of design-related errors made by the human-AI teaming.

4.4.1.2. Operation-related errors

In the aviation environment the use of two people in the flight deck, shared speech frequencies, controllers working as pairs, and the use of 'sign off' in maintenance activities are all examples of means to minimise the likelihood of operation- and organisation-related errors.

According to the SKYbrary website, 'Crew Resource Management (CRM) is the effective use of all available resources' (equipment, procedures and people) 'for flight crew personnel to assure a safe and efficient operation, reducing error, avoiding stress and increasing efficiency. (…) CRM encompasses a wide range of knowledge, skills and attitudes including communications, situational awareness, problem solving, decision making, and teamwork.' (SKYbrary)

By analogy, there is a need to define the notion of human-AI resource management (HAIRM), considering that the introduction of AI is likely to bring some specific problematics, in particular, regarding the communications, situational awareness, problem-solving, decision-making and teamwork.

Objective HF-27: The applicant should design the AI-based system to minimise the likelihood of HAIRM-related errors.

Anticipated MOC HF-27

The errors resulting from the HAIRM can take the form of:

— failure to respect the predefined task allocation pattern;

— Teamwork breakdown (team is not working properly, e.g. due to a communication issue, lack of trust, poor definition of roles and tasks);

— incomplete or incorrect cross-checking process;

— HAI communication issues (for the errors resulting from the communication established between the AI-based system and the end user, refer to the **Objectives HF-11 to HF-13**);

— mismatch in Human/AI-based system respective situational awareness;

— strategic or authority issues related to decision-making

4.4.2. How AI-based systems will affect the methods of errors management

Considering that errors will occur despite the implementation of the **Objectives HF-25 to HF-28**, the introduction of AI will provide new opportunities and ways to manage errors.

Objective HF-28: The applicant should design the AI-based system to be tolerant to end-user errors.

Anticipated MOC HF-28
An AI-based system being tolerant to human errors means that the system is robust enough and will continue performing its intended function despite human errors. On top of that, it should be able to detect and potentially correct these errors so that it can continue to operate as intended.

Objective HF-29: The applicant should design the AI-based system so that in case the end user makes an error while interacting with the AI-based system, the opportunities exist to detect the error.

Objective HF-30: The applicant should design the AI-based system so that once an error is detected, the AI-based system should provide efficient means to inform the end user and correct the error.

4.5. Failure management

		Failure management under end-user responsibility Cooperation		**Failure management under HAIRM** Collaboration		**Failure management under AI-based system responsibility**	
Task/ function	Sub task/function	End user	AI-based system	End user	AI-based system	End user	AI-based system
Attention getting	Emission	Yes	No	Yes: In both directions		No	No
	Acknowledgement	Yes	Yes under conditions	Yes	Yes	No	No
Diagnostic	Situation analysis	Yes	No	Yes	Yes	No	yes
	Identification of the problematics	Yes	No	Yes The end user proposes and the Ais confirms	Yes AI proposes and the end user confirms	No	yes
	Potential solutions proposal	Yes	No	Yes The end user proposes and AI confirms	Yes AI proposes and the end user confirms	No	yes
	Action planning	Yes	No	Yes The end user proposes and AI confirms	Yes AI proposes and the end user confirms	No	yes
Action	Act on the systems	Yes	No except when requested by the end user	Yes: The end user will need a means to act on the system	Yes: The AI-based system will interact without any means on the system	No except if required by operation	Yes

	Require a pre-confirmation	Yes	No	Yes	Yes	No	No
	Perform the action	Yes	No	Yes	Yes	No	Yes
	Monitoring	No	Yes	Yes	Yes	Yes under condition s	Yes, own monitoring
	Verification	No	Yes	Yes	Yes	Yes under condition	Yes, own verification
Post failure managem ent	Need for presentation of status	Yes	Yes	Yes	Yes	Yes	No
	Limitations management	Yes	Monitor and recall	Yes	Yes	No	Yes
	Deferred actions	Yes	No	Yes	Yes	No	Yes

Table 5 — Overview of failure management guidance

4.6. Additional topics under development

Some additional topics are currently under development with a focus on:

— workload management;

— customisation of human-AI interface.

5. AI safety risk mitigation

5.1. AI safety risk mitigation concept

AI SRM is based on the anticipation that the 'AI black box' may not always be opened to a sufficient extent. Indeed, for some applications, it could be unpractical to fully cover all the objectives defined in the explainability and learning assurance building blocks of this guideline. This partial coverage of some objectives could result in a residual risk that may be accommodated by implementing some mitigations called hereafter SRM. The intent of such mitigations is to minimise as far as practicable the probability of the AI/ML constituent producing unintended or unexplainable outputs.

Furthermore, it is also recognised that the use of AI in the aviation domain is quite novel and until field service experience is gained, appropriate safety precautions should be implemented to reduce the risk to occupants, third parties and critical infrastructure.

This could be achieved by several means, among others:

— real-time monitoring of the output of the AI/ML constituent and passivation of the AI-based system with recovery through a traditional backup system (e.g. safety net);

— in a wider horizon, by considering the notion of 'licensing' for an AI-based agent, as anticipated in (Javier Nuñez et al., 2019) and developed further in (ECATA Group, 2019).

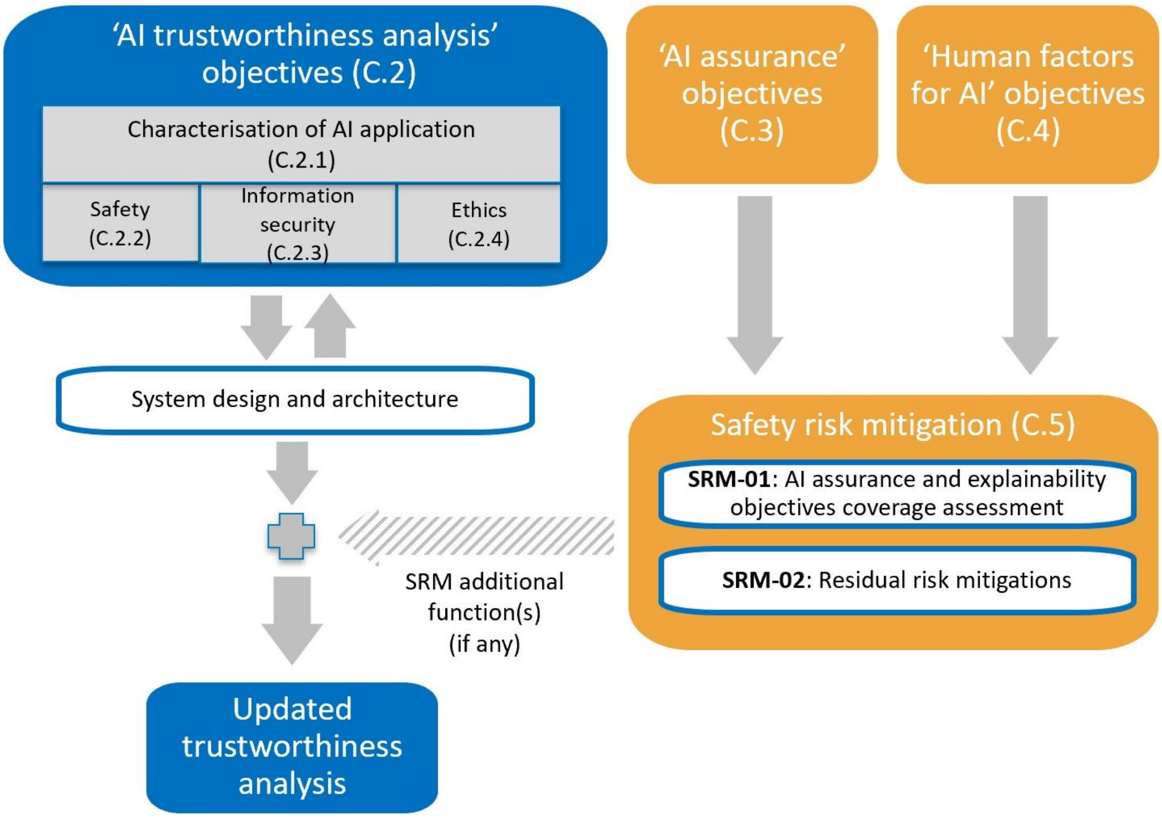


Figure 22 — SRM block interfaces with other building blocks

Note that SRM is solely meant to address a partial coverage of the applicable explainability and learning assurance objectives. SRM is not aimed at compensating partial coverage of objectives belonging to the trustworthiness analysis building blocks (e.g. safety assessment, information security, ethics-based objectives).

5.2. AI SRM top-level objectives

Objective SRM-01: Once activities associated with all other building blocks are defined, the applicant should determine whether the coverage of the objectives associated with the explainability and learning assurance building blocks is sufficient or whether an additional dedicated layer of protection, called hereafter safety risk mitigation (SRM), would be necessary to mitigate the residual risks to an acceptable level.

Anticipated MOC SRM-01: In establishing whether AI SRM is necessary and to which extent, the following considerations should be accounted for:

— coverage of the explainability building block;

— coverage of the learning assurance building block;

— relevant in-service experience, if any;

— AI-level: the higher the level, the more likely it is that SRM will be needed;

— criticality of the AI/ML constituent: the more the ML/AI constituent is involved in critical functions, the more likely it is that SRM will be needed.

In particular, the qualitative nature of some building block mitigations/analysis should be reviewed to establish the need for an SRM.

The SRM strategy should be commensurate with the residual risk/unknown.

Objective SRM-02: The applicant should establish SRM means as identified in Objective SRM-01.

Anticipated MOC SRM-02-1: The following means may be used to gain confidence that the residual risk is properly mitigated:

— monitoring of the output of the AI/ML constituent and passivation of the AI-based system with recovery through a traditional backup system (e.g. safety net);

— when relevant, the possibility may be given to the end user to switch off the AI/ML-based function to avoid being distracted by erroneous outputs.

The SRM functions should be evaluated as part of the safety assessment[19], and, if necessary, appropriate safety requirements should be defined and verified. This may include independence requirements to guarantee an appropriate level of independence of the SRM architectural mitigations from the AI/ML constituent

[19] In the ATM/ANS domain, for non-ATS providers, the safety assessment is replaced by a safety support assessment.

6. Organisations

Prior to obtaining approval of AI applications in the field of civil aviation, organisations that are required to be approved as per the Basic Regulation (Regulation (EU) 2018/1139) might need to introduce adaptations in order to ensure the adequate capability to meet the objectives defined within the AI trustworthiness building blocks (see Figure 2), and to maintain the compliance of the organisation with the corresponding implementing rules.

The introduction of the necessary changes to the organisation would need to follow the process established by the applicable regulations. For example, in the domain of initial airworthiness, the holder of a DOA would need to apply to EASA for a significant change to its design assurance system prior to the application for the certification project.

At this stage, it is worth mentioning that Commission Delegated Regulation (EU) 2022/1645 and Commission Implementing Regulation (EU) 2023/203 (being respectively applicable from 2025 and 2026), on the management of information security risks with a potential impact on aviation safety, require organisations adapt their processes to comply with their requirements. In the context of AI/ML applications, compliance with these Regulations will require that information security aspects during the design, production, and operation phases will be adequately managed and mitigated (e.g. data poisoning in development).

This section introduces some high-level provisions and anticipated AMC with the aim of providing guidance to organisations on the expected adaptations. It provides as well, as an example case, more detailed guidance on the affected processes for holders of a DOA.

6.1. High-level provisions and anticipated AMC

Provision ORG-01: The organisation should review its processes and adapt them to the introduction of AI technology.

Provision ORG-02: In preparation of the Commission Delegated Regulation (EU) 2022/1645 and Commission Implementing Regulation (EU) 2023/203 applicability, the organisation should assess the information security risks related to the design, production and operation phases of an AI/ML application.

Anticipated AMC ORG-02:

Taking advantage of the ENISA report (ENISA, December 2021) on SECURING MACHINE LEARNING ALGORITHMS and possible threats identified in Table 3, the organisation could consider threat scenarios:

— related to unauthorised alterations of the training, validation, and test data sets commonly referred to as 'data set poisoning';

— like 'denial of service' due to inconsistent data or a sponge example, while learning algorithms usually consider input data in a defined format to make their predictions. A denial of service could be caused by input data whose format is inappropriate. It may also happen that a malicious user of the model constructs input data (a sponge example) specifically designed to increase the computation time of the model and thus potentially cause a denial of service.

<u>Provision ORG-03:</u> Implement a data-driven 'AI continuous safety assessment system' based on operational data and in-service events.

Anticipated AMC ORG-03:

The AI continuous safety assessment system should:

— ensure data gathering on safety-relevant areas for AI-based systems;

— perform analyses to support the identification of in-service risks, based on:
  - the organisation scope;
  - a set of safety-related metrics;
  - available relevant data.

The system should be able to refine the identification of risks based on the results of previous interactions with the AI-based systems and incorporating the human evaluation inputs.

When defining the metrics, the data set and gathering methodology should ensure:

— the acquisition of safety-relevant data related to accidents and incidents including near-miss events; and

— the monitoring of in-service data to detect potential issues or suboptimal performance trends that might contribute to safety margin erosion; and

— the definition of target values, thresholds and evaluation periods; and

— the possibility to analyse data to determine the possible root cause and trigger corrective actions.

The following implementing rule requirements, associated AMC and GM may be considered with appropriate adaptations:

For ATS providers:

— ATS.OR.200(2) and (3) Safety management system

— GM1 ATS.OR.200(3)(i) and GM1 ATS.OR.200(3)(iii)

— AMC1 ATS.OR.200(3)(iii)

<u>Provision ORG-04:</u> The organisation should ensure that the safety-related AI-based systems are auditable by internal and external parties, including the approving authorities.

<u>Provision ORG-05:</u> The organisation should adapt the continuous risk management process to accommodate the specificities of AI, including interaction with all relevant stakeholders.

Anticipated AMC ORG-05:

In particular, the applicant should put in place:

— a process to discuss and continuously monitor and assess the AI-based system's adherence to the ethics-based assessment guidance;

— a process for third parties (e.g. suppliers, end users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or bias in the AI-based system.

Provision ORG-06: The organisation should adapt the training processes to accommodate the specificities of AI, including interaction with all relevant stakeholders.

Anticipated AMC ORG-06:

In particular, the applicant should put in place:

— consider the competencies needed to deal with the AI-based systems;

— adaptations to training syllabus to take into account the specificities of AI.

Provision ORG-07: The organisations operating the AI-based systems should ensure that end users' licensing and certificates account for the specificities of AI, including interaction with all relevant stakeholders.

Provision ORG-08: The organisation should establish means (e.g. processes) to continuously assess ethics-based aspects for the trustworthiness of an AI-based system with the same scope as for objective ET-01. This includes the consideration of establishing an AI ethics review board.

6.2. Competence considerations

The inclusion of AI/ML technology in aviation will determine new challenges at all levels from designers to end users. Along with the advantages coming from the progress in technology, several areas of threats will become active.

This section will give consideration to training as a means of mitigation to the threats related to the lack of awareness on AI-based system features.

It is important that every actor in the chain of design, production and operation of aviation systems using AI-based technology receives appropriate information on topics such as:

— Basic concepts of AI;

— AI-based system capability and levels;

— Human factors, including HAII and explainability;

— AI-AI interface;

— Ethics-based assessment;

— Safety management;

— Information security management;

— any other relevant aspect of AI pertaining to the individual post.

At organisation level, each type of organisation should review the threats connected with the use of AI pertaining to the scope activity and develop initial and recurrent programmes aimed to build awareness of their personnel on such topics (refer to **Provision ORG-06**).

The awareness training shall be delivered to all levels of personnel, including top management, to ensure the correct approach to the introduction of AI-based technology in the organisation.

At the individual level, the elements above shall be addressed in the initial training of each domain-specific licence or certificate (refer to **Provision ORG-07**). Furthermore, device- or environment-specific elements shall be considered for the final use cases.

It is equally important that awareness training is addressed to instructors and examiners, as well as to the regulators and inspectors involved in the development or oversight of organisations and products.

6.3. Design organisation case

This section aims to provide an example, for the case of DOA holders by identifying those processes that might need to be assessed and adapted.

The following figure illustrates the potentially affected DOA processes and the key activities in relation to the implementation of AI/ML technologies:

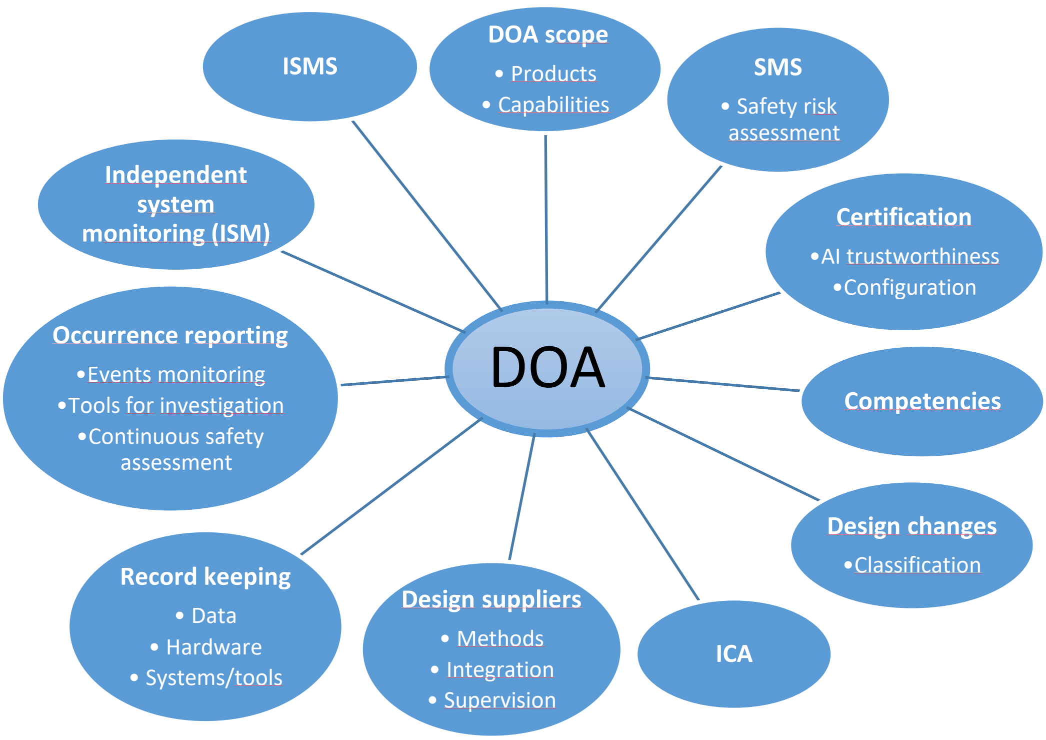


Figure 23 — DOA processes potentially affected by the introduction of AI/ML

Although almost all DOA processes are affected, the nature of the adaptation would be different depending on the interrelation of the process and the specificities of the AI technology.

The certification process would need to be deeply adapted to introduce new methodologies that will ensure compliance with the AI trustworthiness objectives as introduced in the previous sections of this guidance. Similarly, new methodologies might be required for the record-keeping of AI-related data, for the independent system monitoring (ISM) process with regard to both compliance with and adequacy of procedures, and for the continuous safety assessment of events when the root cause might be driven by the AI-based system.

With regard to design changes, new classification criteria may be required when an approved type design related to AI is intended to be changed.

Other processes such as competencies would need to be implemented considering the new AI technologies and the related certification process.

Finally, the DOA scope would need to reflect the capabilities of the organisation in relation to product certification and to privileges for the approval of related changes.

D. Proportionality of the guidance

1. Concept for modulation of objectives

Two main criteria can be used to anticipate proportionality in the objectives from the guidance that is proposed in Chapter C of this document: the AI Level (per **Objective CL-01**) and the criticality (assurance level per **Objective SA-01**) of the item containing the ML model.

A modulation of the objectives of this document, based on these two criteria, has been introduced in the next section.

Notes:

— With the current state of knowledge of AI and ML technology, EASA anticipates a limitation on the validity of applications when AI/ML constituents include IDAL A or B / SWAL 1 or 2 / AL 1, 2 or 3 items. Moreover, no assurance level reduction should be performed for items within AI/ML constituents. This limitation will be revisited when experience with AI/ML techniques has been gained.

— Future work on Level 3 is likely to increase the number of objectives.

2. Risk-based levelling of objectives

Applicability by Assurance Level	
●	The objective should be satisfied with independence.
○	The objective should be satisfied.
	The satisfaction of the objective is at the applicant's discretion.

Applicability by AI Level	
	The objective should be satisfied for AI level 1A, 1B, 2A and 2B.
(hatched)	*The objective should be satisfied for AI level 1B, 2A and 2B.*
(orange)	The objective should be satisfied for AI level 2A and 2B.
(blue)	The objective should be satisfied for AI level 2B.

Building block	Objectives	Assurance Level				
		AL 1 DAL A SWAL1	AL 2 DAL B -	AL 3 DAL C SWAL2	AL 4 - SWAL3	AL 5 DAL D SWAL4
Trustworthiness analysis	**CO-01:** The applicant should identify the list of end users that are intended to interact with the AI-based system, together with their roles, their responsibilities and their expected expertise (including assumptions made on the level of training, qualification and skills).	○	○	○	○	○
	CO-02: For each end user, the applicant should identify which high-level task(s) are intended to be performed in interaction with the AI-based system.	○	○	○	○	○
	CO-03: The applicant should determine the AI-based system taking into account domain-specific definitions of 'system'.	○	○	○	○	○
	CO-04: The applicant should define and document the ConOps for the AI-based system, including the task allocation pattern between the end user(s) and the AI-based system. A focus should be put on the definition of the OD and on the capture of specific operational limitations and assumptions.	○	○	○	○	○
	CO-05: The applicant should document how end users' inputs are collected and accounted for in the development of the AI-based system.	○	○	○	○	○
	CO-06: The applicant should perform a functional analysis of the system.	○	○	○	○	○
	CL-01: The applicant should classify the AI-based system, based on the levels presented in Table 2, with adequate justifications.	○	○	○	○	○
	SA-01: The applicant should perform a safety (support) assessment for all AI-based (sub)systems, identifying and addressing specificities introduced by AI/ML usage.	●	●	○	○	○
	ICSA-01: The applicant should identify which data needs to be recorded for the purpose of supporting the continuous safety assessment .	●	●	○	○	○
	ICSA-02: The applicant should use the collected data to perform a continuous safety assessment. This includes: — the definition of target values, thresholds and evaluation periods to guarantee that design assumptions hold; — the monitoring of in-service events to detect potential issues or suboptimal performance trends that might contribute to safety margin erosion, or, for non-ATS providers, to service performance degradations; and — the resolution of identified shortcomings or issues.	●	●	○	○	○
	IS-01: For each AI-based system and its data sets, the applicant should identify those information security risks with an impact on safety, identifying and addressing specific threats introduced by AI/ML usage.	○	○	○	○	○
	IS-02: The applicant should document a mitigation approach to address the identified AI/ML-specific security risk. Note: Beyond the applicability defined here for any domain, further levelling may be introduced in domains defining specific security assurance levels (SALs). See Section D.3.	○	○	○	○	○

Building block	Objectives	Assurance Level				
		AL 1 DAL A SWAL1	AL 2 DAL B -	AL 3 DAL C SWAL2	AL 4 - SWAL3	AL 5 DAL D SWAL4
	IS-03: The applicant should validate and verify the effectiveness of the security controls introduced to mitigate the identified AI/ML-specific security risks to an acceptable level. Note: Beyond the applicability defined here for any domain, further levelling may be introduced in domains defining specific security assurance levels (SALs). See Section D.3.	○	○	○	○	○
	ET-01: The applicant should perform an ethics-based trustworthiness assessment for any AI-based system developed using ML techniques or incorporating ML models.	○	○	○	○	○
	ET-02: The applicant should ensure that the AI-based system bears no risk of creating human attachment, stimulating addictive behaviour, or manipulating the end user's behaviour.	○	○	○	○	○
	ET-03: The applicant should ensure that the AI-based system presents no capability of adaptive learning.	○	○	○	○	○
	ET-04: The applicant should comply with national and EU data protection regulations (e.g. GDPR), i.e. involve their Data Protection Officer (DPO), consult with their National Data Protection Authority, etc.	○	○	○	○	○
	ET-05: The applicant should ensure that procedures are in place to avoid creating or reinforcing unfair bias in the AI-based system, regarding both the data sets and the trained models.	○	○	○	○	○
	ET-06: The applicant should perform an environmental impact analysis, identifying and assessing potential negative impacts of the AI-based system on the environment and human health throughout its life cycle (development, deployment, use, end of life).	○	○	○	○	○
	ET-07: The applicant should define measures to reduce or mitigate the impacts identified under Objective ET-06.	○	○	○	○	○
	ET-08: The applicant should identify the need for new competencies for users and end users to interact with and operate the AI-based system, and mitigate possible training gaps (link to **Provision ORG-06, Provision ORG-07**).	○	○	○	○	○
	ET-09: The applicant should perform an assessment of the risk of de-skilling of the users and end users and mitigate the identified risk through a training needs analysis and a consequent training activity (link to **Provision ORG-06, Provision ORG-07**).	○	○	○	○	○

Building block	Objectives	Assurance Level				
		AL 1 DAL A SWAL1	AL 2 DAL B -	AL 3 DAL C SWAL2	AL 4 - SWAL3	AL 5 DAL D SWAL4
AI assurance	**DA-01:** The applicant should describe the proposed learning assurance process, taking into account each of the steps described in Sections C.3.1.2 to C.3.1.12, as well as the interface and compatibility with development assurance processes.	○	○	○	○	○
	DA-02: Documents should be prepared to encompass the capture of the following minimum requirements: — safety requirements allocated to the AI/ML constituent; — information security requirements allocated to the AI/ML constituent; — functional requirements allocated to the AI/ML constituent; — operational requirements allocated to the AI/ML constituent, including ODD and AI/ML constituent performance monitoring, detection of OoD input data and data-recording requirements; — non-functional requirements allocated to the AI/ML constituent (e.g. performance, scalability, reliability, resilience, etc.); and — interface requirements.	○	○	○	○	○
	DA-03: The applicant should describe the system and subsystem architecture, to serve as reference for related safety (support) assessment and learning assurance objectives.	○	○	○	○	
	DA-04: Each of the captured requirements should be validated.	●	●	○	○	○
	DA-05: The applicant should document evidence that all derived requirements have been provided to the (sub)system processes, including the safety (support) assessment.	○	○	○	○	○
	DA-06: The applicant should document evidence of the validation of the derived requirements, and of the determination of any impact on the safety (support) assessment and (sub)system requirements.	○	○	○	○	○
	DA-07: Each of the captured (sub)system requirements allocated to the AI/ML constituent should be verified.	●	●	○	○	○
	DM-01: The applicant should define the set of parameters pertaining to the AI/ML constituent ODD.	○	○	○	○	○
	DM-02: The applicant should capture the DQRs for all data pertaining to the data management process, including but not limited to: — the data needed to support the intended use; — the ability to determine the origin of the data; — the requirements related to the annotation process; — the format, accuracy and resolution of the data; — the traceability of the data from their origin to their final operation through the whole pipeline of operations; — the mechanisms ensuring that the data will not be corrupted while stored or processed, — the completeness and representativeness of the data sets; and — the level of independence between the training, validation and test data sets.	○	○	○	○	○
	DM-03: The applicant should capture the requirements on data to be pre-processed and engineered for the inference model in development and for the operations.	○	○	○	○	○

Building block	Objectives	Assurance Level				
		AL 1 DAL A SWAL1	AL 2 DAL B -	AL 3 DAL C SWAL2	AL 4 - SWAL3	AL 5 DAL D SWAL4
AI assurance	**DM-04:** The applicant should ensure the validation to an adequate level of the correctness and completeness of the ML constituent ODD.	●	●	○	○	○
	DM-05: The applicant should ensure the validation of the correctness and completeness of requirements on data to be pre-processed and engineered for the trained and inference model, as well as of the DQRs on data.	●	●	○	○	○
	DM-06: The applicant should identify data sources and collect data in accordance with the defined ODD, while ensuring satisfaction of the defined DQRs, in order to drive the selection of the training, validation and test data sets.	○	○	○	○	○
	DM-07: Once data sources are collected, the applicant should ensure the high quality of the annotated or labelled data in the data set.	●	●	○	○	○
	DM-08: The applicant should define the data preparation operations to properly address the captured requirements (including DQRs).	○	○	○	○	○
	DM-09: The applicant should define and document pre-processing operations on the collected data in preparation of the training.	○	○	○		
	DM-10: When applicable, the applicant should define and document the transformations to the pre-processed data from the specified input space into features which are effective for the performance of the selected learning algorithm.	○	○	○		
	DM-11: If the learning algorithm is sensitive to the scale of the input data, the applicant should ensure that the data is effective for the stability of the learning process.	○	○	○		
	DM-12: The applicant should distribute the data into three separate and independent data sets which will meet the specified DQRs: — the training data set and validation data set, used during the model training; — the test data set used during the learning process verification, and the inference model verification.	○	○	○	○	○
	DM-13: The applicant should ensure validation and verification of the data, as appropriate, all along the data management process so that the data management requirements (including the DQRs) are addressed.	●	●	○	○	○
	DM-14: The applicant should perform a data and learning verification step to confirm that the appropriate data sets have been used for the training, validation and verification of the model and that the expected guarantees (generalisation, robustness) on the model have been reached.	●	●			
	LM-01: The applicant should describe the AI/ML constituents and the model architecture.	○	○	○	○	○

Building block	Objectives	Assurance Level				
		AL 1 DAL A SWAL1	AL 2 DAL B -	AL 3 DAL C SWAL2	AL 4 - SWAL3	AL 5 DAL D SWAL4
AI assurance	**LM-02:** The applicant should capture the requirements pertaining to the learning management and training processes, including but not limited to: — model family and model selection; — learning algorithm(s) selection; — cost/loss function selection describing the link to the performance metrics; — model bias and variance metrics and acceptable levels; — model robustness and stability metrics and acceptable levels; — training environment (hardware and software) identification; — model parameters initialisation strategy; — hyper-parameters identification and setting; — expected performance with training, validation and test sets.	○	○	○	○	○
	LM-03: The applicant should document the credit sought from the training environment and qualify the environment accordingly.	○	○	○		
	LM-04: The applicant should provide quantifiable generalisation guarantees.	○	○	○		
	LM-05: The applicant should document the result of the model training.	○	○	○	○	○
	LM-06: The applicant should document any model optimisation that may affect the model behaviour (e.g. pruning, quantisation) and assess their impact on the model behaviour or performance.	○	○	○		
	LM-07: The applicant should account for the bias-variance trade-off in the model family selection and should provide evidence of the reproducibility of the training process.	●	●	○		
	LM-08: The applicant should ensure that the estimated bias and variance of the selected model meet the associated learning process management requirements.	●	●	○	○	○
	LM-09: The applicant should perform an evaluation of the performance of the trained model based on the test data set and document the result of the model verification.	●	●	○	○	○
	LM-10: The applicant should perform a requirements-based verification of the trained model behaviour and document the coverage of the AI/ML constituent requirements by verification methods.	●	●	○	○	○
	LM-11: The applicant should provide an analysis on the stability of the learning algorithms.	●	●	○		
	LM-12: The applicant should perform and document the verification of the stability of the trained model.	●	●	○	○	○
	LM-13: The applicant should perform and document the verification of the robustness of the trained model in adverse conditions.	●	●	○	○	○
	LM-14: The applicant should verify the anticipated generalisation bounds using the test data set.	●	●	○		
	IMP-01: The applicant should capture the requirements pertaining to the implementation process.	○	○	○	○	○
	IMP-02: Any post-training model transformation (conversion, optimisation) should be identified and validated for its impact on the model behaviour and performance, and the environment (i.e. software tools and hardware) necessary to perform model transformation should be identified.	○	○	○		

Building block	Objectives	Assurance Level				
		AL 1 DAL A SWAL1	AL 2 DAL B -	AL 3 DAL C SWAL2	AL 4 - SWAL3	AL 5 DAL D SWAL4
AI assurance	**IMP-03:** The applicant should plan and execute appropriate development assurance processes to develop the inference model into software and/or hardware items.	○	○	○		
	IMP-04: The applicant should verify that any transformation (conversion, optimisation, inference model development) performed during the trained model implementation step has not adversely altered the defined model properties.	○	○	○		
	IMP-05: The differences between the software and hardware of the platform used for training and the one used for verification should be identified and assessed for their possible impact on the inference model behaviour and performance.	○	○	○		
	IMP-06: The applicant should perform an evaluation of the performance of the inference model based on the test data set and document the result of the model verification.	○	○	○	○	○
	IMP-07: The applicant should perform and document the verification of the stability of the inference model.	●	●	○	○	○
	IMP-08: The applicant should perform and document the verification of the robustness of the inference model in adverse conditions.	●	●	○	○	○
	IMP-09: The applicant should perform a requirements-based verification of the inference model behaviour when integrated into the AI/ML constituent and document the coverage of the ML constituent requirements by verification methods.	●	●	○	○	○
	CM-01: The applicant should apply all configuration management principles to the AI/ML constituent life-cycle data, including but not limited to: — identification of configuration items; versioning; baselining; — change control; reproducibility; — problem reporting; — archiving and retrieval, and retention period.	○	○	○	○	○
	QA-01: The applicant should ensure that quality/process assurance principles are applied to the development of the AI-based system, with the required independence level.	●	●	●	●	●
	EXP-01: The applicant should identify the list of stakeholders, other than end users, that need explainability of the AI-based system at any stage of its life cycle, together with their roles, their responsibilities and their expected expertise (including assumptions made on the level of training, qualification and skills).	○	○	○		
	EXP-02: For each of these stakeholders (or groups of stakeholders), the applicant should characterise the need for explainability to be provided, which is necessary to support the development and learning assurance processes.	○	○	○		
	EXP-03: The applicant should identify and document the methods at AI/ML item and/or output level satisfying the specified AI explainability needs.	○	○	○		
	EXP-04: The applicant should provide the means to record operational data that is necessary to explain, post operations, the behaviour of the AI-based system and its interactions with the end user.	○	○	○	○	○

Building block	Objectives	Assurance Level				
		AL 1 DAL A SWAL1	AL 2 DAL B -	AL 3 DAL C SWAL2	AL 4 - SWAL3	AL 5 DAL D SWAL4
Human factors for AI	***EXP-05:*** *For each output of the AI-based system relevant to task(s) (per Objective CO-02), the applicant should characterise the need for explainability.*	○	○	○	○	○
	EXP-06: *The applicant should present explanations to the end user in a clear and unambiguous form.*	○	○	○	○	○
	EXP-07: *The applicant should define relevant explainability so that the receiver of the information can use the explanation to assess the appropriateness of the decision / action as expected.*	○	○	○	○	○
	EXP-08: *The applicant should define the level of abstraction of the explanations, taking into account the characteristics of the task, the situation, the level of expertise of the end user and the general trust given to the system.*	○	○	○	○	○
	EXP-09: *Where a customisation capability is available, the end user should be able to customise the level of details provided by the system as part of the explainability.*	○	○	○	○	○
	EXP-10: *The applicant should define the timing when the explainability will be available to the end user taking into account the time criticality of the situation, the needs of the end user, and the operational impact.*	○	○	○	○	○
	EXP-11: *The applicant should design the AI-based system so as to enable the end user to get upon request explanation or additional details on the explanation when needed.*	○	○	○	○	○
	EXP-12: *For each output relevant to the task(s), the applicant should ensure the validity of the specified explanation, based on actual measurements (e.g. monitoring) or on a quantification of the level of uncertainty.*	○	○	○	○	○
	EXP-13: *The AI-based system should be able to deliver an indication of the degree of reliability of its output as part of the explanation based on actual measurements (e.g. monitoring) or on a quantification of the level of uncertainty.*	○	○	○	○	○
	EXP-14: The AI-based system inputs should be monitored to be within the operational boundaries (both in terms of input parameter range and distribution) in which the AI/ML constituent performance is guaranteed, and deviations should be indicated to the relevant users and end users.	○	○	○	○	○
	EXP-15: The AI-based system outputs should be monitored to be within the specified operational performance boundaries, and deviations should be indicated to the relevant users and end users.	○	○	○	○	○
	EXP-16: The training and instructions available for the human end user should include procedures for handling possible outputs of the ODD and performance monitoring.	○	○	○	○	○
	EXP-17: Information concerning unsafe AI-based system operating conditions should be provided to the human end user to enable them to take appropriate corrective action in a timely manner.	○	○	○		

Building block	Objectives	Assurance Level				
		AL 1 DAL A SWAL1	AL 2 DAL B -	AL 3 DAL C SWAL2	AL 4 - SWAL3	AL 5 DAL D SWAL4
Human Factors for AI	**HF-01:** The applicant should design the AI-based system with the ability to build its own individual situational awareness.	O	O	O	O	O
	HF-02: The applicant should design the AI-based system with the ability to allow the end user to ask questions and to answer questions from the end user, in order to reinforce the end-user individual situational awareness.	O	O	O	O	O
	HF-03: The applicant should design the AI-based system with the ability to modify its individual situational awareness on end-user request.	O	O	O	O	O
	HF-04: If a decision is taken by the AI-based system, the applicant should design the AI-based system with the ability to request from the end-user a cross-check validation. Corollary objective: The applicant should design the AI-based system with the ability to cross-check and validate a decision made by the end user automatically or on request	O	O	O	O	O
	HF-05: For complex situations under normal operations, the applicant should design the AI-based system with the ability to identify suboptimal strategy and propose through argumentation an optimised solution. Corollary objective: The applicant should design the AI-based system with the ability to accept rejection required by the end user on the proposal.	O	O	O	O	O
	HF-06: For complex situations under abnormal operations, the applicant should design the AI-based system with the ability to identify the problem, share the diagnosis including the root cause, the resolution strategy and the anticipated operational consequences. Corollary objective: The applicant should design the AI-based system with the ability to consider the arguments shared by the end user.	O	O	O	O	O
	HF-07: The applicant should design the AI-based system with the ability to detect poor decision-making by the end user in a time-critical situation.	O	O	O	O	O
	HF-08: The applicant should design the AI-based system with the ability to take the appropriate action outside of a collaboration scheme, in case of detection of poor decision-making by the end user in a time-critical situation.	O	O	O	O	O
	HF-09: The applicant should design the AI-based system with the ability to negotiate, argue, and support its positions.	O	O	O	O	O
	HF-10: The applicant should design the AI-based system with the ability to accept the modification of task allocation / task adjustments (instantaneous/short-term).	O	O	O	O	O
	HF-11: The applicant should design the AI-based system with the ability to understand through the end-user responses or his or her action that there was a misinterpretation from the end user.	O	O	O	O	O
	HF-12: The applicant should design the AI-based system with the ability to notify the end user that he or she misunderstood the information provided through spoken natural language.	O	O	O	O	O

Building block	Objectives	Assurance Level				
		AL 1 DAL A SWAL1	AL 2 DAL B -	AL 3 DAL C SWAL2	AL 4 - SWAL3	AL 5 DAL D SWAL4
Human Factors for AI	**HF-13:** In case of misinterpretation, the applicant should design the AI-based system with the ability to resolve the misunderstanding through repetition, modification of the modality of interaction and/or with the provision of additional rationale on the initial information.	○	○	○	○	○
	HF-14: The applicant should design the AI-based system with the ability to making room for dialogue turns by other participants, keeping silent when needed and not hindering the user (while user is engaged in a complex activity), sticking to information that answers a given question by other participants, etc.).	○	○	○	○	○
	HF-15: In case spoken natural language is used, the applicant should design the AI-based system with the ability to provide information regarding the associated AI-based system capabilities and limitations.	○	○	○	○	○
	HF-16: The applicant should design the syntax of the spoken procedural language so that it can be learned easily by the end user.	○	○	○	○	○
	HF-17: The applicant should design the AI-based system with the ability to transition from verbal natural language to verbal procedural language depending on its own perception of the performance of the dialogue, the context of the situation and the characteristics of the task.	○	○	○	○	○
	HF-18: The applicant should design the gesture language syntax so that they are intuitively associated with the command that they are supposed to trigger.	○	○	○	○	○
	HF-19: The applicant should design the AI-based system with the ability to filter the intentional gesture used for language from non-intentional gesture, such as spontaneous gestures that are made to complement verbal spoken language.	○	○	○	○	○
	HF-20: In case gesture non-verbal language is used, the applicant should design the AI-based system with the ability to recognise the end-user intention.	○	○	○	○	○
	HF-21: In case gesture non-verbal language is used, the applicant should design the AI-based system with the ability to acknowledge the end-user intention with appropriate feedback.	○	○	○	○	○
	HF-22: In case spoken natural language is used, the applicant should design the AI-based system so that it can be deactivated for the benefit of other modalities in case of degradation of the interaction performance.	○	○	○	○	○
	HF-23: The applicant should design the AI-based system with the ability to combine or adapt the interaction modalities depending on the characteristics of the task, the operational event, and/or the operational environment.	○	○	○	○	○
	HF-24: The applicant should design the AI-based system with the ability to automatically adapt the modality of interactions to the end-user states, the situation, the context and/or the perceived end user's preferences.	○	○	○	○	○
	HF-25: The applicant should design the AI-based system to minimise the likelihood of design related error made by the end-user .	○	○	○	○	○
	HF-26: The applicant should design the AI-based system to minimise the likelihood of design related error made by the Human-AI Teaming.	○	○	○	○	○

Building block	Objectives	Assurance Level				
		AL 1 DAL A SWAL1	AL 2 DAL B -	AL 3 DAL C SWAL2	AL 4 - SWAL3	AL 5 DAL D SWAL4
Human Factors for AI	**HF-27:** The applicant should design the AI-based system to minimise the likelihood of HAIRM related error.	○	○	○	○	○
	HF-28: The applicant should design the AI-based system to be tolerant to end-user error .	○	○	○	○	○
	HF-29: The applicant should design the AI-based system so that in case the end user make an error while interacting with AI-based system, the opportunities exist to detect the error .	○	○	○	○	○
	HF-30: The applicant should design the AI-based system so that once an error is detected, the AI-based system should provide efficient means to inform the end-user and correct the error .	○	○	○	○	○

Safety risk mitigation	**SRM-01**: Once activities associated with all other building blocks are defined, the applicant should determine whether the coverage of the objectives associated with the explainability and learning assurance building blocks is sufficient or if an additional dedicated layer of protection, called hereafter safety risk mitigation (SRM), would be necessary to mitigate the residual risks to an acceptable level.	●	●	○	○	
	SRM-02: The applicant should establish SRM means as identified in Objective SRM-01.	●	●	○	○	

3. Additional risk-based levelling of information-security-related objectives

The following applies to domains where information security measures may be assigned a security assurance level (SAL) (e.g. for the product certification domain, see AMC 20-42).

Applicability by Security Assurance Level	
●	The objective should be satisfied with independence.
○	The objective should be satisfied.
	The satisfaction of the objective is to be negotiated between the applicant and the competent authority.

Objectives	SAL 3	SAL 2	SAL 1
IS-01: For each AI-based system and its data sets, the applicant should identify those information security risks with an impact on safety, identifying and addressing specific threats introduced by AI/ML usage.	○	○	○
IS-02: The applicant should document a mitigation approach to address the identified AI/ML-specific security risk.	●	○	
IS-03: The applicant should validate and verify the effectiveness of the security controls introduced to mitigate the identified AI/ML-specific security risks to an acceptable level.	●	○	

E. Annex 1 — Anticipated impact on regulations and MOC for major domains

The EASA Basic Regulation, beyond its main objective to establish and maintain a high uniform level of civil aviation safety in the Union, further aims to promote innovation, particularly by laying down requirements and procedures that are performance-based.

Considering the potential application of AI/ML solution in all the domains under the remit of the Agency, EASA intends to define a common policy that can be applied to the whole of the EU civil aviation regulatory framework, rather than issue domain-specific guidance.

This Annex provides an analysis of the anticipated impact on aviation regulations and on the means of compliance to the current regulations for the various impacted domains.

1. Product design and operations

1.1. Anticipated impact of the introduction of AI/ML on the current regulations

In the product design and certification domain, the current implementing rules (Part 21) and CSs already offer an open framework for the introduction of AI/ML solutions.

In particular, requirements such as CS 25/27/29.1301, 1302, 1309, 1319 or SC-VTOL.2500, 2505, 2510 are considered to still be valid for evaluating the safety of AI-based systems, provided additional means of compliance and standards are developed to answer the gap identified in the building blocks of the AI Roadmap.

For AI Level 1A, 1B or 2A applications, no impact on the EU aviation regulatory framework in relation to certification is deemed necessary. For higher AI Levels (2B and 3), this assumption will need to be revisited when working on further updates to this document.

In the Air Operations domain, the current regulatory framework (Regulation (EU) No 965/2012 (Air OPS Regulation) in its general parts related to organisation requirements (Part-ORO) contains provisions based on safety management principles that allow operators to identify risks, adopt mitigating measures and assess the effectiveness of these measures in order to manage changes in their organisation and their operations (ORO.GEN.200). This framework permits the introduction of AI/ML solutions; however, certain existing AMC and GM will need to be revised and new AMC and GM will need to be developed in relation to AI/ML applications.

More specific provisions in the Air OPS Regulation, related to specific type of operations and specific categories of aircraft, may also need to be revised depending on the specific AI Level 1 or 2A application.

AI Level 2B is expected to have a more significant impact on the Air OPS Regulation, particularly for all aspects related to HAT and task sharing. The specific rules on operational procedures and aircrew along with the associated AMC and GM will need to be revised as a minimum.

AI Level 3A will require a deeper assessment on their regulatory impact on Air Operations particularly on the requirements for air crew. This assumption will need to be revisited when working on further updates to this document.

1.2. Anticipated impact of AI/ML guidance on the current AMC/MoC framework

1.2.1 Summary

The objectives identified in this document are anticipated to provide a sufficient framework in view of approving Level 1 AI applications (both for 1A and 1B) and Level 2A. For Level 2B a more significant impact on the acceptable means of compliance framework is expected, including the creation of specific industry standards to support the guidance on the novel concept of HAT.

For AI Level 3, additional means of compliance will need to be identified when working on further updates to this document.

Although the technology available today may be sufficient to support Levels 1 and 2A AI applications, the ramping up to AI of Level 2B and later on to levels 3 will likely require further breakthroughs in the capability of communication and reasoning, in order to enable more autonomous AI.

The anticipated MOC will surely need to be completed based on the discussions triggered within certification projects, as well as based on industrial standards such as the ones that are under development in the working groups EUROCAE/SAE WG-114/G-34, ISO/IEC SC42 and CEN CENELEC JTC21.

For the first applications, it will be necessary to establish the certification framework addressing the installation and certification of AI-based systems for a given project. That could be achieved by the preparation of Certification Review Items (CRI) using the guidelines from this document.

1.2.2 Detailed analysis

Trustworthiness analysis

From a safety and security assessment perspective, the current guidance (e.g. AMC25.1309, AC 27.1309, AC 29.1309 or MOC VTOL.2510) is fully applicable, as reflected in Sections C.2.2 and C.2.3. The Ethical guidelines provided in Section C.2.4 are mostly novel and constitute one of the impacts of considering AI/ML solutions compared to traditional product certification approaches.

Learning assurance

When dealing with development assurance, the current means of compliance for system, software and hardware development assurance are not sufficient to address the specificities of learning processes (i.e. data management + learning assurance), and need to be complemented through the guidelines for AI assurance (Section C.3) when dealing with the development of the AI/ML-based subsystem. For other (sub)systems not developed with or not embedding AI/ML solutions, the current applicable system, software and hardware development assurance guidance still applies.

Explainability

The need for development explainability is specific to the use of AI/ML solutions is a new MOC. It builds however on some existing guidance; in particular, the applicable human factors guidance already used in certification could provide a sufficient layer of MOC for Level 1A AI/ML applications.

Human factors guidance

From Level 1B on, the concept of operational AI explainability is the cornerstone of the AI-specific human factors. The concept of Human AI-teaming is then developed for Levels 2A and 2B.

Safety risk mitigation

The need for residual risk assessment is dependent on the capacity of the applicant to meet the applicable objectives of the learning assurance and AI explainability building blocks. Even if the risk mitigation foresees the use of traditional MOC (e.g. safety nets), the development of novel methods of mitigation will need to be investigated.

Part 21 AMC/GM for design

The technical particularities of AI technology might require a need to adapt or introduce new AMC & GM related to the following Part 21 points:

— 21.A.3A 'Failures, malfunctions and defects' with regard to potentially new methodologies needed for the analysis of data required to identify deficiencies in the design of AI/ML constituents;

— 21.A.31 'Type design' with regard to guidance in the identification of the AI-related data that constitutes the type design;

— 21.A.33 'Inspections and tests' and 21.A.615 'Inspection by the Agency' with regard to guidance to ensure adequate Agency review of data and information related to the demonstration of compliance;

— 21.A.55, 21.A.105 and 21.A.613 'Record-keeping' with regard to guidance in the identification of the AI-related design information that needs to be retained and accessible;

— 21.A.91 'Classification of changes to a type-certificate' with regard to guidance in the major/minor classification of changes to AI-related approved type design.

2. ATM/ANS

2.1. Current regulatory framework relevant to the introduction of AI/ML

In addition to the Basic Regulation, Regulation (EU) 2017/373, applying to providers of ATM/ANS and other air traffic management network functions, lays down common requirements for:

(a) the provision of ATM/ANS for general air traffic, in particular for the legal or natural persons providing those services and functions;

(b) the competent authorities and the qualified entities acting on their behalf, which perform certification, oversight and enforcement tasks in respect of the services referred to in point (a);

(c) the rules and procedures for the design of airspace structures.

Regulation (EU) 2017/373 was supplemented regarding interoperability with Regulation (EC) No 552/2004[20], that has been repealed with a transition period expiring on 12 September 2023. In this context, it was necessary to introduce a new regulatory framework in relation to ATM/ANS systems and ATM/ANS constituents (referred to as 'ATM/ANS equipment') that ensures the safe, interoperable, and efficient provision of ATM/ANS services. The new regulatory framework proposed with Opinion No 01/2023 will enable the conformity assessment of certain ATM/ANS equipment by means of certification or declaration(s) as well as the approval of organisations involved in their design and/or production.

These Regulations open the path to the use of Level 1 and Level 2 AI in ATM/ANS. For higher AI Level 3, this assumption will need to be revisited when working on further updates to this document.

2.2. Anticipated impact of AI/ML guidance on the current AMC and GM

2.2.1. Summary

EASA has issued a comprehensive set of AMC and GM to the ATM/ANS (Regulation (EU) 2017/373) supporting ATM/ANS service providers in complying with the requirements of the Regulation.

The objectives identified in this document are anticipated to provide an initial framework in view of approving Level 1 and Level 2 AI applications (i.e. Level 1A, 1B, 2A and 2B), to be used by applicants to define their processes in order to achieve these objectives. For Level 3 AI, this assumption will also need to be revisited when working on further updates to this document.

The current AMC will surely need to be completed based on the guidance material delivered, as well as based on industrial standards such as the ones that are under development in the working groups EUROCAE/SAE WG-114/G-34, ISO/IEC SC42 and CEN CENELEC JTC21.

Looking at the set of regulations that are expected to be adopted based on Opinion No 01/2023 on conformity assessment, it is too early at this stage to clearly identify how they will be impacted by the certification or declaration of 'ATM/ANS equipment'. However, it can be anticipated that some impacts could be envisaged at the level of the detailed certification specifications or associated special conditions.

2.2.2. Detailed analysis

The following is an initial list of the Regulation (EU) 2017/373 AMC which could need adaptations:

ANNEX III — Part-ATM/ANS.OR – AMC6 ATM/ANS.OR.C.005(a)(2) Safety support assessment and assurance of changes to the functional system, specifically on the software assurance processes

ANNEX III — Part-ATM/ANS.OR – AMC1 ATM/ANS.OR.C.005(b)(1) Safety support assessment and assurance of changes to the functional system

ANNEX III — Part-ATM/ANS.OR – AMC1 ATM/ANS.OR.C.005(b)(2) Safety support assessment and assurance of changes to the functional system on the monitoring aspects

[20] Note: Regulation (EC) No 552/2004 was repealed by the Basic Regulation, but some provisions remain in force until 12 September 2023. To replace those provisions, a rulemaking task (RMT.0161) has been initiated.

ANNEX IV — Part-ATS – AMC1 ATS.OR.205(b)(6) Safety assessment and assurance of changes to the functional system on the monitoring of introduced changes

ANNEX IV — Part-ATS – AMC4 ATS.OR.205(a)(2) Safety assessment and assurance of changes to the functional system, specifically on the software assurance processes

ANNEX XIII — Part-PERS – AMC1 ATSEP.OR.210(a) Qualification training

Of course, the associated GM could be impacted as well.

3. Aircraft production and maintenance

3.1. Anticipated impact of the introduction of AI/ML on the current regulations

Regulation (EU) No 1321/2014, covering continuing airworthiness and approval of related organisations, is not very specific about technical details and generally contains higher-level requirements. It already addresses the use of software or the use of test equipment and tools (e.g. 'use of a software tool for the management of continuing airworthiness data', 'software that is part of the critical maintenance task'). Software making use of AI and/or ML could be covered under those requirements, including such software within test equipment.

However, the wording, being generic in many areas, still assumes a conventional way of planning and performing maintenance, meaning a *task-based approach*. Maintenance is divided into manageable portions of work (called 'tasks') which means human interference with the product at a defined point in time as a closed action which is signed off by humans when finished, with the product being released to service by explicit human action and signature.

Level 1 AI-based systems, with the human in command and in the specific case of maintenance closing out any activity by human signature and explicit release to service by human action, do not contradict this philosophy.

For Level 2 AI-based systems, this may require more attention, as humans still need to not only oversee, but also to explicitly close off the work performed by the systems with their signature. This may be possible within the frame of the current regulation but may limit the actions which can be carried out by systems.

Level 3 AI-based systems are not in line with the current regulation and would definitely require major changes, as the philosophy of explicit demonstration of airworthiness and release to service by humans would basically change to a withdrawal from service by systems finding lack of airworthiness.

It should also be noted that maintenance is a much more international business with more than a hundred states of registry being responsible compared to type certification with only about a dozen of states of design of large aeroplanes being responsible. This includes states with completely different regulations and hence will probably require a lot of international cooperation to harmonise the applicable regulations, guidance and standards.

3.2. Anticipated impact of AI/ML guidance on the current MoC framework

In the maintenance domain, there is no MoC framework comparable to the one used in certification.

Additionally, a significant part of the approval is done by the competent authorities (NCAs), and the regulation makes specific reference to 'officially recognised standards' (industry standards, national standards) so the complete overall framework of applicable guidance is not that clearly defined, rendering thus the impact of AI/ML not that easy to be evaluated. Industry standards (e.g. SAE) may be used to show compliance with certain requirements.

'Officially recognised standards' as mentioned in the AMC material 'means those standards established or published by an official body, being either a natural or legal person, and which are widely recognised by the air transport sector as constituting good practice'. This allows the use of future standards on AI/ML developed by recognised official bodies (like ASD-STAN, EUROCAE, RTCA, SAE, ASTM, ISO) for demonstrating compliance with certain requirements to the approving authority.

4. Training / FSTD

4.1. Anticipated impact of the introduction of AI/ML on the current regulations

The regulatory requirements for aircrew training are to be found in different Annexes to Regulation (EU) No 1178/2011 (the Aircrew Regulation).

In more detail, regulatory requirements are set in:

— Annex I (Part-FCL) in relation to licensing and training;

— Annex II (Part-ORA) in relation to organisational approvals.

Additional elements of flight crew training pertaining to the crew employed by operators are contained in the Air OPS Regulation.

Those regulations are mainly based on former Joint Aviation Authorities (JAA) regulatory requirements that were drafted almost 2 decades ago. All the structure of licensing and organisation approval therefore refers to traditional methodologies in which the technological contribution is limited to the use of computer-based training (CBT) solutions for the delivery of theoretical elements and to aircraft and flight simulation training devices (FSTDs) to deliver practical flight training elements. Additionally, some reference to distance learning provisions are present allowing certain flexibility for remote training.

The field of support of AI/ML solutions in the training domain may range from organisational aspects to monitoring functions up to more practical solutions in training delivery and performance assessment. The main impact will be on:

— the definition section to include the AI/ML constituents;
— the description of training programme delivery methodologies to address new technologies for administering the training courses;
— the crediting criteria for the use of AI/ML solutions; and
— organisation requirements in which the data management, analysis and correlation may play a role.

In any case, it is advisable that the initial use of AI/ML solutions in Aircrew training should be targeted to ground elements and simulator tasks.

4.2. Anticipated impact of AI/ML guidance on the current AMC/MOC framework

In support of the previous considerations, the AMC for the above-mentioned implementing rules shall be reviewed and updated to foresee the new technological solutions and to address the specificities of AI/ML solutions.

This review could run in parallel to the update of the regulatory framework which is already ongoing to incorporate new technologies and to accommodate emerging needs stemming from:

— new training needs for emerging aircraft concepts and their operations (e.g. VTOL or UAS);

— new training devices (e.g. virtual or augmented reality).

The Aircrew Regulation is not intended to certify products and does not address the design process, therefore all the elements of the ML model:

— trustworthiness analysis;

— learning assurance;

— explainability;

— safety risk mitigation

would need an effort to be created or tailored to the purpose.

5. Aerodromes

5.1. Current regulatory framework relevant to the introduction of AI/ML

In addition to the Basic Regulation, Regulation (EU) No 139/2014[21] lays down requirements and administrative procedures related to:

(a) aerodrome design and safety-related aerodrome equipment;

(b) aerodrome operations, including apron management services and the provision of groundhandling services;

(c) aerodrome operators and organisations involved in the provision of apron management and groundhandling services[22];

(d) competent authorities involved in the oversight of the above organisations, certification of aerodromes and certification/acceptance of declarations of safety-related aerodrome equipment[23].

---

21 As subsequently amended by Commission Regulation (EU) 2018/401 regarding the classification of instrument runways, Commission Implementing Regulation (EU) 2020/469 as regards requirements for air traffic management/air navigation services, Commission Delegated Regulation (EU) 2020/1234 as regards the conditions and procedures for the declaration by organisations responsible for the provision of apron management services, and Commission Delegated Regulation (EU) 2020/2148 as regards runway safety and aeronautical data.

22 For groundhandling services and providers of such services, there are at this stage no detailed implementing rules. These are expected not earlier than 2024.

23 The oversight framework for safety-related aerodrome equipment will be developed in due course but is at the time of writing not yet in place, neither are the European certification specifications for such equipment.

This regulation, in its consolidated form, does not represent a hinderance to the use of Level 1 and 2 AI use cases. For AI Level 3, this statement might be revisited when the need would be brought to the attention of EASA by industry and overseen organisations, as well as manufacturers of safety-relevant aerodrome equipment.

5.2. Anticipated impact of AI/ML guidance on the current AMC and GM

The AMC and GM related to Regulation (EU) No 139/2014 support the implementation of the implementing rule requirements by the organisations concerned.

Most of the AMC and GM do not refer to specific technologies, so they do not impede the approval of Level 1 AI applications. For higher AI Levels (2 and 3), this statement might need to be revisited when the need by industry and overseen organisations, as well as manufacturers of safety-relevant equipment, would be brought to the attention of EASA.

5.3. Preliminary analysis

The following IRs and the related AMC and GM are relevant to the AI use cases further below:

— ADR.OPS.B.015 Monitoring and inspection of movement area and related facilities
— ADR.OPS.B.020 Wildlife strike hazard reduction
— ADR.OPS.B.075 Safeguarding of aerodromes

5.4. Anticipated impact of AI/ML guidance on the current and future CSs for aerodrome design and safety-related aerodrome equipment

The current CSs and Regulation (EU) No 139/2014 provide a comprehensive set of requirements for the design of aerodrome infrastructure and for some aerodrome equipment (as far as it exists stemming from the transposition of Annex 14). Once the future framework for safety-related aerodrome equipment exists, EASA will issue European certification specifications for such equipment. This process will allow for the further introduction of AI/ML solutions at aerodromes, if they fulfil the demands placed on them with respect to safety.

6. Environmental protection

6.1. Current regulatory framework relevant to the introduction of AI/ML

The essential environmental protection requirements for products are laid out in the Basic Regulation Articles 9 and 55 for manned and unmanned aircraft respectively, and in its Annex III. These requirements are further detailed in Part 21 (in particular point 21.B.85) as well as in CS-34 'Aircraft engine emissions and fuel venting', CS-36 'Aircraft noise' and CS-CO2 'Aeroplane CO2 Emissions'. For the majority of manned aircraft, the AMC and GM linked to these requirements are defined in the appendices to ICAO Annex 16 and in Doc 9501 'Environmental Technical Manual'.

6.2. Anticipated impact of AI/ML guidance on the current MOC framework

The AI/ML guidance for Level 1 and 2 systems is anticipated to have no impact on the current MOC framework for environmental protection. The impact of Level 3 AI/ML guidance will be assessed at a later stage. The safety-related guidelines in Chapter C of this document are anticipated to help provide adequate confidence in the functioning of AI/ML applications when demonstrating compliance with environmental protection requirements.

F. Annex 2 — Use cases for major aviation domains

1. Introduction

With the objective of ensuring that its guidelines will remain practical for the applicants, EASA has engaged with the aviation industry and stakeholders, in order to support the elaboration of the guidelines with actual use cases from the major aviation domains.

It is not the intention that each use case is complete and fulfils the full set of objectives described in this guidance document, but rather to evaluate that the objectives and proposed anticipated MOC are practical. This may result in a number of use cases not implementing all AI trustworthiness building blocks.

Before entering into the use cases, Table 6 below provides the audience with a description of how each use case has been classified as per Table 2.

		Domain					
		Aircraft design and operations		ATM/ANS		Aircraft production and maintenance	
EASA AI Roadmap AI Level (subsystem)	Function allocated to the (sub)systems (adapted HARVIS LOAT terminology)	Visual landing guidance system	Pilot assistance – radio frequency suggestion	AI-based augmented 4D trajectory prediction	Time-based separation + Optimum runway delivery	Controlling corrosion by usage-driven inspections	Damage detection in images
Level 1A Human augmentation	Support to information acquisition	camera + pre-processing	ATC radio communication	Data acquisition (FPL and updates, radar + weather)	Data acquisition (weather + radar)	Maintenance, environment, operator / manufacturer databases	infrared camera
	Support to information analysis	Runway object classification + bounding box + tracking/filtering algorithm	Voice recognition	4D trajectory calculation – Curtain + Climb and descent rate	Information preparation (pairs, applicable separation)	Predicted corrosion level + Time to inspect for corrosion	Damage classification
Level 1B Human assistance	Support to decision/action selection	x	Radio frequency suggestion for pilot validation	x		x	Support decision to repair for inspector validation
Level 2A Human-AI cooperation	Cooperative overridable automatic decision/action selection	x	x	x	Trajectory prediction + uncertainty calculation	x	x
	Cooperative overridable automatic action implementation	x	x	x	x	x	x
Level 2B Human-AI collaboration	Collaborative overridable automatic decision/action selection						
	Collaborative and overridable automatic action implementation						
Level 3A Semi-autonomous AI	Overridable automatic decision/action selection	x	x	x	x	x	x
	Overridable automatic action implementation	x	x	x	x	x	x
Level 3B Fully autonomous AI	Non-overridable automatic decision/action selection	x	x	x	x	x	x
	Non-overridable automatic action implementation	x	x	x	x	x	x

Table 6 — Classification applied to use cases

Where:

represents the AI-based system or subsystem; and

The AI/ML constituent is in blue.

2. Use cases — aircraft design and operations

2.1. Visual landing guidance system (derived from the CoDANN report use case)

This use case describes Daedalean's Visual Landing System (VLS), analysed in [CODANN] (Daedalean, 2020) and [FAAVLS] (Federal Aviation Administration, May 2022) reports. The text below references both reports.

2.1.1. Trustworthiness analysis — description of the system and ConOps

2.1.1.1. Description of the system

Objective CO-01: The applicant should identify the list of end users that are intended to interact with the AI-based system, together with their roles, their responsibilities and their expected expertise (including assumptions made on the level of training, qualification and skills).

Objective CO-02: For each end user, the applicant should identify which high-level tasks are intended to be performed in interaction with the AI-based system.

Objective CO-03: The applicant should determine the AI-based system taking into account domain-specific definitions of 'system'.

The VLS provides landing guidance for Part 91 (General Aviation) aircraft on hard-surface runways in daytime visual meteorological conditions (VMC), using a forward-looking high-resolution camera as the only external sensor.

During daytime VMC flight under visual flight rules (VFR), the system recognises and tracks hard-surface runways present in the field of view, and allows the operator to select the one intended for landing or use a pre-configured selection based on a flight plan. Once a runway has been selected and once the aircraft begins its final descent towards it, the VLS provides the position of the aircraft in the runway coordinate frame as well as horizontal and vertical deviations from a configured glide slope, similar to a radio-based instrument landing system (ILS). Uncertainties and validity flags for all outputs are also produced by the system.

See [FAAVLS; Section 1.2, Chapter 5] and [CODANN; Chapter 4] for details.

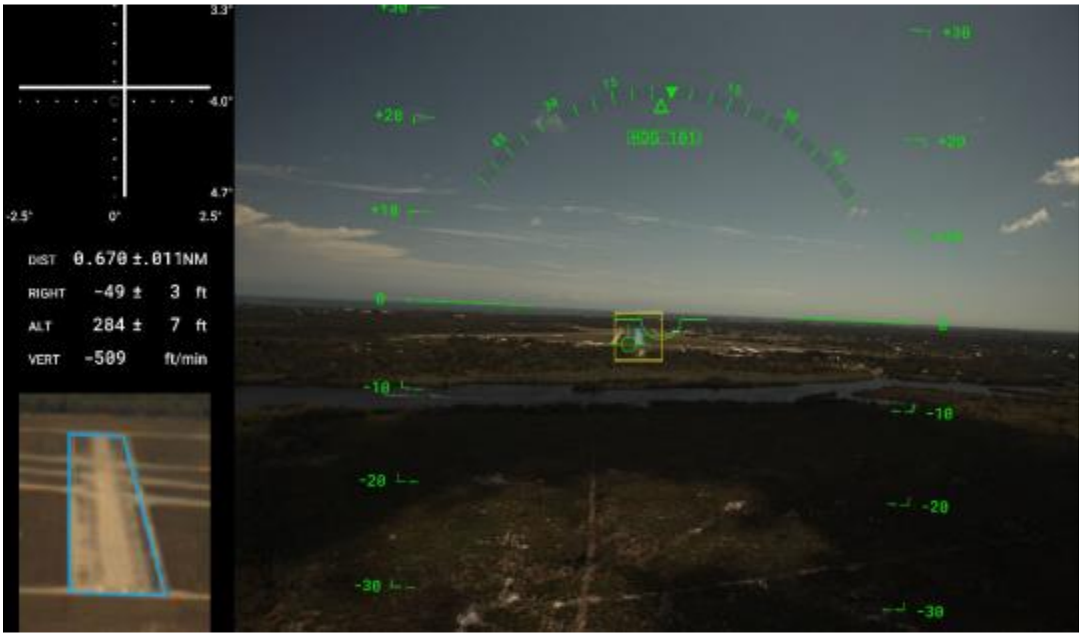


Figure 24— Development view of the system.

The definition of 'system' from ED-79A/ARP-4754A is taken as reference for this airborne application (i.e. a combination of inter-related items arranged to perform a specific function(s)).

2.1.1.2. Concept of operations

Objective CO-04: The applicant should define and document the ConOps for the AI-based system, including the task allocation pattern between the end user(s) and the AI-based system. Focus should be put on the definition of the OD and on the capture of specific operational limitations and assumptions.

See [CODANN; 4.1] and [FAAVLS; Chapter 3], containing detailed descriptions of possible ConOps.

Both reports consider foremost Level 1, limiting to the display of the guidance on a glass cockpit display without involving flight computer guidance.

However, coupling to an onboard autopilot (Levels 2 or 3) is also discussed (but is not part of the flight test campaign).

2.1.1.3. Description of the subsystems involved (inputs, outputs, functions)

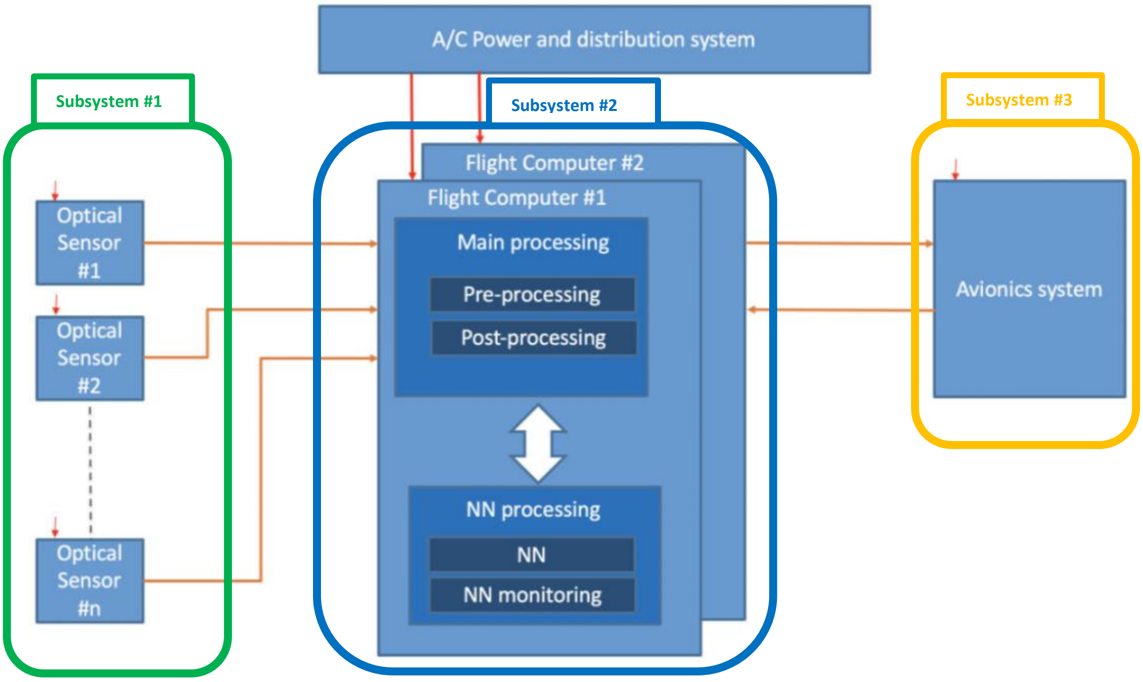


Figure 25 — System breakdown in subsystems and components (source (Daedalean, 2020))

Objective CO-06: The applicant should perform a functional analysis of the system.

The system is composed of three subsystems:

1. Perception, based on a high-resolution camera.
2. Pre-processing (traditional software), image analysis (neural network) and post-processing (traditional software; tracking and filtering).
3. Avionics display system supporting the system's operations.

Subsystem #2 is an AI-based subsystem while subsystems #1 and #3 are traditional subsystems.

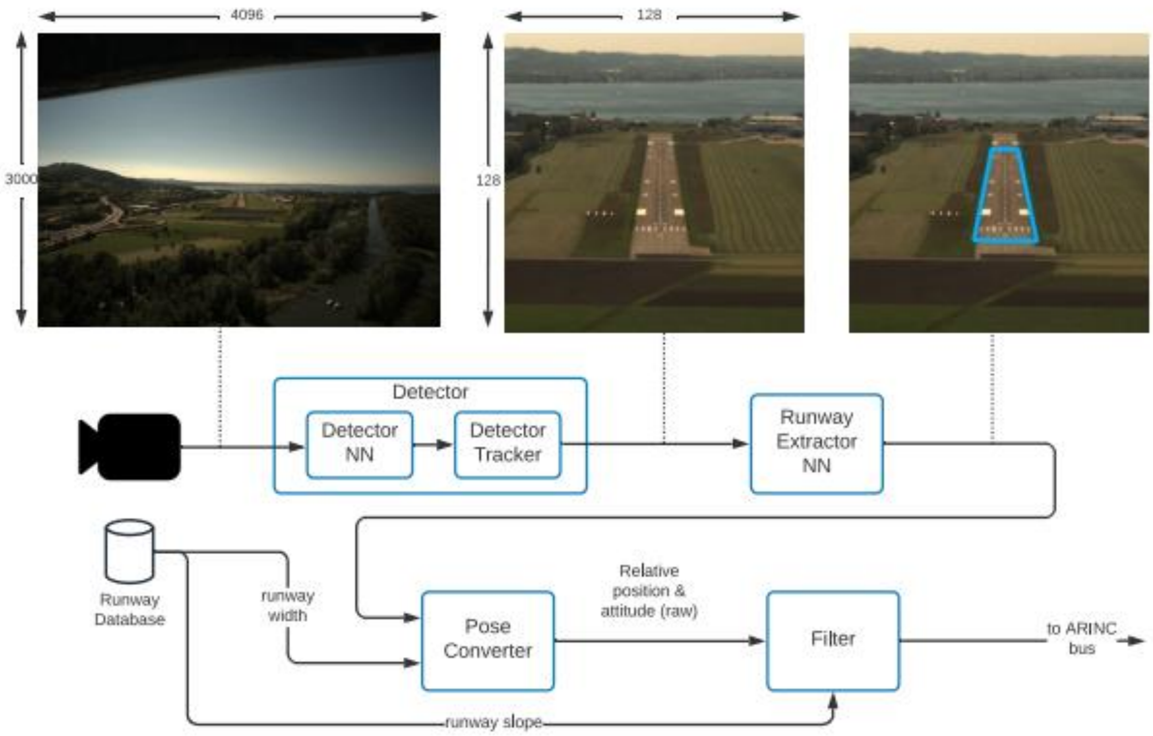


Figure 26 — System overview (source (Daedalean, 2020)

Following [CODANN; 9.2.2] and [FAAVLS; 8.2], a possible functional decomposition of the system is the following:

— F1: To detect a runway

This function is implemented through ML-based perception. At item level the following functions contribute to this system level function:

- F1.1: Capture real-time imagery data of the external environment of the aircraft
- F1.2: To pre-process the image
- F1.3: To detect the runway position in a given image
- F1.4: To track the runway position in an image
- F1.5: To output a crop of an image with a runway inside

— F2: To provide guidance to land on a runway

This function is implemented through ML-based perception and estimation/filtering components. It is the main function analysed in this report. At item level the following functions contribute to this system level function:

- F2.1: To pre-process a cropped image assumed to contain a runway
- F2.2: To compute the runway geometry from a runway crop
- F2.3: To compute the pose with respect to the runway
- F2.4: To filter the pose

- F2.5: To compute and output the lateral/vertical glidepath deviations from the runway

— F3: To monitor the system

At item level the following functions contribute to this system level function:

- F3.1: To monitor sensors
- F3.2: To monitor image characteristics
- F3.3: To continuously monitor internal system health
- F3.4: To test system components at power-up
- F3.5: To determine whether the system output should be enabled
- F3.6: To signal when a landing manoeuvre should be aborted

— F4: To interface with the aircraft systems

- F4.1: To receive the GPS data
- F4.2: To receive the digital terrain elevation data
- F4.3: To receive the phase of flight
- F4.4: To receive electrical power
- F4.5: To provide visual guidance to the pilot
- F4.6: To provide monitoring data to the display

The functional allocation to the subsystems and components can be done as follows:

Subsystem	Constituents and items	Allocated functions
#1	Optical sensor	F1.1 F3.1 F4.4
#3	NN processing	F1.3 F2.2
#4	Avionics display	F4.5 F4.6
#2	Main processing unit	All other functions

Table 1 — Functional allocation to the subsystems, constituents and items

2.1.1.4. Expected benefits and justification for Level 1

The application is intended to provide additional information to the pilot in the form of a runway image displayed in the cockpit from the moment the runway is detected (cruise phase/holding pattern) until the decision to land is confirmed or a go-around is performed.

Objective CL-01: The applicant should classify the AI-based system, based on the levels presented in Table 2, with adequate justifications.

The ***AI Level 1A 'Human augmentation'*** classification is justified by only providing additional/advisory information (***support to information analysis***) to the pilot without any suggestion for action or decision-making.

2.1.2. Trustworthiness analysis

2.1.2.1. Safety assessment

Objective SA-01: The applicant should perform a safety (support) assessment for all AI-based (sub)systems, identifying and addressing specificities introduced by AI/ML usage.

Preliminary FHAs can be found in [CoDANN; 9.2.4] and [FAAVLS; Chapter 8].

For the purpose of this use case discussion, the system can contribute to failure conditions up to Hazardous (as defined in the applicable CSs). More severe failure conditions should be considered in the case of linking the system to an autopilot, but this would trigger another classification for this AI-based system, probably up to Level 2.

Based on the discussions from the CoDANN report (Daedalean, 2020) Chapter 8, two types of metrics are considered for this use case:

— For the evaluation of the binary classification of the runway object, the precision and recall measures can be used to first select the best model and then to evaluate the operational performance.

— For the evaluation of the bounding box, the use of the Jaccard distance can be a useful metric for model selection.

Objective ICSA-01: The applicant should identify which data needs to be recorded for the purpose of supporting the continuous safety assessment.

Inputs (or quantities derived from them) can be used to ensure that input distribution assumptions are verified.

Outputs can be compared against GPS and ILS (when available), to identify discrepancies not explained by the precision level of these systems. This is how the system was evaluated in [FAAVLS; Section 4].

These discrepancies and the uncertainty estimates can be used to select which data to record.

2.1.2.2. Information security

Provision ORG-02: In preparation of the Commission Delegated Regulation (EU) 2022/1645 and Commission Implementing Regulation (EU) 2023/203 applicability, the organisation should assess the information security risks related to the design, production and operation phases of an AI/ML application.

As explained in [FAAVLS; Section 6.2], the data is collected and stored on the company's own infrastructure, along cryptographic checksum reducing information security risks.

Training might involve cloud computing, to access a large number of GPUs; the corresponding risks are mitigated by ensuring data integrity throughout the training process and a verification strategy that depends on the output of the training, but not the process. The same strategy is used to prevent issues when moving from the implementation to the inference environment.

2.1.3. Learning assurance

Objective DA-01: The applicant should describe the proposed learning assurance process, taking into account each of the steps described in Sections C.3.1.2 to C.3.1.12, as well as the interface and compatibility with development assurance processes.

The learning assurance process followed is the W-shaped process described in [CODANN]. See [FAAVLS; Section 6] and [CODANN; Chapter 10].

Due to the scope of the projects and reports, complete system requirements and item requirements capture and validation as requested via **Objective DA-02**, **Objective DA-04**, **Objective DA-05** and **Objective DA-06** were not explicitly captured in [CODANN,FAAVLS], although this would naturally be a central part of a real assessment.

They nevertheless appear implicitly throughout the reports, e.g. when the performance of the end system is analysed from that of its components in [FAAVLS; Chapter 8].

Objective DA-03: The applicant should describe the system and subsystem architecture, to serve as reference for related safety (support) assessment and learning assurance objectives.

The systems and subsystem architectures are carefully described in [FAAVLS; Chapter 5], which is then used in the safety assessment in [FAAVLS; Chapter 8] and throughout the application of the W-shaped process [FAAVLS; Chapter 6].

2.1.3.1. Data management

Objective DM-01: The applicant should define the set of parameters pertaining to the AI/ML constituent ODD.

The input space for this use case is the space of 512 x 512 RGB images that can be captured from a specific camera mounted on the nose of an aircraft, flying over a given region of the world under specific conditions, as defined in the ConOps and in the requirements.

The relevant operating parameters include the following (on the frame level):

— Altitude

— Glide slope

— Lateral deviation

— Distance to runway

— Time of day

— Weather

See [FAAVLS; Section 6.2] for details and [FAAVLS; Section 8.2] for an analysis of the coverage of a data set collected during the project.

With regard to the DQRs, their implementation (including the collection process) and verification are discussed in [FAAVLS; Section 6.2].

Objective DM-06: The applicant should identify data sources and collect data in accordance with the defined ODD, while ensuring satisfaction of the defined DQRs, in order to drive the selection of the training, validation and test data sets.

An analysis of the collected data is present in [FAAVLS; Section 8.2]. The set of operating parameters are first reviewed with respect to the set of requirements and with the ODD, to make a first evaluation of their intrinsic completeness in relation to the use case application. See also [CODANN; 6.2.8].

Objective DM-07: Once data sources are collected, the applicant should ensure the high quality of the annotated or labelled data in the data set.

In the context of this use case, the annotation task consists of marking each of the four runway corners in every image.

The review of the annotation is performed through a manual review and statistical analysis following ISO-2859-1, and comparison with other data sources.

See [FAAVLS; Section 6.2.4].

Objective DM-10: When applicable, the applicant should define and document the transformations to the pre-processed data from the specified input space into features which are effective for the performance of the selected learning algorithm.

This objective is not relevant for this use case, as there is no explicit feature extraction/engineering (use of convolutional neural networks).

Objective DM-11: The applicant should ensure that the data is effective for the stability of the learning process.

The adequacy of data for training is confirmed during the model training phase, described in [FAAVLS; Section 6.4].

Objective DM-12: The applicant should distribute the data into three separate and independent data sets which will meet the specified DQRs:

— the training data set and validation data set, used during the model training;

— the test data set used during the learning process verification, and the inference model verification.

The data is split into training, validation and test sets, carefully taking into account runways and landings (e.g. to prevent the validation or test set from containing only runways that have been trained on, even from any other approach on a same runway).

See [FAAVLS; Section 6.2]

Objective DM-13: The applicant should ensure validation and verification of the data, as appropriate, all along the data management process so that the data management requirements (including the DQRs) are addressed.

— Data completeness and representativeness

The set of operating parameters are first reviewed with respect to the set of requirements and with the ODD, to make a first evaluation of their intrinsic completeness in relation to the use case application.

The approach is completed by the definition of a distribution discriminator *D* using the ODIN method from the paper (Enhancing the reliability of out-of-distribution image detection in neural networks, 2018).

Refer to the CODANN Report (Daedalean, 2020), Section 6.2.8, for more information.

— Data accuracy

To demonstrate that the model was provided with correct data samples during the design phase, several sources of errors need to be shown to be minimal and independent, or else to be mitigated.

First, the systematic errors in the data, also called data bias are identified using statistical testing and mitigated.

In addition, specific attention is paid to single-source errors which could introduce bias in the resulting data sets. This type of error has been avoided by using the same source for data collection in operations as well.

Furthermore, labelling errors have been addressed by involving multiple independent actors in the labelling activity and its validation.

— Data traceability

The data sets undergo a conversion from the raw images format to 8bit RGB, removal of irrelevant information as necessary, and may be modified to enhance the colour, brightness and contrast. These transformations are fully reproducible and a trace of the changes to the origin of each data pair is recorded. This applies also to the annotations.

— Data sets independence

The training/validation and test data sets are created by independent groups. The test data set is not accessible during the design phase.

2.1.3.2. Learning process management

Objective LM-01: The applicant should describe the AI/ML constituents and the model architecture.

See [FAAVLS; Chapter 5], describing the convolutional deep neural network used for runway geometry extraction, including aleatoric uncertainty estimation.

More generally, the full system architecture is described in [FAAVLS; Chapter 5].

Objective LM-02: The applicant should capture the requirements pertaining to the learning management and training processes, including but not limited to:

— model family and model selection;

— learning algorithm(s) selection;

— cost/loss function selection describing the link to the performance metrics;

— model bias and variance metrics, and acceptable levels;

— model robustness and stability metrics, and acceptable levels;

— training environment (hardware and software) identification;

— model parameters initialisation strategy;

— hyper-parameters and parameters identification and setting;

— expected performance with training, validation and test data sets.

The data indicated in Objectives LM-01 and LM-02 is documented, including substantiation for the selection of the model architecture, learning algorithm selection as well as for the learning parameters selection.

See [FAAVLS; Section 6.3].

Objective LM-03: The applicant should document the credit sought from the training environment and qualify the environment accordingly.

The open-source software library TensorFlow is chosen, and the training is run on a compute cluster equipped with NVIDIA GPUs, on Linux-based operating system. See [FAAVLS; Section 6.3] and [CODANN2; Chapter 3]. Following the strategy of the latter, only minimal credit is taken from the training environment, as the verification relies mostly on properties of the inference model in the inference environment.

Objective LM-04: The applicant should provide quantifiable generalisation guarantees.

The approach to providing performance guarantees on the model is a combination of 'evaluation-based' and 'complexity-based' approaches: the former is outlined in [FAAVLS; Section 8.3]. A broad survey of different methods is provided in [CODANN; Section 5.3].

The integration with a classical system (pose filtering) allows to control the time dependency, and prevent errors probabilities from accumulating exponentially; see [FAAVLS; Section 8.6].

Objective LM-05: The applicant should document the result of the model training.

The resulting training curves and performance on the training and validation sets are recorded in the learning accomplishment summary (LAS).

Objective LM-06: The applicant should document any model optimisation that may affect the model behaviour (e.g. pruning, quantisation) and assess their impact on the model behaviour or performance.

No optimisation is performed at the level of the learning process. These optimisations would be applied at the implementation level; see the comments there.

Objective LM-07: The applicant should account for the bias-variance trade-off in the model family selection and should provide evidence of the reproducibility of the training process.

Convolutional deep neural networks are used, which theoretically have low bias but higher variance due to the number of parameters (model complexity); the latter is mitigated through the use of sufficient data.

The Bootstrapping and Jack-knife methods have been used to estimate bias and variance and support the model family selection.

To this purpose, the learning process is repeated several times with variations in the training data set to show that:

— the models have similar performance scores on training and validation data sets;

— the selected model is not adversely impacted by a small change in the training data set.

Objective LM-08: The applicant should ensure that the estimated bias and variance of the selected model meet the associated learning process management requirements.

The learning process is repeated multiple times on various subsets of the training data to show that the models are not highly dependent on a small part of the training data.

The bias of the model is estimated in other objectives, as this represents the model performance.

Objective LM-09: The applicant should perform an evaluation of the performance of the trained model based on the test data set and document the result of the model verification.

The resulting performance of the model on the test data set is recorded in a LAS.

Objective LM-10: The applicant should perform a requirements-based verification of the trained model behaviour and document the coverage of the AI/ML constituent requirements by verification methods.

A requirements-based verification of the model is outlined in [FAAVLS: Section 8.3]. The requirements are expressed as conditions on the distribution of absolute errors on sequences of frames.

Objective LM-11: The applicant should provide an analysis on the stability of the learning algorithms.

The performance of the mode (loss, metrics) is analysed over its training via gradient descent, to rule out behaviours that would be incompatible with good generalisation abilities (e.g. overfitting, underfit-ting, large oscillations, etc.). See [FAAVLS; Section 6.4.1].

Objective LM-12: The applicant should perform and document the verification of the stability of the trained model.

Objective LM-13: The applicant should perform and document the verification of the robustness of the trained model in adverse conditions.

Aspects of the model robustness are analysed through saliency maps in [FAAVLS; Section 6.5.3].

It is crucial to understand how errors at the level of the model will propagate to other components; a sensitivity analysis is carried out in [FAAVLS; Section 8.5.1], quantifying the effect of model errors on the pose estimate.

Objective LM-14: The applicant should verify the anticipated generalisation bounds using the test data set.

See [FAAVLS; Section 6.5.2, Section 8.3] for an analysis of the performance of the model on various kinds of data (training, validation, test; seen or unseen runways).

2.1.3.3. Trained model implementation

Objective IMP-01: The applicant should capture the requirements pertaining to the implementation process.

Objective IMP-02: Any post-training model transformation (conversion, optimisation) should be identified and validated for its impact on the model behaviour and performance, and the environment (i.e. software tools and hardware) necessary to perform model transformation should be identified.

Objective IMP-04: The applicant should verify that any transformation (conversion, optimisation, inference model development) performed during the trained model implementation step has not adversely altered the defined model properties.

Objective IMP-05: The differences between the software and hardware of the platform used for training and the one used for verification should be identified and assessed for their possible impact on the inference model behaviour and performance.

The transition between implementation and inference environment would follow the strategy outlined in [CODANN2; Chapter 3], where most of the verification takes place directly on the inference model, and minimal credit is needed from the implementation environment or transformations to the inference environment.

Due to time constraints, the flight tests from [FAAVLS] did not run on production hardware, but on uncertified COTS (e.g. GPUs) hardware, which is described in [FAAVLS; Section 6.6]. An actual system would also follow the recommendations from [CODANN2; Chapter 3].

With regard to **Objective IMP-06**, **Objective IMP-07**, **Objective IMP-08**, and **Objective IMP-09**, a similar strategy to the corresponding LM objectives would be adopted, on the inference model in the inference environment.

2.1.3.4. Data and learning verification

Objective DM-14: The applicant should perform a data and learning verification step to confirm that the appropriate data sets have been used for the training, validation and verification of the model and that the expected guarantees (generalisation, robustness) on the model have been reached.

An outline of the execution of this objective is present in [FAAVLS; Section 2.3.7].

In particular, it is verified that the test set has not been used during development (since it was not annotated until the test phase) and that the operational space had been correctly identified from the model's behaviour.

2.1.3.5. Configuration management

Objective CM-01: The applicant should apply all configuration management principles to the AI/ML constituent life-cycle data, including but not limited to:

— identification of configuration items;

— versioning;

— baselining;

— change control;

— reproducibility;

— problem reporting;

— archiving and retrieval, and retention period.

All artifacts, from the original data to trained models, are carefully tracked, including: checksums, sources, production process, etc. See [FAAVLS: Section 6.2.2].

This permits configuration management over the full life cycle of the pipeline, including reproducibility, change control, baselining, etc.

2.1.3.6. Quality assurance

Objective QA-01: The applicant should ensure that quality/process assurance principles are applied to the development of the AI-based system, with the required independence level.

ISO-2859-1 is applied to carry out quality assurance of the manual annotations. See [FAAVLS: Section 6.2.4].

The rest of the processes are automated, and the underlying tools are qualified at the required level.

2.1.4. Development & post-ops AI explainability

Objective EXP-03: The applicant should identify and document the methods at AI/ML item and/or output level satisfying the specified AI explainability needs.

Image saliency analysis is used in [FAAVLS; Section 6.5.3] to analyse which parts of a given input image are most important for the output of the neural network. This allows identifying potential undesired behaviour or biases, or possible misidentifications of the input space (e.g. use of non-generic runway markings or adjacent objects).

Objective EXP-04: The applicant should provide the means to record operational data that is necessary to explain, post operations, the behaviour of the AI-based system and its interactions with the end user.

The system inputs and outputs are recorded in real time, including the output of dissimilar sensors for comparison. See [FAAVLS; Section 4.1]. When limited storage space is available, the recording can be limited to the outputs, or to situations where a difference with other sensors or high uncertainty are detected.

2.1.5. AI operational explainability

Objective EXP-06: The applicant should present explanations to the end user in a clear and unambiguous form.

The prototype flight display shows a zoomed-in inlet of the runway and its corners, as detected by the system. The design of the system implies that if the corners are precisely positioned, then the guidance will be accurate. This provides the end user with a powerful explanation of the quality of the output, in addition to the provided measures of uncertainty. See [FAAVLS: Section 4.2.1].

Objective EXP-12: For each output relevant to the task(s), the applicant should ensure the validity of the specified explanation, based on actual measurements (e.g. monitoring) or on a quantification of the level of uncertainty.

The system includes an uncertainty estimation component, estimating both the aleatoric and epistemic uncertainties.

See [FAAVLS; Chapter 5, Section 8.4].

Objective EXP-17: Information concerning unsafe AI-based system operating conditions should be provided to the human end user to enable them to take appropriate corrective action in a timely manner.

When OoD samples are detected or when the system estimates a high uncertainty, the system outputs are disabled, and the system's assistance cannot be used.

2.1.6. Safety risk mitigation

Objective SRM-01: Once activities associated with all other building blocks are defined, the applicant should determine whether the coverage of the objectives associated with the explainability and learning assurance building blocks is sufficient or if an additional dedicated layer of protection, called hereafter safety risk mitigation (SRM), would be necessary to mitigate the residual risks to an acceptable level.

In this use case, it is considered that all objectives related to the trustworthiness analysis, learning assurance and explainability building blocks can be fully covered.

Objective SRM-02: The applicant should establish SRM means as identified in Objective SRM-01.

No SRM mitigations are identified in SRM-01.

2.2. Pilot assistance — radio frequency suggestion

2.2.1. Trustworthiness analysis — description of the system and ConOps

2.2.1.1. Description of the system

Objective CO-03: The applicant should determine the AI-based system taking into account domain-specific definitions of 'system'.

An example of an AI Level 1B application for pilot assistance may be voice recognition and suggestion of radio frequencies.

The application recognises radio frequencies from ATC voice communications and suggests to the pilot a frequency that has to be checked and validated by the pilot before tuning the radio accordingly (e.g. tuning the standby VHF frequencies).

2.2.1.2. Expected benefits and justification for Level 1

The application is expected to reduce workload or help the pilot to confirm the correct understanding of a radio frequency in conditions of poor audio quality.

Objective CL-01: The applicant should classify the AI-based system, based on the levels presented in Table 2, with adequate justifications.

The Level 1B classification is justified by providing support to the pilot in terms of gathering the information and suggesting it to the pilot for validation before any action is taken, i.e. support to decision-making. The frequency may be either displayed to the pilot who then will tune it manually or may be pushed automatically into the avionics after acceptance of the pilot. The two cases will require a different level of assessment.

2.2.2. Trustworthiness analysis — safety and security assessment

Objective SA-01: The applicant should perform a safety (support) assessment for all AI-based (sub)systems, identifying and addressing specificities introduced by AI/ML usage.

A risk of complacency and over-reliance on the applications exists.

Objective IS-01: For each AI-based (sub)system and its data sets, the applicant should identify those information security risks with an impact on safety, identifying and addressing specific threats introduced by AI/ML usage.

If the application is integrated with the avionics with the possibility to exchange data, the check and validation function, as well as data integrity and security aspects, will have to be further assessed.

2.3. Pilot AI teaming — Proxima virtual use case

Most of the use cases proposed currently by industry target Level 1 AI applications. Beyond research, there is to date no compelling example of a Level 2 AI use case that could be taken as reference to illustrate the additional guidance developed in this Issue 02 of the EASA Concept Paper.

Consequently, as a first approach to Level 2 AI, it was decided to develop a virtual use case under the AI task force of the EASA Scientific Committee.

Proxima is an example of a virtual co-pilot meant to support single-pilot operations of large transport aeroplanes, by enabling human-AI teaming and crew resource management capabilities comparable to those of multi-crew operations. What follows is a first description of the capabilities anticipated when the maximum potential of level 2B AI-based systems will be reached.

2.3.1. Trustworthiness analysis — description of the system and ConOps

2.3.1.1. High-level task(s) and AI-based system definition

Objective CO-01: The applicant should identify the list of end users that are intended to interact with the AI-based system, together with their roles, their responsibilities and their expected expertise (including assumptions made on the level of training, qualification and skills).

The main end user interacting with Proxima is the pilot-in-command (PIC). A second layer of end users include the air traffic controller (ATCO).

The PIC role and responsibilities are anticipated to be similar to those allocated to the PIC in multi-crew operations. However, level 2B AI is by definition capable of automating certain decisions, thus reducing partially the 'authority' of the PIC for certain tasks. The expertise of the pilot shall be the current one with additional training to deal with the AI-based system and the new type of operations.

The ATCO role, responsibilities and expertise remain strictly identical to current operations; however, with the necessary awareness that he or she is also interacting with an AI-based system.

Objective CO-02: For each end user, the applicant should identify which high-level tasks are intended to be performed in interaction with the AI-based system.

In single-pilot operation aircraft, Proxima and the pilot will share tasks and will have a common set of goals. Through perception and analysis, Proxima will learn from the situations encountered and will be able to continually adapt to the current situation to assist the crew in its decision-making process. Proxima will also have the ability to respond appropriately to displayed information. In addition, it will also identify any mismatch between the information that it has that is relevant to a pilot's decision and the information available to the pilot via displays and other means. It will then respond appropriately.

Proxima can:

- follow pilot activities and displayed information and adjust its support level in view of those activities and the displayed information;
- assess the mental and physical state of the human pilot through sensors and cameras to some degree;

— detect human pilot workload, incapacitation, and make correlation between the situation and the human pilot states to adapt its level of support; and

— monitor human communications and data link with the ground and aircraft position to ensure appropriate flight path management, and intervene where appropriate.

The high-level tasks performed by Proxima in interaction with the human end users can be supported by several types of scenarios. The objective of the scenarios is to create situations where the pilot will be busy flying manually. Such scenarios serve as a means to foster pilot's mental representation of the HAII with Proxima.

For the PIC, the high-level tasks are oriented by four main subparts: Fly, Navigate, Communicate, Management of systems, as proposed here:

— Proxima capable of performing automatic configuration of the aircraft including gear extension.

— Proxima in charge of the Navigation (FMS inputs)

— Proxima in charge of the Communication

— Proxima in charge of identification and management of failure

For the ATCO, the high-level tasks will be limited to 'communicate' and report to the PIC in case of doubt on the proper functioning of the AI-based system (based on its inputs).

2.3.1.2. Concept of Operations (ConOps)

Objective CO-04: The applicant should define and document the ConOps for the AI-based system, including the task allocation pattern between the end user(s) and the AI-based system. A focus should be put on the definition of the OD and on the capture of specific operational limitations and assumptions.

List of potential end users: See Objective CO-01 above.

List of high-level tasks for each end user: See Objective CO-02 above.

End user centric descriptions of the operational scenario(s):

For the purpose of the use case, the following detailed scenario is selected to exemplify the interaction between Proxima and the PIC:

<u>Aircraft SPO, automated flight control system failure and AP loss</u>

The scenario takes place in a flight from Paris Charles De Gaulle to Frankfurt. Pilot will be flying a commercial aircraft under SPO and acts as a PIC. The scenario starts in the air FL100 in descent. The pilot is performing the approach UNOKO 25N arrival followed by an ILS APPR to RWY 25R in navigating with the flight management system (FMS).

During the approach, the aircraft will experience a failure of the automated flight control system leading to a disengagement and loss of autopilot (AP) that requires constant input of PIC flying manually. The pilot will decide to continue the approach manually up to landing. Proxima will monitor aircraft parameters, the still functioning AP, as well as pilot state and any deviation. Depending on what Proxima detects, it will perform a number of actions such as interpreting information, initiating a conversation, acting on aircraft systems, communicating with the ATCO, etc.so fulfilling the high-level tasks.

How the end users could interact with the AI-based system:

User interface	HMI	Proxima		
		Reception	Output	AI capabilities
Speech interface	Speech input	Language recognition Speech recognition	Natural Language Procedural language	- Conversation - Questions/Answers - Argumentation / Negotiation - Follow-up questions - Corrections - Explanations - Acknowledgements
Gesture interface	Spatial hands gesture Head movements User behaviour (movement, posture)	Cameras Sensors	appropriate action	Gesture recognition combined with natural language understanding
Contact interface Galvanic Response	Keyboard CCD Touchscreens Skin contact with aircraft controls	Conventional hardware systems 'Sweat' rate – skin conductivity	Haptic information	Pilot state detection
Haptic	Control column, throttle leavers, switches, etc.	Monitoring of force, grip strength, speed etc. used when activating controls	Aural warning	Pilot state monitoring
Facial expression interface	Emotions Lips movements Pupil diameter Blink rate/duration	Cameras Eye-tracking	appropriate action	Pilot State detection Workload detection/fatigue
Neural computer interface	Brain activity signals Heart activity signals	Receptors	Control actuations	Workload detection
Aural interface	Aural	Voice comms – ground air, air ground	Voice comms air to ground	aircraft state intervention
Eye tracking	Gaze position – eye tracker	Eye fixation points	Colocation of displayed information – Synoptic screens	Interpretation of information requirements

Table 7 — Proxima user interface possibilities

2.3.1.3. Expected benefits and justification for Level 2B

The application is expected to reduce workload or help the pilot to confirm the correct understanding of a radio frequency in conditions of poor audio quality.

Objective CL-01: The applicant should classify the AI-based system, based on the levels presented in Table 2, with adequate justifications.

The Level 2B classification is justified by the fact that:

— Proxima and the PIC have a set of common goals and share tasks;

— Proxima is capable of using natural language, of taking decisions and implement actions that are overseen and overridable by the pilot; and

— Proxima is capable of having some authority in decision-making, to share situational awareness and react in real time to changes by adjusting strategies and task allocation.

3. Use cases — ATM/ANS

3.1. AI-based augmented 4D trajectory prediction — climb and descent rates

The objective of the use case is to improve the accuracy of a predicted 4D trajectory by better estimating the climb and descent rates with the use of deep learning techniques. To this purpose, a DNN is introduced to replace the software item in charge of the estimation of the climb and descent rates.

Note: The objectives referred to in this use case are traceable (in numbering and text) to the ones developed in the first issue of the EASA Concept Paper 'First usable guidance for Level 1 ML applications' from December 2021 (and may not match certain of the updated objectives in the present document).

3.1.1. Description of ConOps and systems involved in the use case

Objective CO-04: The applicant should define and document the ConOps for the AI-based system, including the task allocation pattern between the end user(s) and the AI-based system. A focus should be put on the definition of the OD and on the capture of specific operational limitations and assumptions.

3.1.1.1. Introduction

All information in this section has been derived from both the ATFCM Users Manual (EUROCONTROL, 2020) and the IFPS Users Manual (EUROCONTROL, 2020).

A 4D trajectory of a flight during pre-tactical phase, tactical phase, or when the flight is airborne is a fundamental element for correct network impact assessment and potential measures to be taken on congested airspace.

The 4D trajectory is (re)calculated in the context of many different services delivered by the Network Manager. Many different roles are interested in the 4D trajectory. Many different triggering events can generate the computation of a 4D trajectory.

Note: 4D trajectory and flight profile are to be considered as synonyms in this document.

3.1.1.2. Roles

Four different categories of end users with the following roles are involved in the operations of the 4D trajectory:

— (Aircraft operator (AO)) flight dispatcher;

— ATCO, with the area or en-route (ATC in this document) and the aerodrome or tower (TWR in this document);

— Flow management position (FMP); and

— Network Manager (NM) tactical team: The NM tactical team is under the leadership of the Deputy Operations Manager in charge of managing the air traffic flow and capacity management (ATFCM) daily plan during the day of operation. The tactical team is formed by the tactical Senior Network Operations Coordinator, the Network Operations Controllers, the Network Operations Officer and the Aircraft Operator Liaison Officer on duty.

3.1.1.3. 4D trajectory before flight departure

— Initial 4D trajectory based on flight plan (flight plan or filed flight plan (FPL))

A first version of the 4D trajectory is computed on the reception of a valid FPL by the AO.

The 4D trajectory is distributed to all ATCOs and TWR responsible for the ATC where the flight takes place.

— Reception of a change message (CHG)

When an individual FPL has been filed but it is decided, before departure, to use an alternative routing between the same aerodromes of departure and destination, the AO may decide to send a CHG for any modification.

Reception of a CHG triggers the re-calculation of the 4D trajectory and distribution to all ATCOs and TWRs responsible for the ATC where the flight takes place.

— Reception of a delay(ed) message (DLA)

On receipt of a DLA by the AO, the initial flight plan processing system (IFPS) shall re-calculate the 4D trajectory of that flight based on the revised estimated off-block time (EOBT).

— ATFCM solutions to capacity shortfalls

Where overloads are detected and the collaborative decision-making (CDM) process is initiated, different ATFCM solutions should be considered between the NM and the respective FMP(s).

This consists in:

- optimisation of the utilisation of available capacity, and/or utilisation of other available capacities (rerouting flows or flights, flight Level (FL) management) or advancing traffic; and/or
- regulation of the demand.

Most of the time, such ATFCM solutions will generate computation of 4D trajectories for the flights impacted.

3.1.1.4. 4D trajectory all along the life cycle of the flight

— Updating Central Airspace and Capacity Database (CACD) Data in Predict / Enhanced Tactical Flow Management System (ETFMS)

Updates to a subset of the environmental data (i.e. taxi time, runway in use for departures and arrivals, time to insert in the sequence (TIS), time to remove from the sequence (TRS), etc.) will trigger the re-computation of the flight profile of the aircraft concerned.

Taxi time updates and actual SID used by aircraft originating from A-CDM (from EOBT-3h up to target take-off time (TTOT)) are communicated to the ETFMS via departure planning information (DPI) messages for each individual aircraft.

The above parameters may be updated for each different (active) runway and the flight profiles are re-computed using this information.

— Airport CDM

Most advanced airports have engaged with NM in a CDM process aiming at improving the quality of information based on which decisions are made, then leading to enhanced operational efficiency and facilitating optimum use of available capacity.

Some of the DPI messages received by the ETFMS will have as a consequence the re-computation of the 4D trajectory for this specific flight (e.g. taxi time updates and actual SID used by aircraft originating from A-CDM (from EOBT-3h up to TTOT)).

— ETFMS flight data message (EFD) / publish/subscribe flight data (PSFD)

The EFD is basically an extract of flight data that is available in the ETFMS of which the flight profile is the most important part.

The EFD is sent by ETFMS to ANSPs of flight data processing areas (FDPAs) that request such information.

In the last years, EFDs have been complemented with PSFDs accessible via the NM B2B services.

3.1.1.5. 4D trajectory after departure

— Flight data information

On departure, the AO should send a departure message (DEP). Some AOs are sending aircraft (operator) position report (APR) messages to ETFMS. This data will then be used by the ETFMS to update the 4D trajectory in the current flight model (current tactical flight model (CTFM)) of the flight and also all other times (estimated times over (ETOs)) in the flight profile are updated accordingly.

Upon the flight's entry into the NM area, the flight's profile is then updated by first system activation (FSA) and correlated position report (CPR) messages where applicable.

For trans-Atlantic flights, flight notification message (FNM) from Gander and message from Shanwick (MFS) are messages that are received which provide an estimate for the oceanic exit point. MFS and FNM are processed first by integrated IFPS, that sends then the information to ETFMS. IFPS also sends it to AOs.

These estimates are used by the ETFMS to update the corresponding flight profiles.

— Correlated position reports (CPRs)

A flight may deviate from its last computed profile triggering a profile recalculation.

3.1.1.6. other usage of 4D trajectory

— Network simulations

The NM is responsible for the management of strategic ATFCM plans. Such plans rely on many simulations running in parallel and involve FMPs and AOs. Some simulation can imply the 4D trajectory calculations for flows under scrutiny.

— Post OPS analysis and reporting

The NM regularly reports on its activities and deliveries.

Among these post-operations activities, some reports elaborate on alternative 4D trajectories of the flown ones for further analysis in terms of flight efficiency (improved use of airspace, fuel consumption, etc.).

3.1.1.7. Measures

Considering a normal day of operations with:

- 30 000 flights;
- 5 000 000 CPR messages received;
- multiplicity of scenarios being launched in the context of ATFCM operations;
- new requests coming from A-CDM airports,

a rough estimation gives **300 000 000** of 4D trajectories computed every day.

3.1.2. Expected benefits and justification for Level 1

Objective CL-01: The applicant should classify the AI-based system, based on the levels presented in Table 2, with adequate justifications.

The ***AI Level 1A 'Human augmentation'*** classification is justified by only augmentation of the precision of the climb and descent phases, which participate to the computation of the 4D trajectory distributed to the roles involved with the flight profile. All decisions based on the predicted 4D trajectory are performed by a human or a machine with many indirections to the flight profile. It is then considered that this augmentation (***support to information analysis***) does not suggest any action or decision-making.

3.1.3. Trustworthiness analysis

3.1.3.1. Safety support assessment

Objective SA-04: The applicant should perform a safety support assessment for any change in the functional (sub)systems embedding a constituent developed using AI/ML techniques or incorporating learning algorithms, identifying and addressing specificities introduced by AI/ML usage.

The following describes the process that has been supporting the safety support assessment of the use case. The execution of the process takes into account the existence of a baseline safety support case (BSSC) for the NM services currently in the operations.

For reasons of conciseness, only the main outcomes of the process are presented in this document. For more information, please refer to Section 4.1 of the full report available by EUROCONTROL.

— Safety support assessment process

The safety support assessment of the change has been carried out in compliance with the requirements included in Regulation (EU) 2017/373 and its associated AMC and GM for service providers other than ATS providers.

The first step is the understanding and scoping of the change. It includes determination of the changed/new components of the NM functional system (FS), impacted (directly and indirectly) components of the NM FS, interfaces and interactions, and its operational context.

The second step of the safety support assessment used the failure mode and effect analysis (FMEA) technique to identify functional system failures. These failures can cause the services to behave in a non-specified manner, resulting in a different to the specified service output (e.g. lost, incorrect, delayed). Failure modes are linked (traceable) to the degraded mode(s) that can be caused by the failure. Where appropriate, internal (safety support requirements) and external mitigations (assumptions) have been derived to reduce or prevent undesired failure effects.

The third step of the safety support assessment, the degraded mode causal analysis, has been performed by means of facilitated structured brainstorming. It enabled the identification of the potential contribution of the changed and impacted elements of the NM FS to the occurrence of the degraded modes, as well as the establishment of safety support requirements to control the occurrence of the degraded modes and hence the service behaviour.

The fourth step will be the provision of the needed arguments and justification to demonstrate compliance with the safety support requirements.

— Safety support requirements

The table below contains the inventory of the safety support requirements, i.e. the necessary means and measures derived by the safety support assessment to ensure that NM operational services will behave as specified following the implementation of AI for the estimation of aircraft climb and descent rates. This table provides traceability to the mitigated service degraded modes and to the service performance.

No transition safety support requirements have been derived as the implementation of AI for the aircraft climb and descent rate estimation does not require a transition period.

ID	Safety support requirement	Mitigated degraded mode	Impacted service performance
R-01	Curtain shall implement alternative way of prediction calculation (e.g. based on fallback BADA table).	DGM06 DGM10 DGM11 DGM15 DGM17 DGM19	integrity availability
R-02	The AI/ML constituent shall return an error code in case it is able to detect an incorrect prediction.	DGM10	integrity
R-03	Curtain shall implement means to detect incorrect prediction provided by the AI/ML constituent.	DGM10	integrity
R-04	Curtain shall perform validation check of the AI prediction using a set of established criteria.	DGM10 DGM15 DGM19	integrity
R-05	Rules for use of alternative prediction computation by curtain shall be implemented.	DGM-10	integrity
R-06	Learning assurance shall be applied to the AI module to optimise the model generalisation.	DGM10	integrity
R-07	Carry out adequate tests of the AI module.	DGM10	integrity
R-08	Carry out focused TensorFlow tests.		
R-09	Measure the time to obtain a prediction and trigger alarm in case a defined threshold has been reached.	DGM06 DGM11 DGM17	availability
R-10	Design and execute dedicated test to refine the prediction validity threshold.	DGM10 DGM15 DGM19	integrity
R-11	Carry out load tests (at development and verification level).	DGM06 DGM11 DGM17	availability
R-12	Ensure resources (e.g. memory, disk space, CPU load) monitoring in operations.		
R-13	Comply with the SWAL4 requirement for IFPS/ETFMS.	DGM10 DGM15 DGM19	integrity

Table 8 — Safety support requirements

— Behaviour in the absence of failures

To ensure the completeness of the change argument, there is a need to analyse the behaviour of changed and impacted components of the NM FS in the absence of failures in order to prove that the NM services continue to behave as specified in the respective service specifications.

As a result of this analysis, the following safety support requirements have been placed on the changed and impacted by the change FS elements:

- **R-14**. The AI/ML constituent shall use industry-recognised technology (e.g. deep neural network) for training the prediction model. The use of TensorFlow shall be considered.
- **R-15**. The AI/ML constituent shall ensure correct generalisation capabilities which shall be verified by means of pre-operational evaluation with real flight plan data and, if necessary, improved.
- **R-16**. The AI/ML constituent shall expose an interface which shall be consumed by Curtain.
- **R-17**. The AI/ML constituent shall be able to process up to 100 requests per second. Curtain shall send a prediction request to the AI/ML constituent upon identification of the need to build a new or update an existing 4D trajectory.
- **R-18**. Curtain shall process the climb and descent rate predictions delivered by the AI/ML constituent.

— Assumptions

The table below contains the list of assumptions made during the safety support assessment that may apply and impact on the effectiveness and/or availability of the mitigation means and measures. It traces the assumptions and conditions to the associated degraded modes where they have been raised. The table also provides justification why the assumptions are correct and valid.

ID	Assumption/ Condition	Degraded Modes	Justification
A-01	Exhaustion of system resources will not only affect the AI module, but Curtain and other system processes, too.	DGM06 DGM11 DGM17	The AI module, Curtain and other critical system processes use the same computing resources (disk, memory and CPU).
A-02	By design, consecutive incorrect rate prediction for different flights cannot occur.	DGM10 DGM19	Successive incorrect rate predictions due to AI design issues will be identified during the software development and integration testing phase, and the AI predictive model will be enhanced consequently.
A-03	Failure of Curtain to compute an alternative prediction cannot occur for all flights.	DGM10 DGM19	This is a legacy function that has been proven in operation since years.

Table 9 — Use-case assumptions

— Safety support requirements satisfaction

This section will provide the needed assurance that the safety support requirements listed above are implemented as required in order to ensure that NM services (flight planning, ATFCM and centralised code assignment and management system (CCAMS)) will continue to behave only as specified in the respective service specifications.

3.1.3.2. Information security considerations

Objective IS-01: For each AI-based (sub)system and its data sets, the applicant should identify those information security risks with an impact on safety, identifying and addressing specific threats introduced by AI/ML usage.

The following describes the process that has been supporting the security assessment conducted on the use case.

For reasons of conciseness, only the main outcomes of the process are presented in this document. For more information, please refer to Section 4.2 of the full report available by EUROCONTROL.

— Approach to security assessment

The high-level security assessment is based on the following works:

- Microsoft:
  - AI/ML Pivots to the Security Development Lifecycle Bug Bar[24]
  - Threat Modeling AI/ML Systems and Dependencies[25]
  - Failure Modes in Machine Learning[26]
- A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View (Liu, 2018)
- MITRE Adversarial ML Threat Matrix[27].

The objective is to establish different potential attack paths and identify possible shortcomings.

As illustrated in Figure 27, we are considering the following security threats to the ML life cycle:

- Poisoning attacks: Those aim at corrupting the training data so as to contaminate the machine model generated in the training phase, aiming at altering predictions on new data.
- Evasion, impersonate & inversion attacks: Those aim at recovering the secret features used in the model through careful queries or other means.

[24] Source: https://docs.microsoft.com/en-us/security/engineering/bug-bar-aiml

[25] Source: https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml

[26] Source: https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

[27] Source: https://github.com/mitre/advmlthreatmatrix/blob/master/pages/adversarial-ml-threat-matrix.md. Latest commit: Oct 23, 2020.

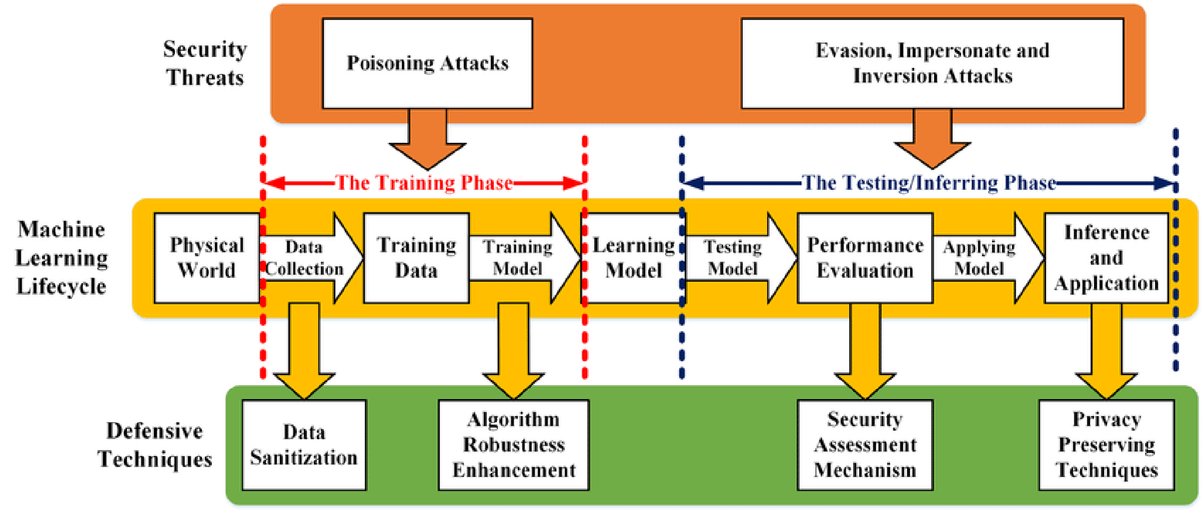


Figure 27 — Illustration of defensive techniques of ML

The 'Threat Modeling AI/ML Systems and Dependencies' questionnaires developed by Microsoft were used to capture the various aspects of the project and facilitate the security assessment. The 'Adversarial ML Threat Matrix' developed by MITRE was further used to focus the exercise on ML-specific techniques.

— System model for security assessment

Figure 28 is a simplified modelisation of the interaction between the different elements of the system. It represents the principal data exchanges taking place in the system.

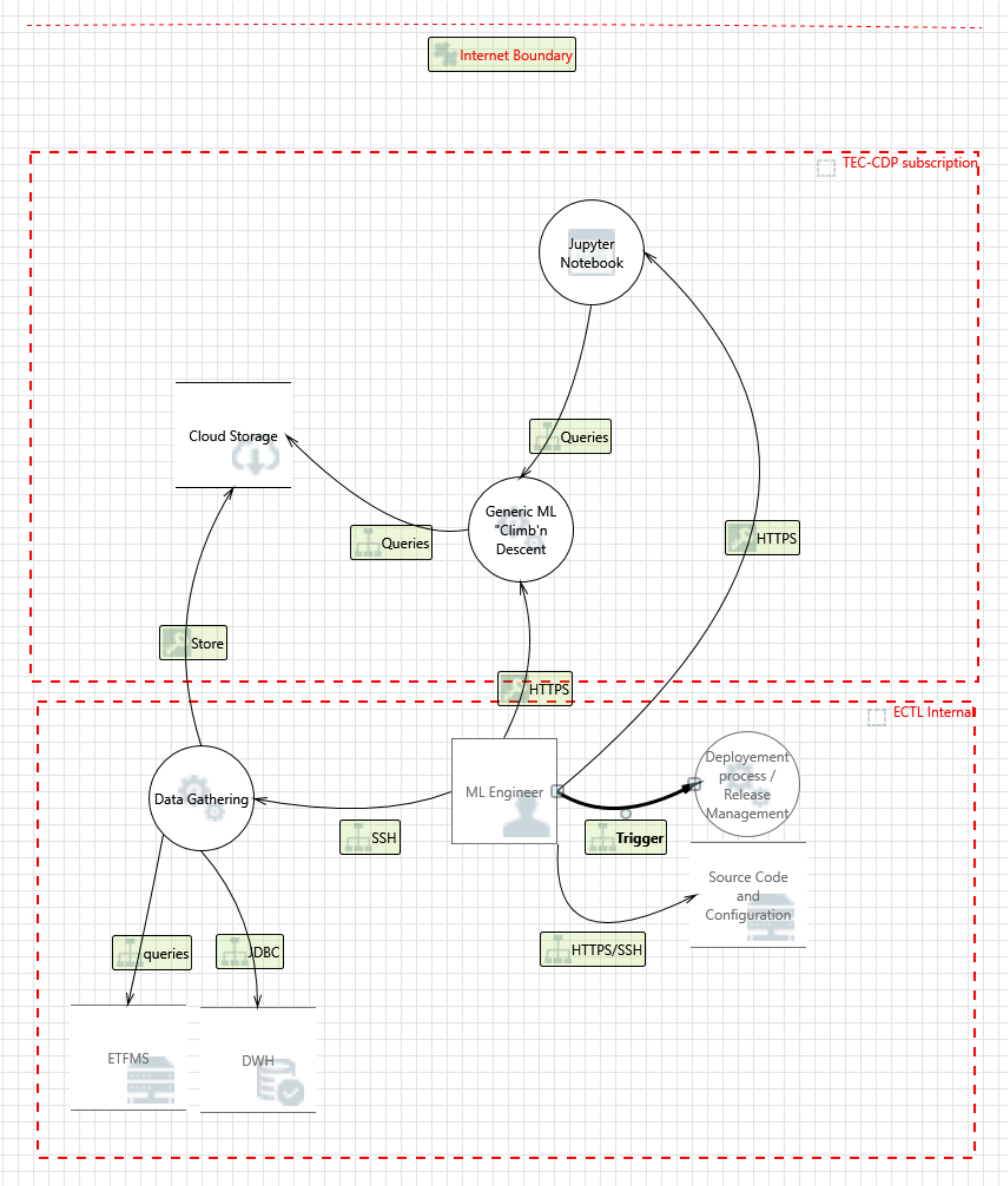


Figure 28 — Modelisation of the 'climb and descent' ML system

— Assumptions for security assessment

After an interview with the team in charge of the use case and considering the safety support case, the following considerations apply:

- The system considered is limited to the development phase of the model. The transfer to operations follows a dedicated workflow outside the scope.

- All data processed is post operations (no data confidentiality requirements, traffic light protocol (TLP):GREEN)
- The system is not considered as an operational system and does not present time-sensitive information.
- Safety support requirements and mitigations are in place, including the non-regression test.
- All involved communication networks are considered private with no interactive access to/from the internet.

Security and risks that are not inherent to the activities relating to the learning process are not considered in this assessment. Therefore, the applicable ratings for confidentiality, integrity and availability are:

- Confidentiality: Low
- Integrity: High
- Availability: Low

— Specific risks assessed

- Model poisoning: the threat was considered as mitigated by the assumptions: the isolation of the ML system vis-a-vis any external component whether from network or access permissions is considered sufficient mitigation.
- Data poisoning of training data: the threat was considered as mitigated by the assumptions: the isolation of the ML system vis-a-vis any external component whether from network or access permissions as well as the controlled source for all training data is considered sufficient mitigation.
- Model stealing: the threat was considered as mitigated by risk management: while there is no specific mitigation in place against the threat, it would not harm the organisation if it was to occur (no value loss).
- Denial of service on any component: the threat was considered as mitigated by the operational model: unavailability of the training data or ML environment has no operational impact and only results in limited financial costs.

Other risks have been considered during the analysis but are not considered pertinent in view of the operational model in place (for example, defacement, data exfiltration, model backdoor, etc.).

3.1.4. Learning assurance (in particular data management considerations)

Objective DA-01: The applicant should describe the proposed learning assurance process, taking into account each of the steps described in Sections C.3.1.2 to C.3.1.12, as well as the interface and compatibility with development assurance processes.

Most of the activities expected to be performed as per the 'learning assurance' have been executed. The following will make the demonstration of this statement.

3.1.4.1. Data preparation

a. Data collection

Objective DM-03: The applicant should identify data sources and collect data in accordance with the defined ODD, while ensuring satisfaction of the defined DQRs, in order to drive the selection of the training, validation and test data sets.

— Data sources

Almost 3 years of data (from 01/01/2018 until 30/09/2020) was extracted from the NM Ops data warehouse from the ARU[28] schema. This contains basically all flights in the NM area for the last 3 years, and these were taken into the data set.

Weather information was taken from the UK Met office Sadis source, stored in the operational FTP server under the Met directory. EUROCONTROL has had a long-standing contract with the UK Met office to provide this data.

Objective DM-04: Once data sources are collected, the applicant should ensure the high quality of the annotated or labelled data in the data set.

— Data labelling

The data labels[29] are also extracted from the ARU data set.

— Rates of climb and descent by performance slice

In a first step, the rate of climb between consecutive points of the point profile was calculated.

For a given flight phase, the time T for which a flight arrives at the flight level F, if there is no point at this flight level in the profile, can be approximated by linear interpolation:

$$T = T_{prev} + \frac{T_{next} - T_{prev}}{F_{next} - F_{prev}}(F - F_{prev})$$

[28] Due to the mainframe phase-out, this system was converted to Unix under the heading of the ARU System (Archive System on Unix). Once most functions were migrated to Unix, the system was renamed to Data Warehouse System (DWH).

[29] Data labelling is a key part of data preparation for machine learning because it specifies which parts of the data the model will learn from.

where *prev* and *next* stand for the point of the profile respectively before and after the flight level.

If there is a point at the requested flight level, we simply use its time over.

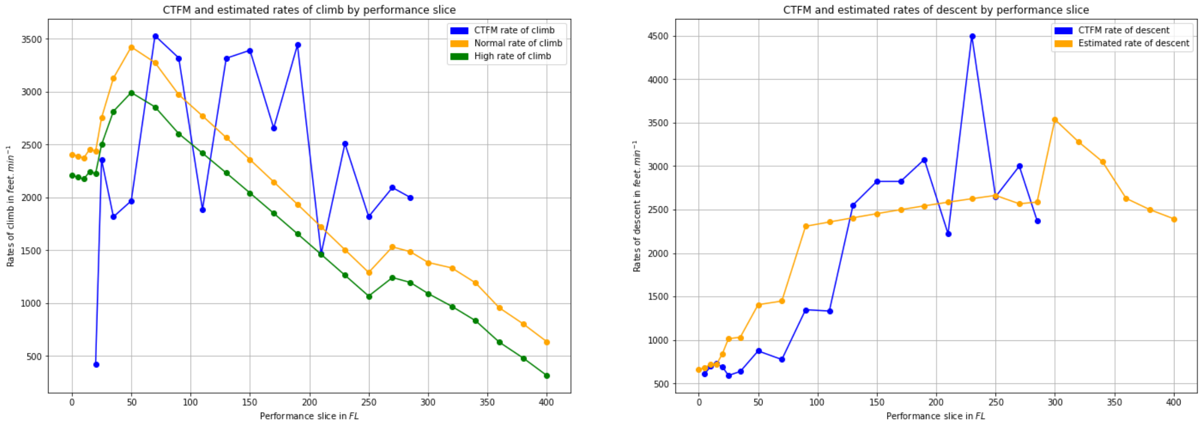


Figure 29 — Climb and descent rates by performance slice

— Managing aleatory uncertainty — removing high-frequency noise on CPR data

It was observed that the calculated climb rates appear to have a lot of high-frequency noise overlaid on the signal and so we removed it by applying a low-pass filter to that in the form of a simple moving average window function of width 5.

b. Data pre-processing

Objective DM-06: The applicant should define and document pre-processing operations on the collected data in preparation of the training.

— Data cleaning

Several data cleaning operations were performed, including the removal of yo-yo flights[30] (polluting the quality of the model), and the removal of the data corresponding to the cruise phase of the flight.

— Outliers

All data samples with climb rates that were calculated to be greater than 1 000 ft/min (likely to be not physically realistic and related to inaccuracy in the radar plots) were removed from the data set. Around 0.1 % of the 400 million samples were removed during this operation.

Objective DM-08: The applicant should ensure that the data is effective for the stability of the model and the convergence of the learning process.

— Data normalisation

All data was normalised by centring on zero by subtracting the mean and given similar ranges by dividing by the standard deviation of that feature.

30 Yo-yo flight: flight with multiple climb and descent phases in the flight profile.

c. Feature engineering

Objective DM-07: When applicable, the applicant should define and document the transformations to the pre-processed data from the specified input space into features which are effective for the performance of the selected learning algorithm.

Feature engineering was managed via a pipeline. The pipeline's purpose is to enrich the data with various calculated features required for the subsequent operation.

Firstly, the SID and STAR are extracted from the flown route and attached to separate fields to the flight information so that they can be used as independent features.

The representations of coordinates in the database was string format rather than decimal format and these were converted into decimal degrees.

Several operations were made on the weather forecast data source. For more information, please refer to the full report available by EUROCONTROL.

Several additional calculated weather-forecast-related features were then produced, namely wind speed and wind direction relative to the aircraft.

Some further features were then added. It was discovered that using the latitude and longitude of the aerodrome of departure and destination as well as the first and last point of the climb and descent was more effective than any other encoding of these values. For example, an embedding layer was used to encode the categorical values e.g. the ICAO names for aerodromes of departure and destination, but this was not nearly as effective as the vector encoding as latitude and longitude.

This resulted in a model with some 40 features which was saved in a parquet file which when loaded was around 100 gigabytes in RAM.

The permutation importance (a similar method is described in Breiman, 'Random Forests', Machine Learning, 45(1), 5-32, 2001) for these features was then calculated. This was a very heavy calculation taking several days on a GPU to complete.

Permutation importance:

Climb		Descent	
Weight	**Feature**	**Weight**	**Feature**
494468.7164 ± 269.1501	PERF_CAT_LOWER_FL	392129.5391 ± 248.7002	PERF_CAT_LOWER_FL
217568.8688 ± 138.2701	FTFM_CLIMB_RATE	211356.3405 ± 95.6282	FTFM_DESC_RATE
138494.9605 ± 44.0213	FTFM_MAX_FL	133131.4453 ± 68.8156	FTFM_DESC_FIRST_PT_LAT
114020.7645 ± 86.3738	FLT_DEP_AD	85637.1216 ± 64.1071	FTFM_DESC_LAST_PT_PT_LAT
109271.3590 ± 243.7783	FLT_DEP_AD_LAT	85262.9041 ± 138.5218	FLT_FTFM_ADES_LAT
105701.0231 ± 96.9098	FTFM_CLIMB_FIRST_PT_LAT	80916.0368 ± 71.9405	FLT_FTFM_ADES
95154.7142 ± 86.0832	ICAO_ACFT_TY_ID	72740.5408 ± 34.9251	FTFM_DESC_FIRST_PT_LNG
86846.6291 ± 88.8068	FTFM_CLIMB_FIRST_PT_LNG	70372.2655 ± 109.2796	FTFM_DESC_LAST_PT_LNG
86710.6489 ± 193.9731	FLT_DEP_AD_LNG	69247.5777 ± 83.0451	FLT_FTFM_ADES_LNG

Climb		Descent	
Weight	Feature	Weight	Feature
23296.1818 ± 26.1849	FTFM_CLIMB_DURATION	43342.9997 ± 56.8700	FTFM_MAX_FL
21731.4291 ± 59.1714	AO_ICAO_ID	37916.0572 ± 130.2117	FTFM_DESC_DURATION
20337.5237 ± 73.7881	FTFM_CLIMB_FIRST_PT	32727.9660 ± 55.2942	FTFM_DESC_LAST_PT
18971.2889 ± 22.4656	FLT_FTFM_ADES_LAT	12746.5049 ± 19.2558	ETA_DAYOFYEAR
18136.2638 ± 26.9874	FLT_FTFM_ADES_LNG	11355.1165 ± 65.0552	AIRAC_CYCL
18026.4043 ± 22.2186	FTFM_DESC_LAST_PT_PT_LAT	9524.1099 ± 37.4795	ICAO_ACFT_TY_ID
16417.4972 ± 20.0458	FTFM_DESC_LAST_PT_LNG	6437.3164 ± 30.2539	AO_ICAO_ID
15343.8757 ± 44.8245	ETA_DAYOFYEAR	5731.4322 ± 19.5940	FLT_REG_MARKING
15176.5899 ± 32.8208	FLT_REG_MARKING	5658.8823 ± 21.7385	FTFM_CLIMB_FIRST_PT_LAT
15034.2075 ± 24.5128	FTFM_CLIMB_LAST_PT_LNG	5400.5508 ± 40.4232	FTFM_CLIMB_LAST_PT_LNG
14964.0634 ± 29.0470	FTFM_CLIMB_LAST_PT_LAT	5119.9972 ± 15.9033	FTFM_CLIMB_FIRST_PT_LNG

Table 10 — Extract of candidate features by importance (20 out of 40)

When the permutation importance of a feature is low, this means the feature is not very decisive for obtaining a result.

d. Hosting for data preparation and model training

Data preparation was hosted under Microsoft Azure. The model training was hosted in a Cloudera Machine Learning (CML) environment. This is Cloudera's cloud-native ML service, built for CDP. The CML service provisions clusters, also known as *ML workspaces,* that run natively on Kubernetes.

ML workspaces support fully-containerised execution of Python, R, Scala, and Spark workloads through flexible and extensible *engines*.

This facility allows automating analytics workloads with a job and pipeline scheduling system that supports real-time monitoring, job history, and email alerts.

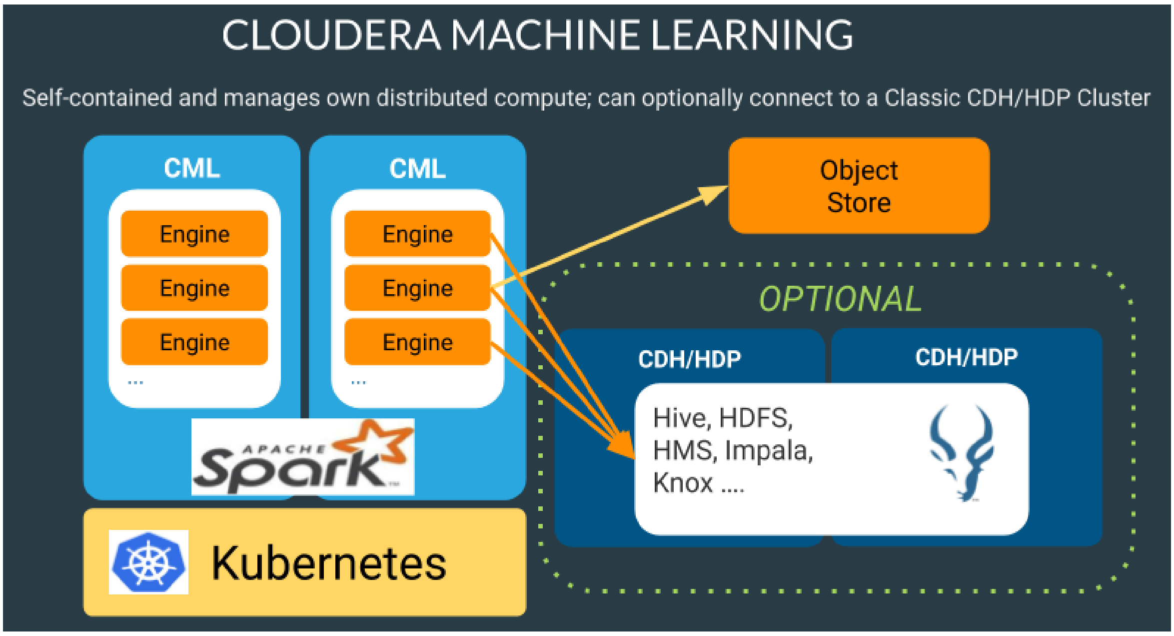


Figure 30 — Cloudera machine learning environment

For more information, please refer to the full report available by EUROCONTROL, or contact the teams at EUROCONTROL in charge of such an environment.

3.1.4.2. Data validation

a. Data completeness

Objective DM-10: The applicant should ensure validation and verification of the data, as appropriate, all along the data management process so that the data management requirements (including the DQRs) are addressed.

The period which has been considered for the data in the data set (3 years of archived data from the DWH), and the inherent quality of the DWH via its usage by thousands of stakeholders on a daily basis, ensure the completeness of the data for the use case.

b. Data accuracy

Data accuracy has been established through the different activities performed during the data management phase. In particular, incorrect or non-representative data has been removed from the data set during data cleaning (e.g. removal of yo-yo flights), or when identifying outliers (flights with unrealistic climb or descent rates).

c. Data traceability

All operations performed on the source data set extracted from the DWH were orchestrated via scripting and pipelining in different python modules. All code is under configuration management, ensuring full traceability and capability to reproduce featured input and labelled data for subsequent training.

d. Data representativeness

The 4D trajectory applies to the ECAC area. The DWH archives all information which has been processed by IFPS/ETFMS, then ensuring that the data set fully covers this geographical area.

e. Data allocation — data independence

Objective DM-09: The applicant should distribute the data into three separate and independent data sets which will meet the specified DQRs:

— the training data set and validation data set, used during the model training;

— the test data set used during the learning process verification, and the inference model verification.

There are roughly 370 million data samples in the data set. The test set was chosen at random and had 5 % set-aside.

The validation set was a further 20 % of the remaining.

Considering the large amount of data samples, keeping 5 % of all data for the test set represents 25 million samples in the test data set, which is enough to provide a statistically valid result. The same remark applies to the validation data set.

3.1.4.3. Learning process management

Objective LM-01: The applicant should describe the AI/ML constituents and the model architecture.

Objective LM-02: The applicant should capture the requirements pertaining to the learning management and training processes.

a. Model selection

A DNN was selected.

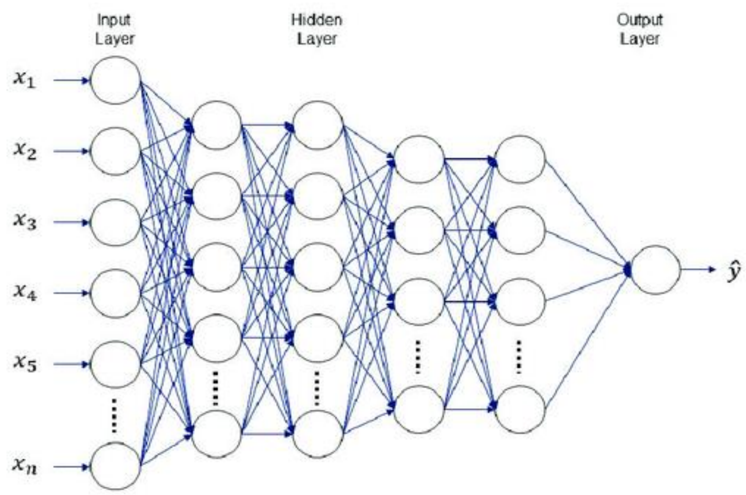


Figure 31 — DNN structure

Multiple architectures were tested during hyper-parameter tuning. The most successful architecture for the hidden layers was as follows.

Layer number	number of neurons
1	512
2	512
3	256
4	256
5	128
6	64

Table 11 — Internal architecture of the DNN

The table below summarises the main decisions/configurations made/applied at the end of the training process:

Title	Information / Justification
Activation function	The **PReLU** activation function was chosen for a number of its advantages in DNNs; particularly, avoidance of the vanishing gradients problem as was the case with standard ReLU, but in addition the avoidance of the dying neuron problem.
Loss function selection	Several loss function strategies were studied during the learning and training process. Finally, it was decided to use **'mean absolute error'** which appears to give the best results on the test set.
Initialisation strategy	The **Glorot** initialisation technique was chosen for initialising the values of the weights before training.
Hyper-parameter tuning	Hyper-parameter tuning was a recurrent activity all along the learning process management and the model training.

Table 12 — Key elements of the DNN

b. Hosting the model predictor

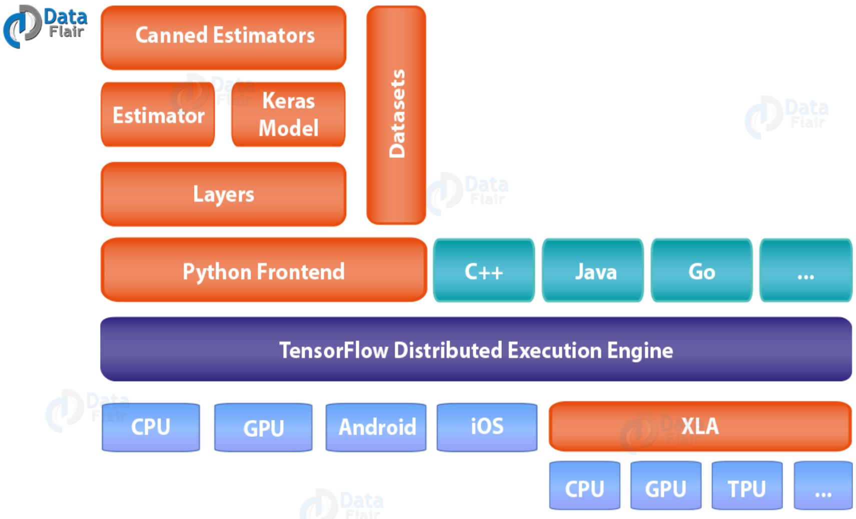


Figure 32 — Tensorflow component model and dependencies

The above diagram represents the TensorFlow component model and dependencies. The predictive models were developed using Keras Python interfaces to TensorFlow — see above on the left side.

The model training pipeline based on Python and Keras produces a saved model in protobuf format and associated model weights files. This is done in the cloud as described above.

3.1.4.4. Model training

a. Feature set

The following table represents the current list of features which were used for the training:

Feature	Feature
AO_ICAO_ID	float32
ETA_DAYOFYEAR	float32
FLT_DEP_AD_LAT	float32
FLT_DEP_AD_LNG	float32
FLT_FTFM_ADES_LAT	float32
FLT_FTFM_ADES_LNG	float32
FLT_REG_MARKING	float32
FTFM_CLIMB_RATE	float32
ICAO_ACFT_TY_ID	float32
PERF_CAT_LOWER_FL	float32

Table 13 — List of features as an input to model training

Objective LM-05: The applicant should document the result of the model training.

b. Learning curves

The figure below depicts a learning curve when using the feature set and the labelled data:

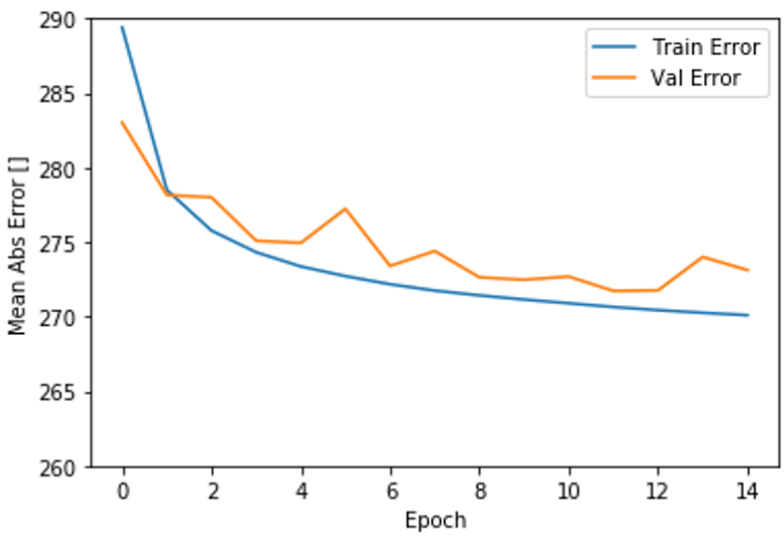


Figure 33 — Model training (mean absolute error)

3.1.4.5. Learning process verification

Objective LM-09: The applicant should perform an evaluation of the performance of the trained model based on the test data set and document the result of the model verification.

a. 3D histogram plot of the predicted values

The below figures show two plots for all of the test data of climb rate predictions against actually observed climb rates.

The dispersion is greatly reduced with the trained ML model.

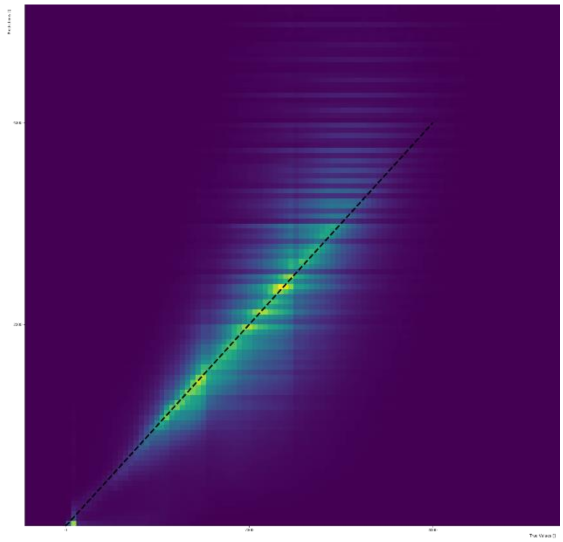


Figure 34 — Predicted climb rate (with BADA) v actual from CTFM

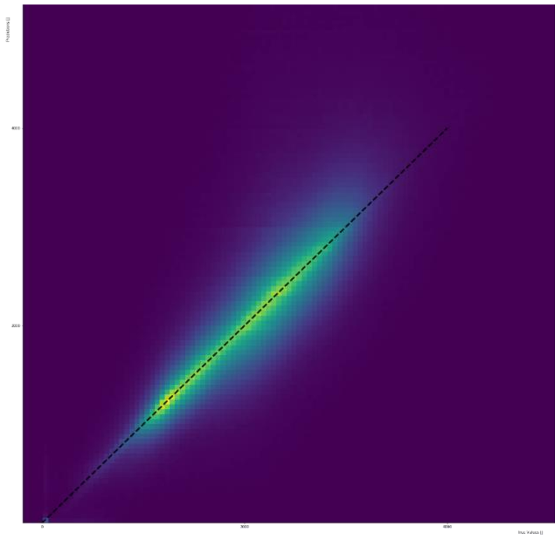


Figure 35 — Predicted climb rate (with ML) v actual from CTFM

b. Comparison of error rates between current (FTFM) and new ML calculation

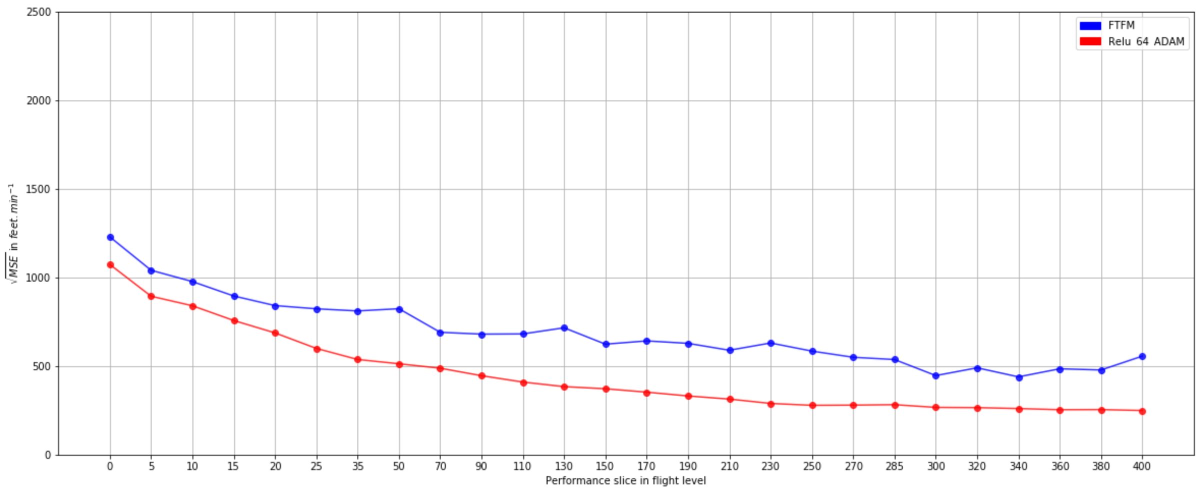


Figure 36 — Mean square error on actual climb rates (with low-pass filter)

3.1.4.6. Implementation

Objective IMP-03: For each transformation step, the environment (i.e. software tools and hardware) necessary to perform model transformation should be identified and any associated assumptions or limitations captured and validated.

a. System architecture

Depending on the context where the 4D trajectory calculation is performed, the AI/ML library could be called from different processes. The following is the logical architecture of ETFMS. The 4D trajectory is calculated within the 'profiler process':

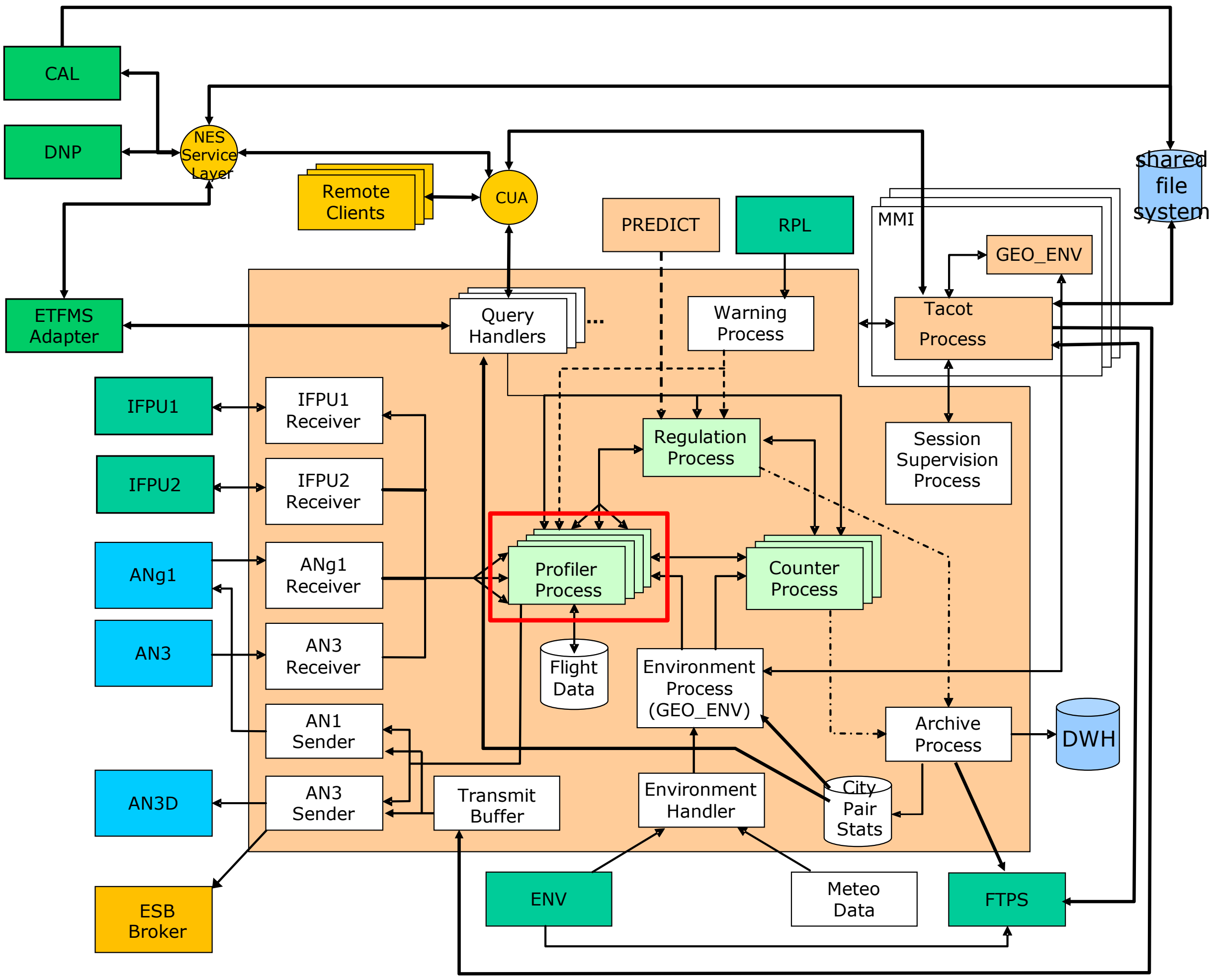


Figure 37 — ETFMS logical architecture

The 'profiler process' computes the flight profile or 4D trajectory. For performance reasons, several processes can co-exist in ETFMS. An algorithm statically associates a flight with a 'profiler process' to allow parallelism.

The 'profiler process' is mission-critical. Its failure induces an ETFMS failure.

The flight load is distributed equally by a hashing algorithm amongst the number of 'profiler processes'. Once a flight has been associated with a given instance of a process, for the sake of data consistency, this instance is the only one that manages the flight; all messages relating to the flight are directed to it.

The 'profiler process' embeds the Curtain software package.

The Curtain software package has been adapted to use the AI/ML constituent.

b. AI/ML constituent as a library

— General information

A prediction is a numerical value provided by a TensorFlow model. The inputs are an ordered list of fields and, usually, after transformation and normalisation, are passed to the model which returns a value, the prediction. The library should be supported with additional information: the TensorFlow model resulting from training, the statistics from the training data (mainly mean and standard deviation) used by the normalisation, and the conversion from categorical value to numerical value used to include categories in the prediction. The library is also configured with a description of the fields, categories, eventual ways to validate the input and output, and, in the case of invalid input, how to replace them by acceptable values.

A prediction is provided by a predictor. The API lets the user create and register one or more predictors with a given name. It is possible to remove an existing predictor but also to swap two predictors (they exchanged their names) as a shortcut to remove and re-create. Creation implies moving in memory several lookup tables, so swapping can improve performance in some cases.

Each predictor is linked to one or more TensorFlow models, provided as TensorFlow .pb and checkpoint files.

As a lot is triggered by configuration, there is a function in the API to print the global configuration (input data and pre-computed lookup tables) from a predictor. Another function will try to analyse the predictor in order to see if it is consistent (at least one model, at least one field, etc.).

The API is a C API and will provide different functions, structures to represent input data and enumerations for code values.

— Workflow

The library implemented a workflow which is generic and can be reused for different AI/ML use cases.

The figure below depicts the workflow for prediction which was implemented:

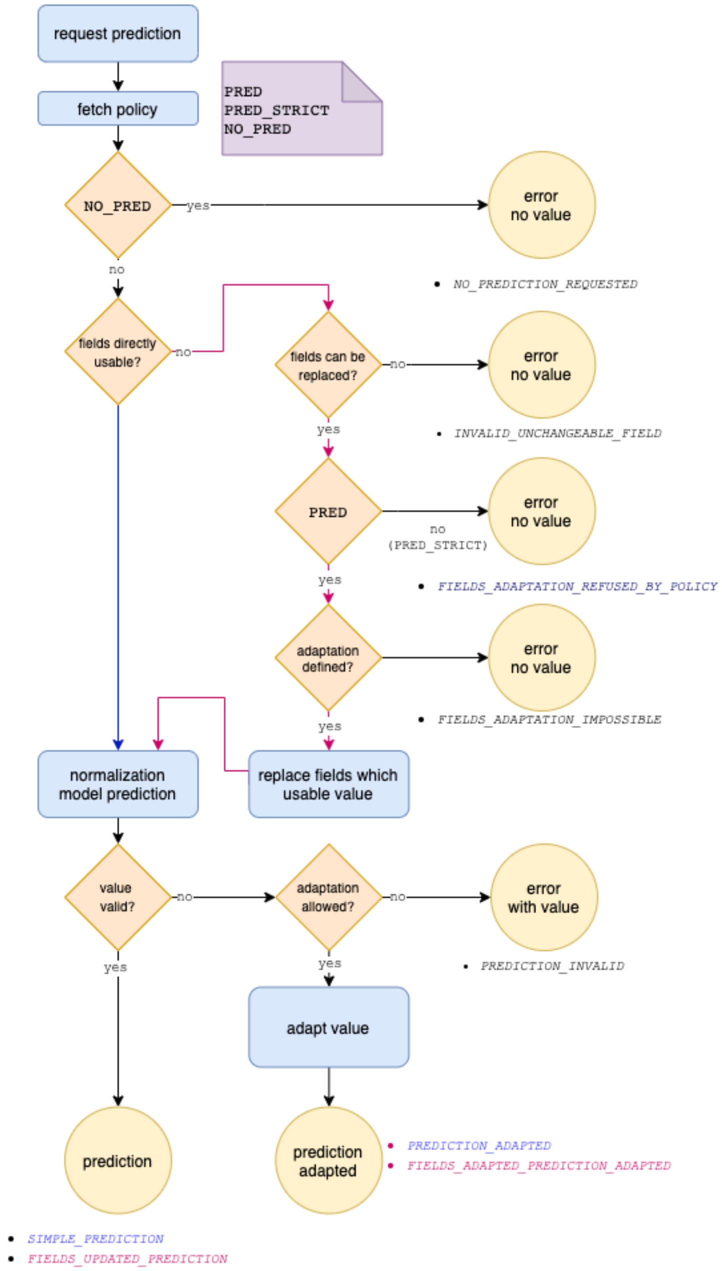


Figure 38 — Workflow for prediction delivered by the AI/ML library

The saved models were used in the ETFMS operational context via the C/C++ API.

This library was delivered to the ETFMS, and an Ada binding is produced so that the predictions could be provided by a simple in-process call in the same address space.

The reason for this is the need for very low latency and high bandwidth to ML predictions as the trajectory calculations in ETFMS are particularly performance-sensitive. It is not feasible or desirable to use a traditional technique of providing an ML REST-based server to provide the predictions as the latency of the network connection would make the predictions useless in this context.

c. Executable model architecture

The following depicts part of the NM operational infrastructure running the ETFMS.

ETFMS is located on-premise. It is part of a mission-critical cluster (called RED). ETFMS operational instance (ptacop1 or ptacop3) is part of the sub cluster RED_03, which contains four virtual machines (red011 to red014).

These virtual machines are based on **Linux Red Hat Enterprise Server** Operating System.

All the virtual machines of the RED Cluster are spread over the 6 **HP DL560 G10** Servers (rambo, rocky, rufus, romeo, rusty, roger), all based on 36 CPS and 1,5 TB of RAM.

The hypervisor used to manage the virtual machines is **VMWare ESXi**.

Storage for this cluster is spread over NASes (rednas10 to 13).

3.1.4.7. Inference model verification

Objective IMP-06: The applicant should perform an evaluation of the performance of the inference model based on the test data set and document the result of the model verification.

a. Verification of improvements at network level

The most appropriate way to assess the performance of the AI/ML constituent was to analyse the impact on the network situation. This analysis is possible based on some tools capable of replaying specific situations which have occurred in the past (also known as PREQUAL). For more information, please refer to the full report available by EUROCONTROL.

The table below demonstrates significant improvements on the network for two separate dates in 2020:

	10/08/2020		**14/02/2020**		**18/12/2020**	
	Avg	Max	Avg	Max	Avg	Max
BL	283 051	1 860 046	544 428	3 226 165	285 194	1 783 651
ML	265 889	1 655 747	514225	2 880 420	272 071	1 632 486
Improv.	6,06 %	10,98 %	5,54 %	10,71 %	4,60 %	8,48 %

Table 14 — Improvements on the network

Objective IMP-07: The applicant should perform a requirements-based verification of the inference model behaviour and document the coverage of the ML constituent requirements by verification methods.

In addition to verification of the improvement brought at network level, verification activities have taken place from various perspectives, including system resilience.

b. Robustness

Objective IMP-08: The applicant should perform and document the verification of the robustness of the inference model.

At the date of this report, the robustness of the AI/ML constituent remains to be investigated. It will be progressively assessed via additional testing at the limits (e.g. how will the model perform when being faced to abnormal data like an unknown airport or unknown aircraft type).

c. Resilience

Based on the system requirements identified for Curtain, and the target architecture, should the model face robustness limitations, then the legacy climb and descent computation would continue to deliver the service even in a less performant mode of operation. All these measures ensure resilience at system level.

3.2. Time-based separation (TBS) and optimised runway delivery (ORD) solutions

3.2.1. Trustworthiness analysis — description of the system and ConOps

3.2.1.1. Description of the system

Objective CO-03: The applicant should determine the AI-based system taking into account domain-specific definitions of 'system'.

Headwind conditions on final approach cause a reduction of the aircraft ground speed which for distance-based separation results in increased time separation for each aircraft pair, a reduction of the landing rate, and a lack of stability of the runway throughput during arrival operations. This has a negative impact not only on the achieved capacity, but also on the predictability of operations, time and fuel efficiency, and environment (emissions). The impact on predictability for core hubs is particularly important at the network level. The service disruption caused by the reduction in achieved runway throughput compared to declared capacity in medium and strong headwinds on final approach has a significant impact on the overall network performance. It is also particularly exacerbated if this occurs on the first rotation of the day because of the impact on all the other rotations throughout the day.

Time-based separation (TBS) on final approach is an operational solution, which uses time instead of distance to safely separate aircraft on their final approach to a runway.

In order to apply this concept, approach and tower ATCOs need to be supported by a separation delivery tool which:

— provides a distance indicator (final target distance (FTD)), enabling to visualise, on the surveillance display, the distance corresponding to the applicable TBS minima, and taking into account the prevailing wind conditions;

— integrates all applicable separation minima and spacing needs.

This separation delivery tool, providing separation indicators between arrival pairs on final approach, also enables an increase in separation performance when providing a second indicator (initial target distance (ITD)): a spacing indicator to optimise the compression buffers ensuring optimum runway delivery (ORD). Both indicators are shown in Figure 39.

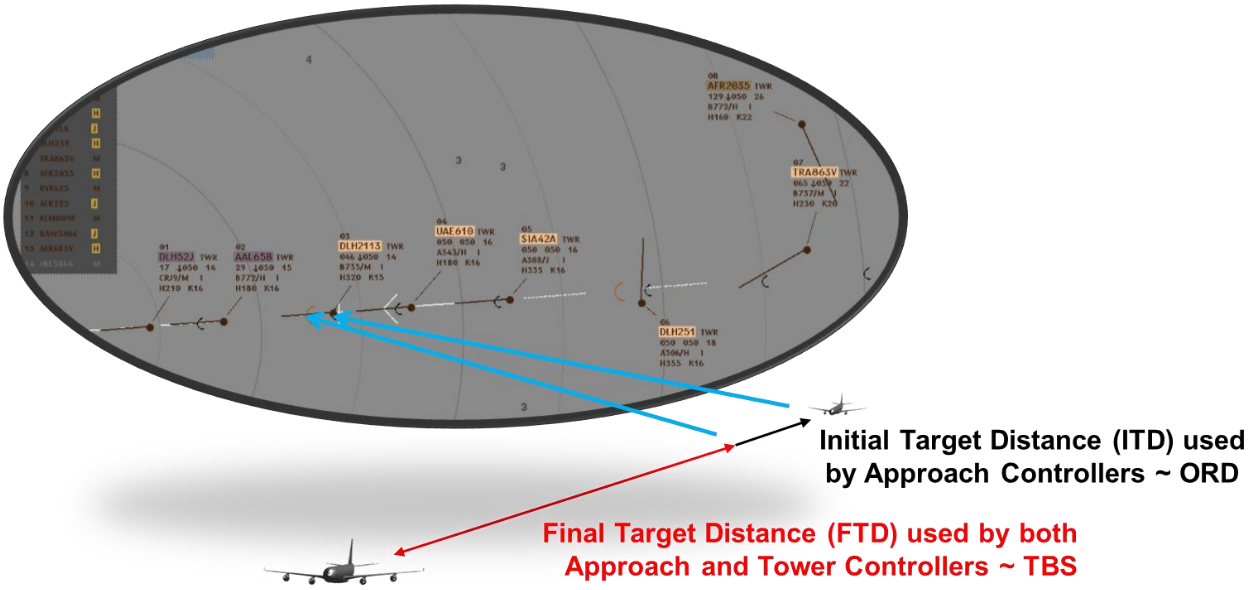


Figure 39 — Representation of FTD and ITD in the ATCO's separation delivery tool

The move from distance (DBS)- to time (TBS)-based rules allows efficient and safe separation management requests to properly model/predict aircraft ground speed and behaviour in short final approach and the associated uncertainty. A too conservative definition of buffer in the indicator calculation can lead to a reduction of efficiency, whereas making use of advanced ML techniques for flight behaviour prediction allows improvements of separation delivery compared to today while maintaining or even reducing the associated ATCO workload.

The Calibration of Optimised Approach Spacing Tool (COAST) is a EUROCONTROL service to ANSPs for safely optimising the calculation of TBS-ORD target distance indicators through the training and validation of ML models and a methodology to use them. A description of COAST can be found in https://www.eurocontrol.int/publication/eurocontrol-coast-calibration-optimised-approach-spacing-tool-use-machine-learning.

Those models can then be integrated in the indicator calculation modules of a TBS-ORD ATC separation tool. The inference model functional architecture in such a tool is depicted in Figure 40.

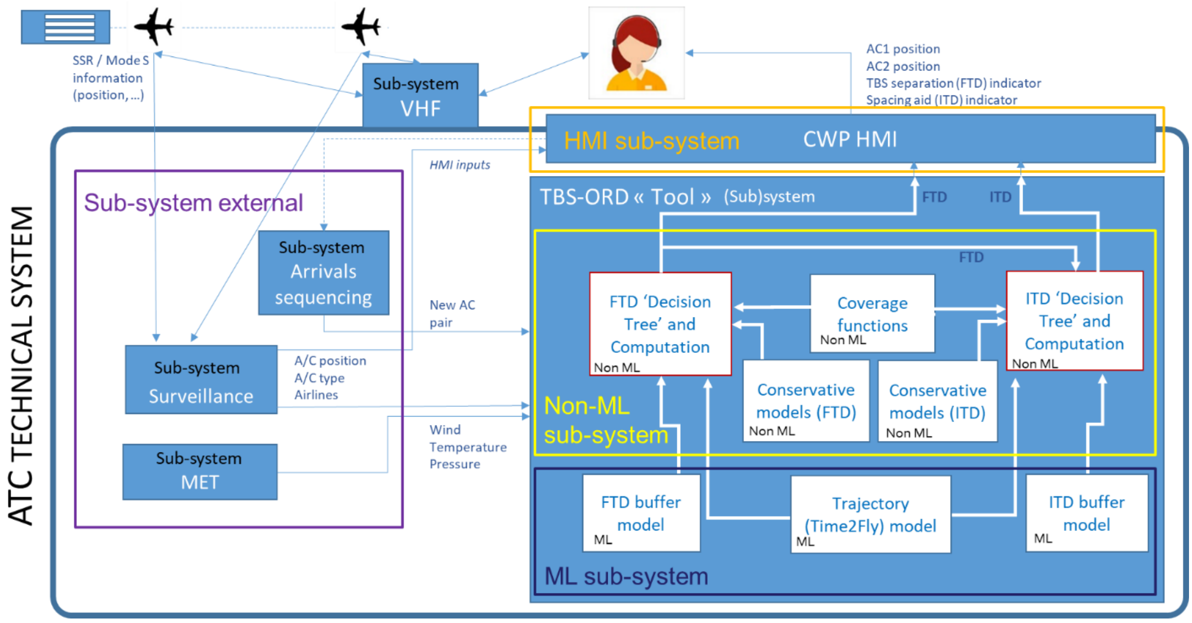


Figure 40 — TBS-ORD system functional architecture

To build the different models, EUROCONTROL has developed a pipeline that creates the different components defined in the architecture based on each airport's historical data. The execution of the ML toolbox pipeline can be summarised as follows: a set of flights and couples are read from the database. The data set is split in three data subsets, called training, validation and test data sets, which are given as an input to the ML toolbox. The toolbox computes ML time-to-fly, FTD and ITD buffer models using the training set and calculates the coverage (defining when ML models can be used) on the validation set. A conservative time-to-fly model is also developed and calibrated based on the training set whereas conservative FTD and ITD buffer models are calibrated based on the validation set. The whole process is then assessed on the independent test data set. For every execution of the pipeline, ad hoc model performance reports are generated. They include information about:

— time-to-fly estimation quality;

— buffers estimation accuracies;

— coverage functions explanation; and

— FTD/ITD and resulting time separations.

The toolbox requires the following historical data as inputs:

— **flight data**: aircraft type, category (RECAT), airline, runway, landing time, time-to-fly profile and optionally *origin airport*

— **weather data**: surface head-, cross- and total wind, vertical profile of head- and cross-wind, temperature and pressure on the glide

— **constraints information**: wake distance- and time-based separation, runway occupancy time (ROT) spacing and surveillance (MRS) minima

The models can also be trained without having access to some optional features (e.g. origin airport, pressure).

Design criteria are defined for FTD and ITD:

— for FTD, to ensure compliance with applicable wake, surveillance and ROT separation/spacing constraints;

— for ITD, to prevent FTD infringement after leader deceleration to stabilised approach speed.

The ML toolbox creates a set of ML models and functions:

— Predictive time-to-fly models

— Predictive buffer models

— A set of coverage functions and the associated decision trees

— A set of conservative models:

  - conservative models for time-to-fly
  - conservative models for buffers

3.2.1.2. Concept of operations

Objective CO-04: The applicant should define and document the ConOps for the AI-based system, including the task allocation pattern between the end user(s) and the AI-based system. A focus should be put on the definition of the OD and on the capture of specific operational limitations and assumptions.

Operational design domain (ODD)

The TBS-ORD separation delivery tool supports and is used by the approach and tower controllers in delivering the required separation or spacing on approach to the runway landing threshold. It calculates the indicators and displays target distances on the approach and tower controller working positions (CWPs).

The target distances indicators include:

— **Final target distance (FTD):** a separation indicator which displays the required separation / spacing to be delivered to the required delivery point (DP). The separation indicator corresponds to the minimum distance separation to be applied between the leader and the follower when the leader is overflying the separation DP.

— **Initial target distance (ITD):** a spacing indicator which displays the required separation when the leader aircraft is at a prescribed glide speed (e.g. 160 kt) before deceleration to final approach speed such that the FTD will be obtained at the separation DP.

Figure 41 shows a representation of the FTD and ITD indicators during final approach phase.

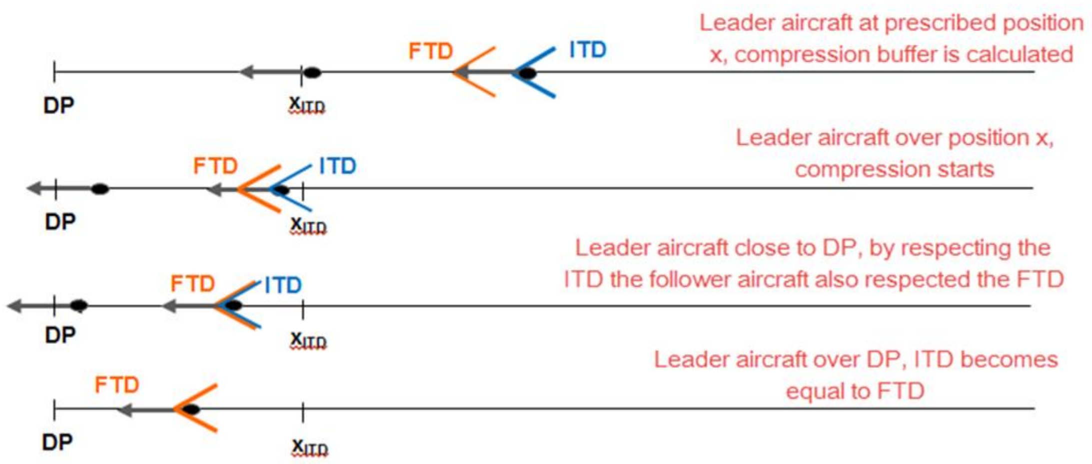


Figure 41 — Evolution of FTD and ITD indicators during final approach phase

Within COAST, a set of models are developed using advanced big data and ML techniques to predict aircraft performance (in terms of trajectory / time to fly) on the final approach as well as the safety buffers required to account for uncertainties relating to aircraft performance and wind. The COAST output consists of a number of different models that are used by a TBS-ORD separation delivery tool to calculate the FTD and ITD per aircraft pair. These ML models are used instead of using more traditional analytical techniques.

The models calibrated through COAST consist of:

— **Trajectory/time-to-fly ML model** — predicting the trajectory/time-to-fly profile of a flight on the final approach.

— **FDT and ITD buffer ML models** — predicting the buffer to be added in the FTD and ITD calculation. The time-to-fly model indeed predicts an average expected profile for an aircraft. Its predictions allow the computation of expected separation (FTD) and compression (ITD) indicators. However, uncertainties exist on the time-to-fly predicted profiles of both the leader and follower and also due to uncertainty on the aircraft airspeed profiles and on the wind conditions.

— **Time-to-fly, FTD buffer and ITD buffer conservative models** — For certain flight conditions, where there is little data and the confidence in the quality of the ML models is limited, conservative models are developed. When a flight or an aircraft pair is not considered covered (identified by an independent performance evaluation of the predictability of the models using an independent test data set), a fallback is needed. To do so, conservative models are defined for both trajectory / time-to-fly and buffers. They are calibrated based on the training data set.

In the process, the ML models are thus only used if a sufficient amount of consistent data is available to train the model and if it leads to ensuring that the FTD and ITD design criteria will be met. If this is not the case, then the fallback is the use of the conservative models. This approach allows the calculation of indicators to cover any aircraft in any situation, approaching the airport under safe conditions.

Figure 42 and Figure 43 show the coverage decision tree for respectively the separation indicator (FTD) and the compression spacing indicator (ITD).

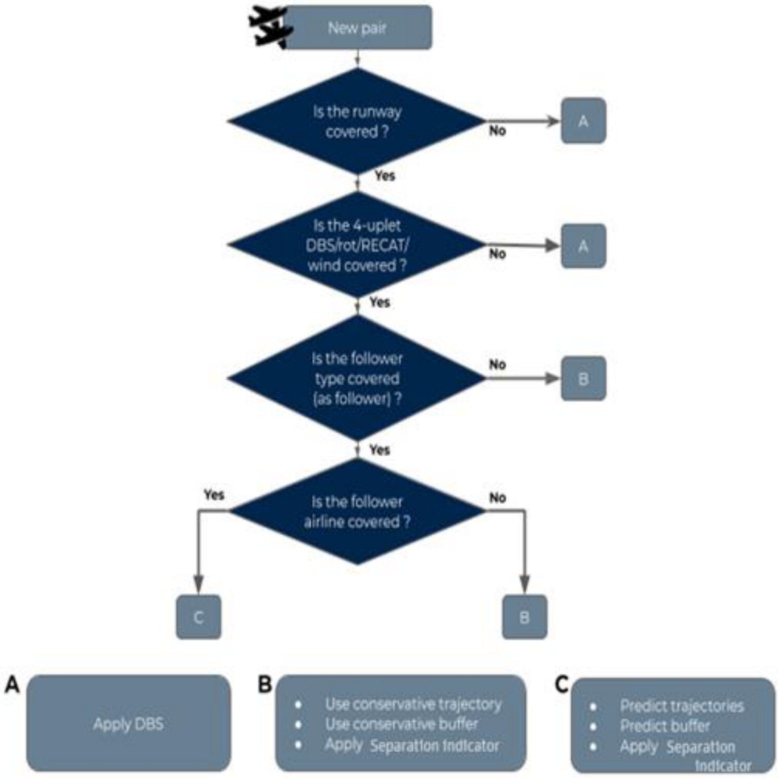


Figure 42 — Separation indicator (FTD) coverage decision tree

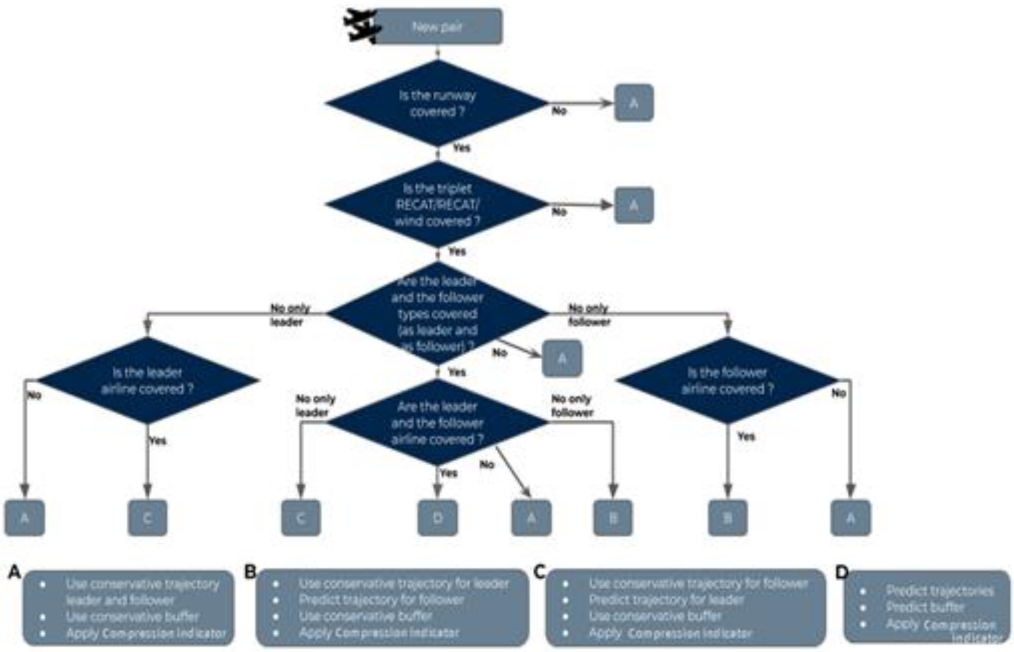


Figure 43 — Spacing indicator (ITD) coverage decision tree

OOD – out-of domain detection and solution

Conservative models

We have to ensure the reliability and safety of models. When a flight or a pair is not covered, a fallback is needed. To do so, conservative models are defined for both time-to-fly and buffer models. The time-to-fly conservative model is calibrated based on the training set whereas FTD and ITD buffer conservative models are calibrated based on the validation set. They can also be used for previously unseen cases. This conservative approach guarantees that safety is preserved with a limited operational cost in terms of over-spacing. The uncovered pairs are by definition too rare to dramatically affect the overall performance.

Regarding time-to-fly, two types of conservative models must be defined: conservatively slow time-to-fly models for the leaders, and conservatively fast time-to-fly models for the followers.

Coverage functions

The coverage functions are intended to decide in which cases we can rely on the trained ML time-to-fly and buffer models. This decision is taken comparing the obtained FTD and ITD separation performance compared to the FTD and ITD design criteria based on the empirical validation data set.

3.2.1.3. Description of the ML model data management process (inputs, outputs, functions)

The data ingestion in the ML pipeline is carried out following these steps:

1. Data loader: The data loader imports adequately the airport raw data files containing historical flights and meteorological information into the structured database.
2. Pre-processing: Even if raw data files have been stored and organised in the database using the data loaders, they are still not suited for feeding the ML toolbox. Indeed, several steps are needed such as de-noising the data, filtering some unsuited data and creating new complex features from raw data. The pre-processing procedure is made up of seven steps:
   - Detect the landing time and the landing runway
   - Compute the wind/temperature/pressure vertical profiles associated with every single flight
   - Detect the glide measurements and threshold their speeds
   - Associate the runway surface headwind and crosswind with all flights
   - Filter out the flights according to their final true air speed
   - Generate a time-to-fly profile for all flights
   - Create leader-follower couples

Once the data is ready for its management, the different ML models consider the cleaned data as follows:

- Time-to-fly data management

  Time-to-fly, FTD and ITD and buffer models are estimated on a training data set. For FTD and ITD buffer model training, flight couples (i.e. pairs) are required. Those couples are built assuming that arrivals taking place on the same runway in a certain time interval could have been an aircraft pair. A classic random train/test split on the flight couples is not suitable. The guarantee of independence between two subsets randomly drawn from the input data set is not straightforward. Since a flight can be simultaneously a leader in one or several couples and a follower in other couples, we need to ensure that couples involving the same flight are not distributed among training and test sets simultaneously. To solve this problem, the split of the input data set is made on a daily basis. All the couples landing on the same day are assigned to the same set. The split between training, validation and test sets follows the same rule as for time-to-fly, FTD and ITD buffer models. If both leader and follower are in the time-to-fly training set, the couple is in the buffers training set. If both are in the test set, the couple is assigned to the buffers test set.

- Time-to-fly data management

  In order to check the quality of the output, and then to ensure confidence in the models, the performance of the time-to-fly models is evaluated on an independent test data set. This independence is fundamental to avoid introducing bias in the evaluation.

- FTD buffer data management

  The features for a given couple are built by merging the flight and weather features of the leader and the follower aircraft. Additionally, the time and distance separation rules are concatenated to the resulting data vector.

  The targets are defined using the time-to-fly prediction on each follower flight. For each couple and each constraint, a predicted distance to fulfil a constraint is computed. It is then compared to the actual distance (using ground truth data) to compute the leeway (the maximal reduction to apply on the prediction without violating the constraints).

  We use a script to produce features and targets for both predictive and fully conservative FTD buffers. It is invoked twice:

  - Once for the predictive FTD buffers, using the **predictive time-to-fly** model
  - Once for the fully conservative FTD buffers, using the **conservative follower time-to-fly** model

- ITD buffer data management

  The features for a given couple are built by merging the flight and weather features of both, the leader and the follower aircraft. Additionally, the time and distance separation rules are concatenated to the resulting data vector.

For each couple, the FTD must have been computed previously, since one of the constraints is related to the respect of the FTD.

The targets are defined using the time-to-fly prediction on each leader and follower flights. For each couple and each constraint, a predicted distance to fulfil a constraint is computed. It is then compared to the actual distance (using ground truth data) to compute the leeway (the maximal reduction to apply on the prediction without violating the constraints).

3.2.1.4. ML model design process (AI techniques)

Time-to-fly learning — description

Since a flight can be present in several couples, and in order to avoid having these couples split between train and test (which will result in a loss of independence), we consider that all flights landing on the same day must either be train flights or test flights. The data set is then split on the landing day. Features and targets are stored separately.

Features and targets

The predictive model has the inputs and outputs described in Table 15.

Features	Target
Per each flight: Flight data: • Aircraft type • Category (RECAT) • Airline • Origin airport • Runway • Landing time (hour and day) Weather data: • Headwind at runway threshold • Crosswind at runway threshold • Total wind at runway threshold • Altitude headwind profile on the glide • Altitude crosswind profile on the glide • Altitude temperature profile on the glide • Altitude pressure profile on the glide	A vector containing 80 time-to-fly values, in seconds, at points located on the glide, between 0 and 20 km from the runway threshold. Those points are horizontally spaced by 250 metres.

Table 15 — Features and targets of the time-to-fly AI/ML constituent

FTD buffer learning — description

For buffer models, contrary to the time-to-fly case, the targets are not directly computed from raw data. The targets are defined as the maximal distance which can be safely subtracted from the conventional separation distance using the predictive time-to-fly model. Some processing is then needed to create the training and test sets.

We will learn a different model for each time-based separation/spacing constraint, but they will be serialised as a single .onnx file.

Features and targets

The predictive model has the inputs and outputs described in Table 16.

Features	Target
Leader and follower flight data: • Aircraft type • Category (RECAT) • Airline • Landing runway • Landing time and day Weather data: • Headwind at runway • Surface crosswind • Surface total wind • Down-sampled altitude headwind profile on the glide • Down-sampled crosswind profile on the glide • Down-sampled temperature profile on the glide • Down-sampled pressure profile on the glide Separation constraints: • Distance- and time-based wake separation minima • Leader ROT • Time-to-fly prediction for the follower aircraft	Four additional distances, in kilometres, to be added to the expected separation distance to meet the considered separation/spacing constraints

Table 16 — Features and targets of the FTD buffer ML constituent

ITD buffer learning — description

For ITD buffer models, contrary to the time-to-fly case and analogously to the FTD buffer one, the targets are not directly computed from raw data. The targets are defined as the maximal distance which can be subtracted from the separation distance predicted using the predictive time-to-fly model. Some processing is then needed to create the training and testing sets.

We will learn a different model for each separation/spacing constraint, but they will be serialised as a single .onnx file.

Features and targets

The predictive model has the inputs and outputs described in Table 17.

Features	Target
Leader and follower flight data: • Aircraft type • Category (RECAT) • Airline • Landing runway • Landing time and day Weather data: • Headwind at runway • Surface crosswind • Surface total wind • Down-sampled altitude headwind profile on the glide • Down-sampled crosswind profile on the glide • Down-sampled temperature profile on the glide • Down-sampled pressure profile on the glide Separation constraints: • Distance- and time-based wake separation minima • Leader ROT TTF output for leader and follower aircraft. • FTD output for the aircraft pair	Two additional distances, in kilometres, to be added to the expected separation distance to meet the considered constraints

Table 17 — Features and targets of the ITD buffer ML-constituent

3.2.1.5. Expected benefits and justification for Level 1B

The use of ML models to calculate time-based separation and spacing indicators makes it possible to use statistical behaviour in their computation and hence harmonise error rates and enhance prediction accuracy, which results in improved operational efficiency.

The benefits related to the use of FTD and ITD are further illustrated by providing three scenarios:

1. Current situation with no optimisation: In this scenario, there is neither FTD (allowing dynamic separation reduction) nor ITD (providing optimised spacing indication). A conservative spacing buffer is thus applied before leader deceleration (starting at deceleration fix (DF)) in order to cope with compression uncertainty resulting in a separation delivered at threshold showing some margin compared to the minimum.

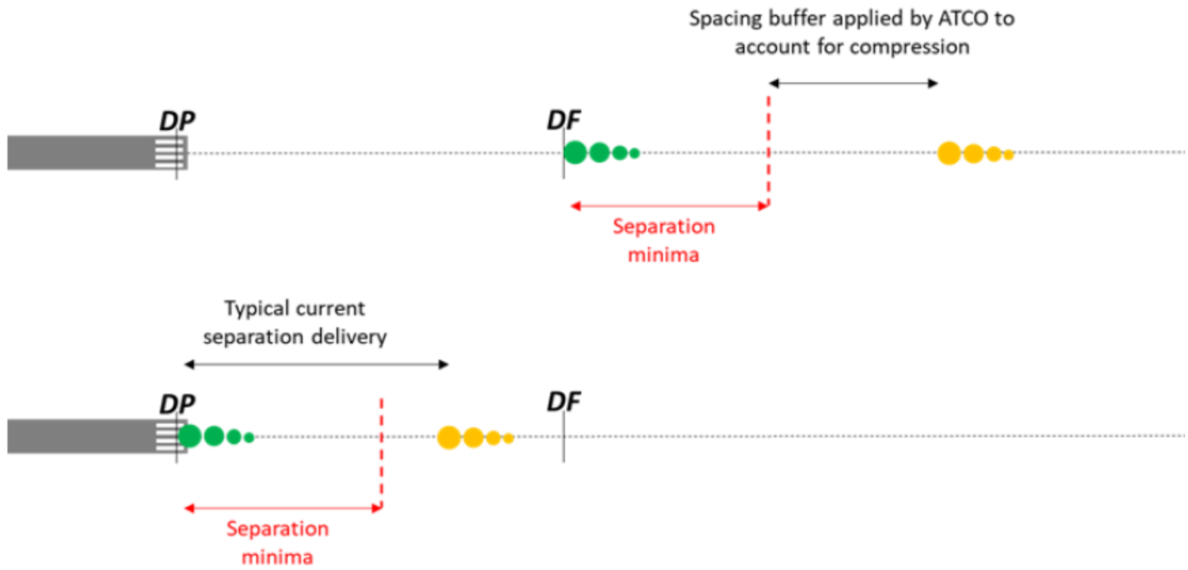


Figure 44 — Current ATCO support tool for separation and spacing

2. Use of ITD without change of separation minima mode: In this scenario, the use of the ITD allows optimised spacing of the flight before leader deceleration (starting at DF) resulting in a DBS delivered at threshold with higher accuracy.

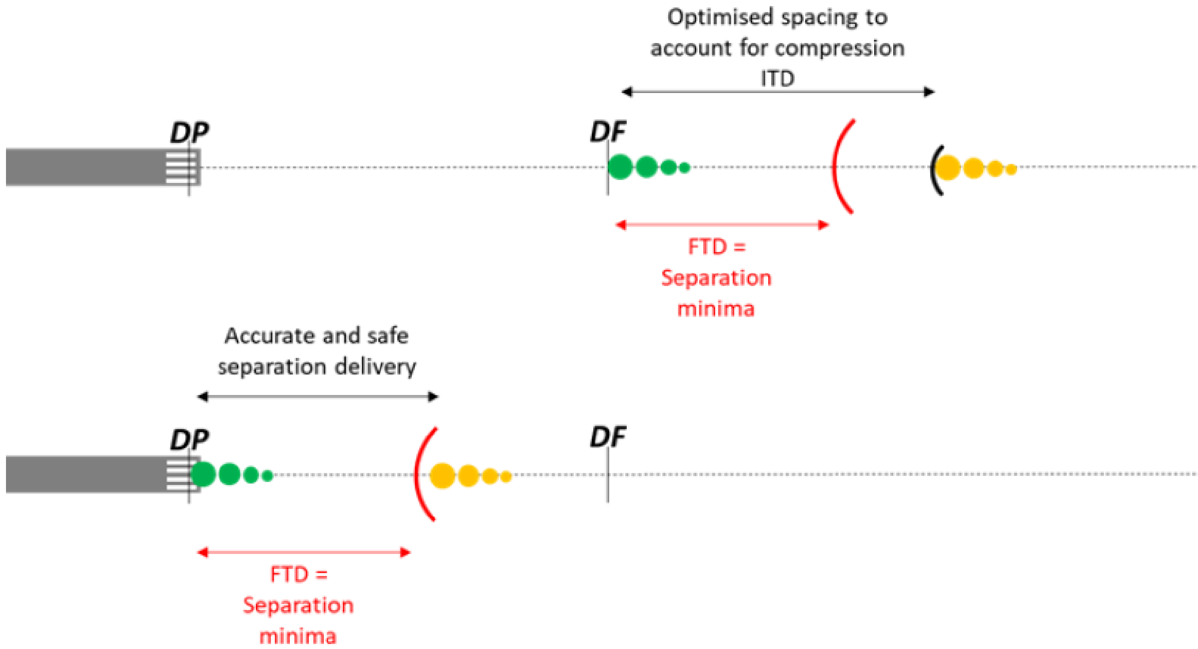


Figure 45 — DBS mode + ATC support tool for separation and spacing

3. Use of ITD with reduced separation minima: In this scenario, the use of FTD, allowing dynamic separation reduction (e.g. applying TBS) combined with ITD, improving the separation delivery accuracy, shows significant decrease in the delivered separation.

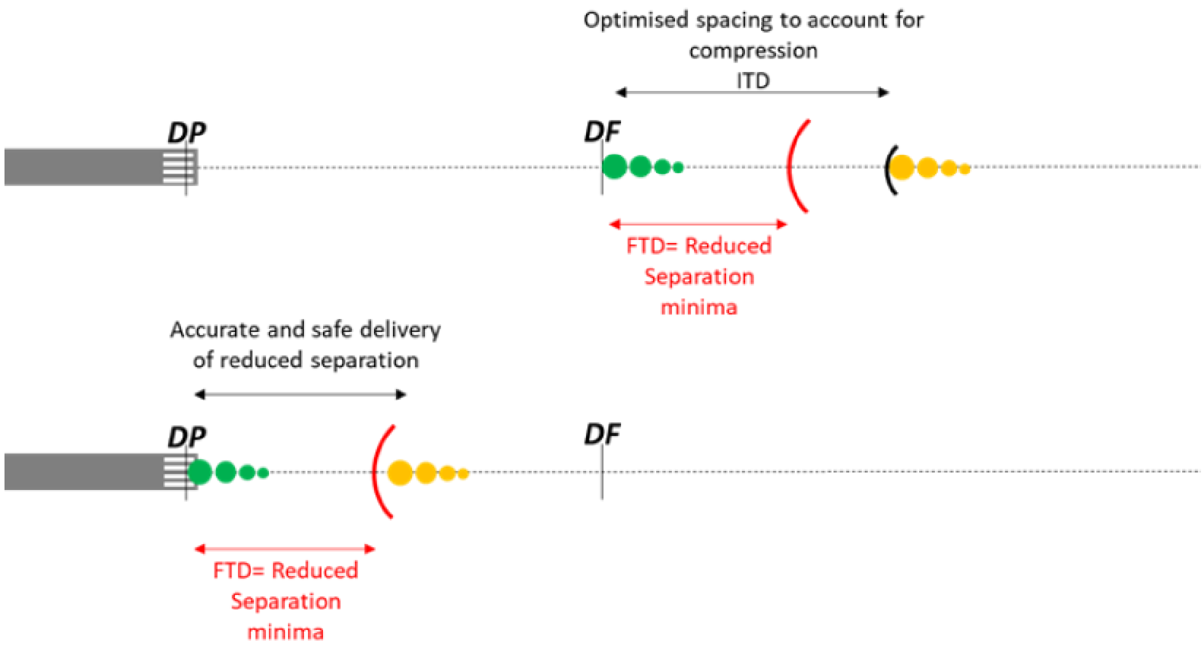


Figure 46 — TBS mode + ATC support tool for separation and spacing

Objective CL-01: The applicant should classify the AI-based system, based on the levels presented in Table 2, with adequate justifications.

Three applications were identified during use case development:

Distance-based separation — optimum runway delivery (DBS — ORD)

In this mode of application, the separation is based on distance as usual, but the tool provides an ITD indicator to the ATCO allowing the optimal spacing of aircraft on final approach.

The system assists the human in this case. The decision is solely the task and responsibility of the ATCO. Therefore, the AI-based system for this application is **Level 1B**.

Time-based separation (TBS)

In this mode of application, the separation is based on time. The separation minimum allowed to be applied by the ATCO is indicated on the CWP by the FTD. The ATCO is asked to target the indicator and is allowed to reduce the separation down to the FTD possibly below the current DBS. The ATCO has no means to verify that the FTD between two aircraft on final approach is safe.

Time-based separation — optimum runway delivery (TBS — ORD)

In this mode of application, the separation is based on time, similarly as in the previous mode, but the tool provides an ITD indicator to the ATCO allowing the optimal spacing of aircraft on final approach.

TBS application with or without ORD (cases b. and c.) provides the human with the information on the applicable separation minimum, which is dynamic (depending on aircraft-specific behaviour and prevailing wind). Because of the complexity of calculation, the decision logic is not provided to the ATCO. This is intrinsically related to the nature of the TBS solution and would also be the case for separation indicators calculated based on non-ML models. Strictly speaking, the responsibility to separate aircraft according to FTD still lies with the ATCO (i.e. the ATCO could still apply larger separation corresponding to DBS). However, the decision on applicable separation/spacing minima is transferred from the current distance-based rules known by the ATCO to an ML-based decision which the ATCO cannot override because of the lack of information. For that reason, these applications could be classified up to **Level 2**.

4. Use cases — aircraft production and maintenance

It should be noted that maintenance to assure continuing airworthiness of products is divided into two fundamentally different levels of activity:

— **Planning and scheduling of maintenance tasks:** this is typically done in by CAMOs.

In the generic wording of GM M.A.708(b)(4) 'the CAMO is responsible for determining what maintenance is required, when it has to be performed, by whom and to what standard in order to ensure the continuing airworthiness of the aircraft.', to determine *what* and *when* is currently decided based on fixed maintenance schedules and monitoring mainly simple usage parameters of the aircraft (e.g. flights, flight hours, calendar time), also including a regular update of the maintenance schedule taking into account in-service experience.

Modern aircraft providing an enormous amount of data in service and other information available (e.g. environmental data) do now provide a data pool which would allow scheduling maintenance much more appropriately and individually; however, to evaluate such big amount of data, sophisticated ML models are required.

— **Performance of maintenance:** this is typically done by approved maintenance organisations (often also referred to as Part-145 organisations, as they are covered in Part-145).

During performance of more complex maintenance tasks, it is normal to make use of special test equipment, today often including software. The use of test equipment containing AI/ML has a high potential to improve the quality of tests and inspections, while also improving efficiency.

In both domains, AI-based systems could be used to augment, support or replace human action, hence two examples are given.

4.1. Controlling corrosion by usage-driven inspections

4.1.1. Trustworthiness analysis

4.1.1.1. Description of the system

Currently the so-called corrosion prevention and control programmes (CPCP) managed at fleet level do control corrosion by scheduled inspections implemented at a fixed threshold and performed at fixed intervals, which are from time to time adjusted depending on the severity of corrosion found during previous inspections.

Today we have detailed data about where the aircraft has been at which point in time, which temperature, rainfall, de-icing agents, corrosion-critical pollutants, etc. it has been exposed to, how it has been utilised, which corrosion findings have been made on other aircraft, and a lot of other usage, utilisation, maintenance, repair, events etc. it has experienced. From this huge data pool an ML model could be trained to evaluate the individual corrosion risk of all relevant locations within each individual aircraft, to allow the CAMO to schedule focused inspections for corrosion at the most appropriate time (when airworthiness is not at risk, the probability of findings is high, and repair is still economic).

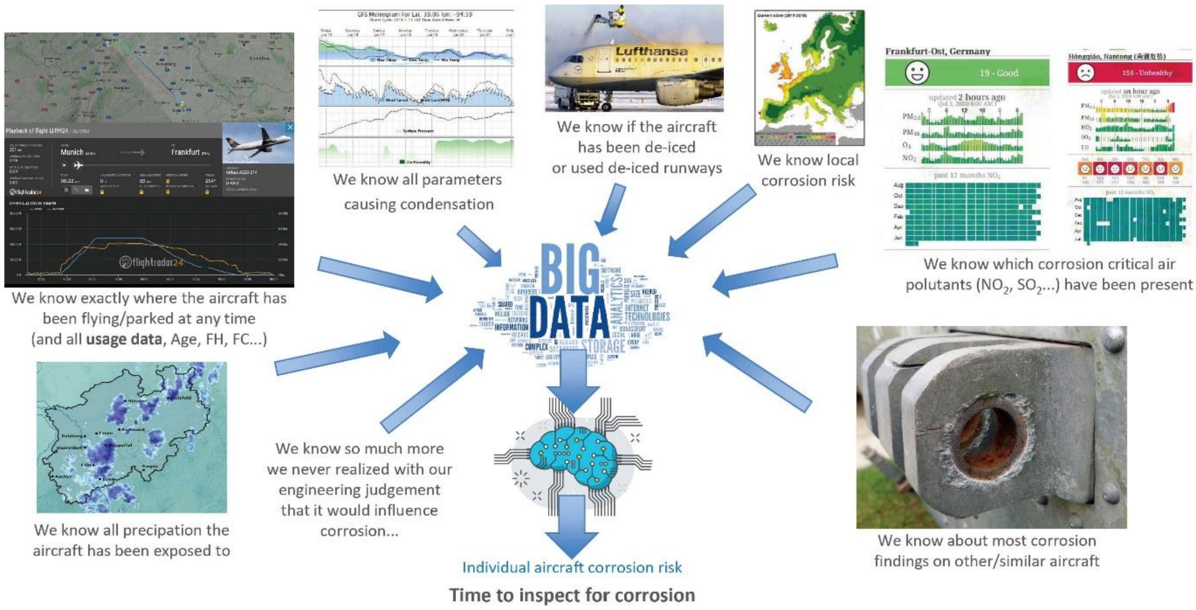


Figure 47 — General philosophy of CPCP by utilisation of data and AI

Objective CO-03: The applicant should determine the AI-based system taking into account domain-specific definitions of 'system'.

A system at the CAMO would constantly receive operational data from the aircraft, either directly through satellite data link (e.g. ACARS), or indirectly as download by the operator or a contracted service provider. Additional data (e.g. weather data, whether de-icing has been performed, occurrences, repairs) would be constantly acquired as well creating a database covering the full-service history of all individual aircraft under the control of the CAMO.

This does already happen today, but to a lower extent and not specifically focusing on corrosion, but is typically more related to system components (which do provide more specific data easily processed by conventional deterministic algorithms).

A special system would then analyse the data collected, making use of an ML model trained on similar data of other aircraft in the past to predict the level of corrosion which is probably present at specific areas within individual aircraft.

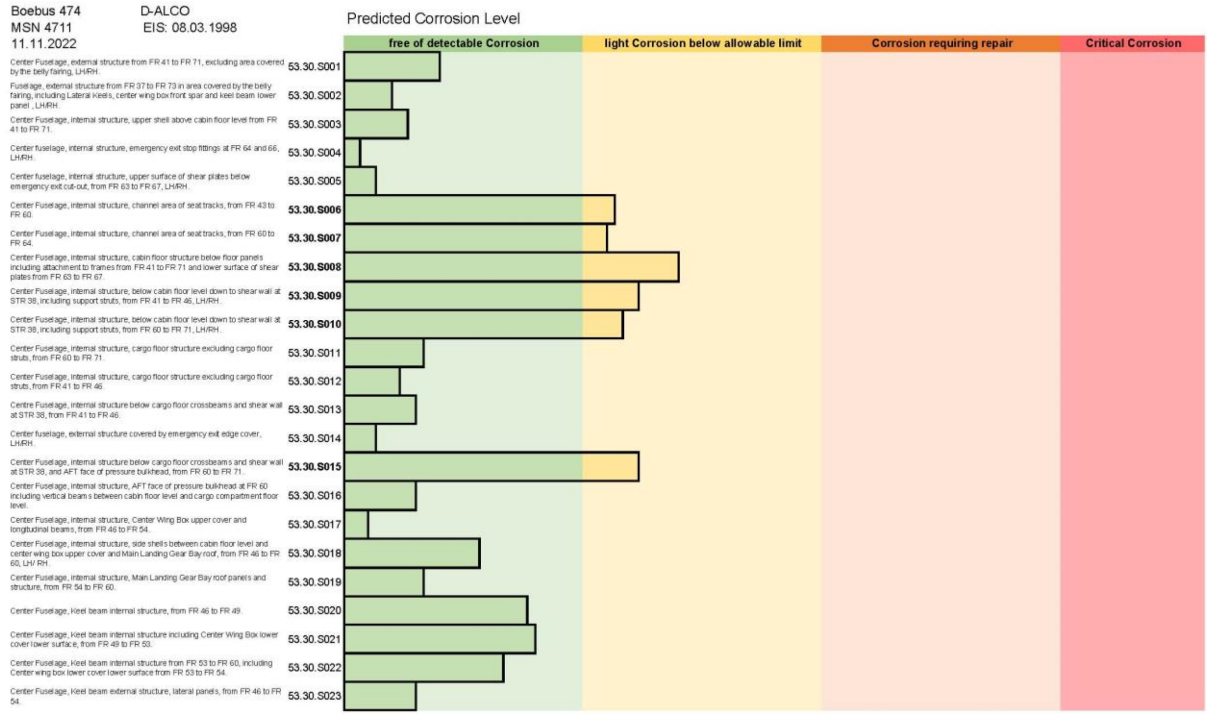


Figure 48 — Example of a possible system output: predicted corrosion in specific areas

4.1.1.2. Description of the system(s) involved (inputs, outputs, functions)

Input:

Usage data of individual aircraft
Environmental data (covering the location at the time of operation)
Operational information (e.g. type of cargo loaded, seafood?)
Findings from inspections (in all of the fleet)

Output:

Corrosion risk level at individual locations of individual aircraft
(output could be in the form of an alert or regular status information)

Type of AI:

Pattern detection in large databases

4.1.1.3. Expected benefits and justification for Level 1

The application is expected to improve corrosion control by identifying areas of specific aircraft which have been exposed to increased corrosion risk and require an earlier inspection to limit the severity of structural degradation, or to identify areas of specific aircraft which have not been exposed to high corrosion justifying a later inspection reducing cost, downtime and the risk of access-induced damage. This would allow the increase of safety while reducing cost at the same time.

For the maintenance planning activity, it is not so easy to determine the role of humans. Whereas the actual inspection at the aircraft is still performed by humans, the planning of such physical human interference with the aircraft could be implemented at a high level of automation.

Maintenance planning is done today already using computers. Even if performed by humans, all maintenance work at the aircraft is scheduled through computer tools. There is however also always a certain level of human involvement; for example, humans decide which mechanic/inspector should perform which of the scheduled tasks. As such all physical human interference with the aircraft requested by the system can always be overridden by humans (they can always inspect an aircraft although not requested, they can always reject the request to inspect).

In a first application, the system would only support the maintenance planning engineer in deciding when to perform a corrosion inspection at a certain area of an individual aircraft, which would make it a Level 1B system. As the decision to perform a specific maintenance task is always following several considerations (e.g. aircraft availability at the place of the maintenance organisation, availability of hangar space, access requirements and the possibility to perform several tasks at the same time), the final decision is always complex, so the system may also be understood as being only Level 1A and only supporting the maintenance engineer by providing and analysing information.

It could however be possible to upgrade the system up to Level 3A, if all those practical and economical aspects of maintenance planning could be ignored, and the system could automatically schedule inspections without any human interference at CAMO level.

The system could be set up with two types of fundamentally different output:

— Providing the maintenance engineer with regular (e.g. weekly) reports of the aircraft status

— Providing the maintenance engineer with a warning if an area reaches a selected alert threshold

This is similar to the concept of installing either an indication or a warning on the flight deck to either allow monitoring by the flight crew or to alert them when required. There are advantages and disadvantages for both concepts and a combination is also possible.

This will finally make the difference between a Level 1A or 1B system.

Objective CL-01: The applicant should classify the AI-based system, based on the levels presented in Table 2, with adequate justifications.

The ***AI Level 1A 'Human augmentation'*** classification is justified by only providing additional/advisory information (***support to information analysis***) to the maintenance engineer without any suggestion for action or decision-making.

4.2. Damage detection in images (X-Ray, ultrasonic, thermography)

4.2.1. Trustworthiness analysis — description of the system and ConOps

Visual inspections and non-destructive testing (NDT) are typical methods to detect damage of aircraft structure.
Those tasks rely on specifically trained inspectors visually detecting damages either by directly inspecting items or by evaluating pictures (e.g. X-Ray pictures). With today's technology, as most pictures are no longer produced physically but digitally, detection of damages is already typically performed on computer screens, either 'offline' in offices after taking them at the aircraft or even 'online' directly at the aircraft using portable test equipment with displays.
This use case could be similarly applied to a variety of images, from optical pictures (photographs) taken by humans, fixed cameras or programmed or potentially autonomously acting machines (such as UAS that are already used successfully in maintenance to detect damages on structures), through sophisticated imaging technology like X-ray or thermography (infrared) up to fully synthetic pictures generated by scanning an area with ultrasonic or eddy current probes. All these inspection methods finally produce digital images which have to be checked for showing damages or defects. Recognising damage shown on digital pictures would be a typical application of AI, similar to some other applications currently widely discussed (runway detection, 'see-and-avoid'). The learning algorithm and the training of the ML model of course would be individually different for the different types of image to be evaluated.

To be able to address specific issues, the example chosen is the analysis of thermographic images, a method when pictures of the aircraft are taken by digital optical means in the infrared range of the light spectrum in combination with production of a temperature difference (typically heating up the appropriate test item and then inspect it in a room colder than the item), allowing the detection of several types of typical damage in composites structures by visualising the local thermal capacity and conductivity of the item. Infrared cameras have advanced enormously in the last two decades and are now as easy to use as any other optical camera.

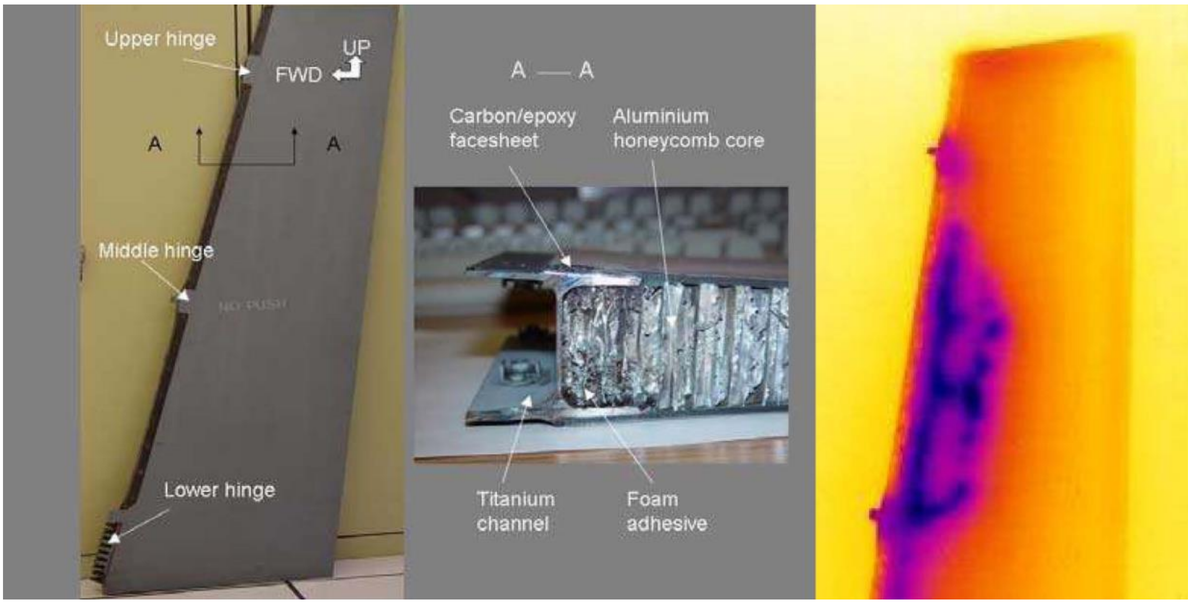


Figure 49 — Thermographic images of a fighter aircraft rudder showing water ingress in honeycomb cells

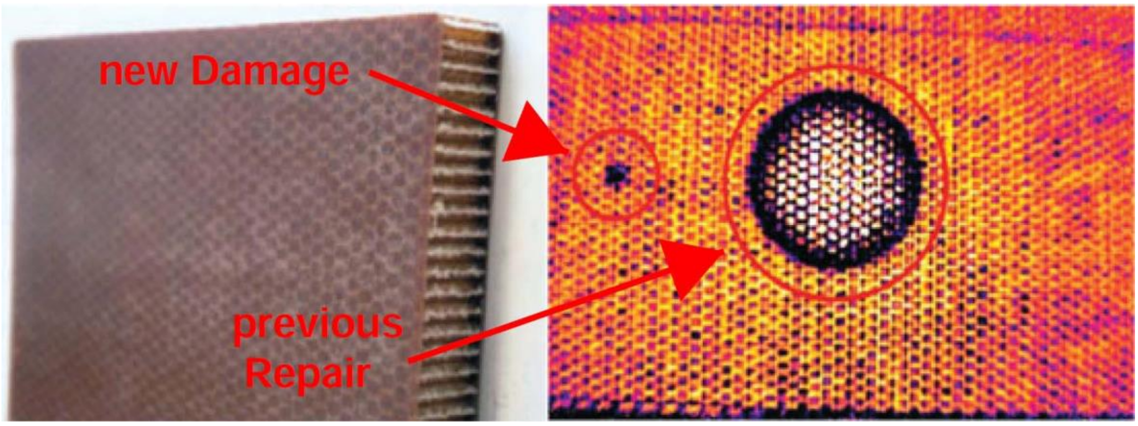


Figure 50 — Optical and thermographic image of a GFRP sandwich panel

4.2.1.1. Description of the system

Objective CO-03: The applicant should determine the AI-based system taking into account domain-specific definitions of 'system'.

A system supporting thermographic inspections of aircraft could be integrated in portable test equipment to be used at the aircraft.

Such a system would not only show, but also analyse the digital image produced with the infrared camera and provide the inspector with additional information and classification of the details seen in the picture by damage type and criticality. A data link to the operator/CAMO/manufacturer databases could be envisaged in the future.

Figure 51 — Portable thermographic test equipment, potentially including an image evaluation system

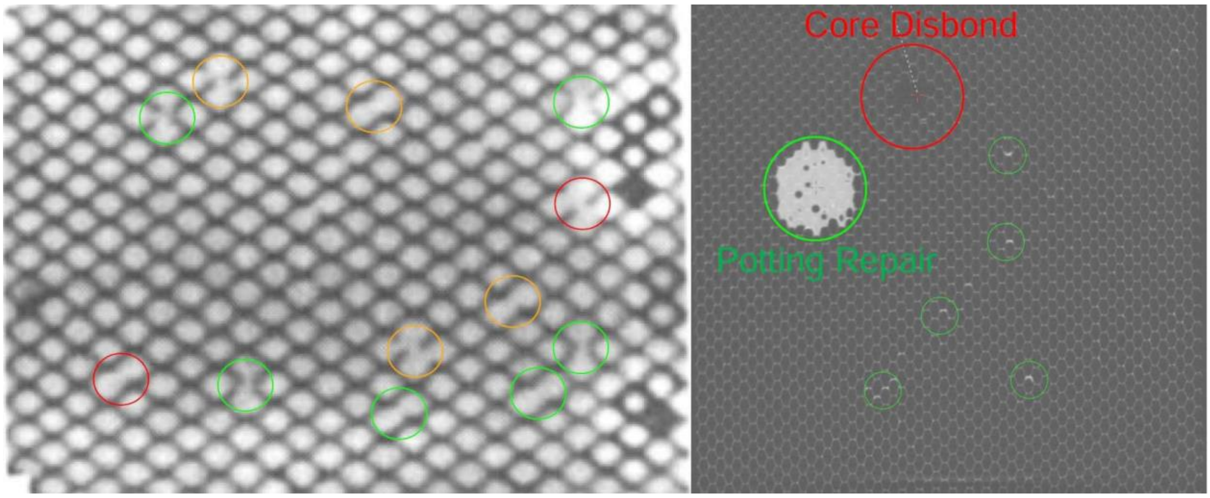


Figure 52 — Example of how the system could mark some areas in images to support inspection of honeycomb sandwich

4.2.1.2. Concept of operations

Objective CO-04: The applicant should define and document the ConOps for the AI-based system, including the task allocation pattern between the end user(s) and the AI-based system. A focus should be put on the definition of the OD and on the capture of specific operational limitations and assumptions.

The terms 'operation' and 'limitation' are not typical in the maintenance domain.

The AI-based system is intended to be used for NDT to inspect aircraft structures. The system needs to be trained on specific types of structures (e.g. monolithic composites, bonded metal), specific materials (e.g. CFRP, aluminium) and specific failures/damages/defects (e.g. delaminations, disbond, water ingress). Each specific system configuration is strictly limited to be used on the appropriate type of structure.

This is comparable to the situation today with human inspectors, who are also just qualified to perform certain NDT methods on certain types of structure. Training the ML model is comparable to the requirements for human inspectors to be specifically trained for the NDT they perform.

Additionally M.A.608 requires that 'Tools and equipment shall be controlled and calibrated to an officially recognised standard.' Specifically for NDT equipment, the individual tools and equipment used have individual sensitivity and detection characteristics. It is therefore normal practice that those are adjusted in line with equipment and aircraft manufacturer instructions in order to be calibrated. To this purpose, defects (type, size) are predefined by the manufacturer by use of a 'standard' (i.e. one or more test pieces with an artificial defect as defined by the aircraft manufacturer). This very same philosophy is applicable for ML. The end user needs to train (calibrate) the ML model (equipment) with a data set (standard) defined by the aircraft manufacturer. Then the end user needs to demonstrate that the trained model is able to correctly classify all the standard samples.

M.A.608 also covers 'verified equivalents as listed in the maintenance organisation manual' to 'the equipment and tools specified in the maintenance data', meaning it is allowed and normal practice not to use the specific NDT method and/or equipment required by the manufacturer, but an alternative method/equipment verified to be equivalent. This implicitly allows the use of equipment making use of AI/ML if it is verified to provide equivalent detection capability. This of course needs to be demonstrated to the approving authority.

4.2.1.3. Description of the system(s) involved (inputs, outputs, functions)

Input:

Digital image from an infrared camera

Output:

Digital picture with highlighted areas of interest

Information about the type and severity of damage found

Type of AI:

Image recognition

4.2.1.4. Expected benefits and justification for Level 1

The application is expected to reduce workload and improve the quality of inspection. A major issue of human performance is the change in attention over the day as a lot of maintenance is performed at night either as line maintenance given that the aircraft flies during the day, or in a 24-hour activity to keep the downtime short. The use of AI-based systems would allow for a more consistent quality of inspections reducing the impact of human factors.

Additionally, the use of an image assessment based on a computer system allows the inspector to be provided with additional information derived from databases, e.g. by recognising which exact location of the aircraft is shown in the picture, to highlight the location of previous repairs or to show modifications and to provide additional information such as the allowable damage size in that area, information which today has to be manually produced by the inspector using the appropriate handbooks.

In a first step, the system would be classified as a Level 1B, as the system would support the inspector to take the decision whether:

— the inspected structure is free of defects;

— it only contains allowable damage; or

— a deeper inspection or a repair is required before the aircraft can return to service.

The final decision and the need to sign off the inspection would remain with the human; the system would just support this.

In a later stage, higher levels would be technically possible but would require a change of the current philosophy about how maintenance is performed, also requiring changes to regulatory requirements.

Objective CL-01: The applicant should classify the AI-based system, based on the levels presented in Table 2, with adequate justifications.

The ***AI Level 1B 'Human assistance'*** classification is justified by providing information to ***support decision/action selection*** to the maintenance engineer.

4.2.1.5. Potential safety impact

A risk of complacency and over-reliance on the applications exists. Inspectors may be biased in their final decision if the system would classify a detail in the image to not show a defect and they may not check as thoroughly as today when being very confident in the performance of the system.

As many inspections are intended to prevent catastrophic failure by detecting damages or defects before they grow to a critical size (damage tolerance concept), non-detection of existing damage can have a safety impact. As long as the final decision is still with the human and the system just provides support, these safety risks exist in combination with human factors, for which safety management systems are already in place.

5. Use cases — training / FSTD

5.1. Assessment of training performance

This use case will be developed in a future revision of this document.

6. Use cases — aerodromes

It needs to be made clear that the scope of the European rules for aerodrome safety address the aviation activities and operational processes on the airside only; and that the so-called landside is not covered by these rules. It is however inside the terminal and in relation to passenger services and passenger management where AI has manifold application areas. For example, AI is integrated with airport security systems such as screening, perimeter security and surveillance since these will enable the aerodrome operator to improve the safety and security of the passengers. Furthermore, border control and police forces use facial recognition and millimetre-wave technologies to scan people walking through a portable security gate. ML techniques are used to automatically analyse data for threats, including explosives and firearms, while ignoring non-dangerous items — for example, keys and belt buckles — users may be carrying. In addition, ML techniques are used by customs to detect prohibited or restricted items in luggage.

On the airside, there are by comparison fewer use cases of AI/ML in the service of aerodrome safety. The most well-known ones are:

6.1. Detection of foreign object debris (FOD) on the runway

The presence of FOD on the runways can end up damaging aircraft, vehicle and equipment, and ultimately can even cause accidents. FOD prevention and the inspection of movement area for the presence of FOD is a core activity of aerodrome operators. Because physical inspections of runways are time-consuming and reduce capacity and are also not free of human detection error, the use of technological solutions for FOD detection has long been attempted. More recently the application of ML by such systems has been included, as this way the detection of FOD and the related alerts would be more reliable. Since there is a considerable market for FOD detection systems and not all systems are of the desired reliability and maturity, it is not advised to single any of them out.

This use case may be further developed in a future revision of this document. EASA would welcome if it could be alerted of any impediments to the evolution of such systems in today's rules for aerodrome safety.

6.2. Avian radars

At airports, the prevention of bird strikes to aircraft is an ongoing challenge. Avian radars can track the exact flight paths of both flocks and individual birds up to 10 km. They automatically detect and log hundreds of birds simultaneously, including their size, speed, direction, and flight path. Bird radar tracks may be presented to tablets of the bird control vehicles in real time, thereby creating situational awareness and allowing for better response by bird control staff. Collection of data related to bird activities may be used to predict future problematic areas, identify specific patterns and support decision-making. Since there is a considerable market for avian radar systems and as not all systems are of the desired reliability and maturity, it is not advised to single any of them out.

This use case may be further developed in a future revision of this document. EASA would welcome if it could be alerted of any impediments to the evolution of such systems in today's rules for aerodrome safety.

6.3. UAS detection systems

Similar to the situation with birds, the surroundings of aerodromes may be affected by the unlawful use of unmanned aircraft. This represents a hazard to aircraft landing and taking off from the runways. UAS detection, tracking and classification, in conjunction with alert and even neutralisation functions by reliable technological solutions will one day provide the desired safety and security for the airport environment; however, as today's technology-based C-UAS solutions are mostly multi-sensor-based, no single technology can perform several functionalities satisfactorily. The improvement of such technologies with ML appears to be the logical evolution.

Since there is a considerable market for such UAS detection systems and as not all systems are of the desired reliability and maturity, it is not advised to single any of them out.

This use case may be further developed in a future revision of this document. EASA would welcome if it could be alerted of any impediments to the evolution of such systems in today's rules for aerodrome safety.

7. Use cases — environmental protection

7.1. Engine thrust and flight emissions estimation

This use case will be developed in a future revision of this document.

8. Use cases — safety management

8.1. Quality management of the European Central Repository (ECR)

This use case will be developed in a future revision of this document.

8.2. Support to automatic safety report data capture

This use case will be developed in a future revision of this document.

8.3. Support to automatic risk classification

This use case will be developed in a future revision of this document.

G. Annex 3 — Definitions and acronyms

1. Definitions

Accessibility — The extent to which products, systems, services, environments and facilities can be used by people from a population with the widest range of user needs, characteristics and capabilities to achieve identified goals in identified contexts of use (which includes direct use or use supported by assistive technologies)[31].

Accountability — This term refers to the idea that one is responsible for their action – and as a corollary their consequences – and must be able to explain their aims, motivations, and reasons. Accountability has several dimensions. Accountability is sometimes required by law. For example, the General Data Protection Regulation (GDPR) requires organisations that process personal data to ensure that security measures are in place to prevent data breaches and report if these fail[32].

Accuracy (of the data) — The degree of conformance between the estimated or measured value and its true value.

Adaptivity (of the learning process) — The ability to improve performance by learning from experience. [In the ML context,] **adaptive learning** refers to learning capability during the operations (see also **online learning**).

Artificial intelligence (AI) — Technology that can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with[33].

AI-based system — A system that is developed with one or more of the techniques and approaches listed in Annex I to the EU AI Act and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with[34].

Artificial neural network (ANN) or neural network (NN) — A computational graph which consists of connected nodes ('neurons') that define the order in which operations are performed on the input. Neurons are connected by edges which are parameterised by weights (and biases). Neurons are organised in layers, specifically an input layer, several intermediate layers, and an output layer. This document refers to a specific type of neural network that is particularly suited to process image data: convolutional neural networks (CNNs) which use parameterised convolution operations to compute their outputs.

Commonly used types of neural networks are to be highlighted:

— **Convolutional neural networks (CNNs)** — A specific type of deep neural networks that are particularly suited to process image data, based on convolution operators. (Daedalean, 2020)

— **Recurrent neural networks (RNNs)** — A type of neural network that involves directed cycles in memory.

31 Source: adapted from (EU High-Level Expert Group on AI, 2020).

32 Source: adapted from (EU High-Level Expert Group on AI, 2020).

33 Source: adapted from (EU Commission, 2021).

34 Source: (EU Commission, 2021)

Attachment — Is the state of strong emotional bond between the end user and the AI-based system[35].

Auditability — Refers to the ability of an AI-based system to undergo the assessment of the system's learning algorithms, data and design processes. This does not necessarily imply that information about business models and intellectual property related to the AI-based system must always be openly available. Ensuring traceability and logging mechanisms from the early design phase of the AI-based system can help enable the system's auditability[36].

Authority — The ability to make decisions and take actions without the need for approval from another member involved in the operations.

Automation — The use of control systems and information technologies reducing the need for human input, typically for repetitive tasks.

Autonomy — Characteristic of a system that is capable of modifying its intended domain of use or goal without external intervention, control or oversight[37].

Advanced automation — The use of a system that, under specified conditions, functions without human intervention[38].

Bias — Different definitions of bias have to be considered depending on the context:

— **Bias** (in the data) — The common definition of data bias is that the available data is not representative of the population or phenomenon of study.

— **Bias** (in the ML model) — An error from erroneous assumptions in the learning [process]. High bias can cause a learning algorithm to miss the relevant relations between attributes and target outputs (= underfitting).

Big Data — A recent and fast evolving technology, which allows the analysis of a big amount of data (more than terabytes), with a high velocity (high speed of data processing), from various sources (sensors, images, texts, etc.), and which might be unstructured (not standardised format).

Completeness — A data set is complete if it sufficiently (i.e. as specified in the DQRs) covers the entire space of the operational design domain for the intended application.

Compromise of AI/ML application components — Refers to the compromise of a component or developing tool of the AI/ML application (ENISA, December 2021). Example: compromise of one of the open-source libraries used by the developers to implement the learning algorithm[39].

Concept of operations (ConOps) — A ConOps is a human-centric document that describes operational scenarios for a proposed system from the users' operational viewpoint.

[35] Source: adapted from WordReference.com LLC
[36] Source: adapted from (EU High-Level Expert Group on AI, 2020).
[37] Source: adapted from ISO/IEC 22989:2022(en), 3.1.7.
[38] Source: adapted from ISO/IEC 22989:2022(en), 3.1.7.
[39] Source: adapted from (ENISA, December 2021).

Corner case (see also edge case) — Relates to a situation that, considering at least two parameters of the AI/ML constituent ODD, occurs rarely on all of these parameters (i.e. low representation of the associated values in the distribution for those parameters).

Cost function — A function that measures the performance of an ML model/constituent for given data and quantifies the error between predicted values and ground-truth values.

Critical maintenance task — A maintenance task that involves the assembly or any disturbance of a system or any part on an aircraft, engine or propeller that, if an error occurred during its performance, could directly endanger the flight safety.

Data-driven AI — An approach focusing on building a system that can learn a function based on having trained on a large number of examples.

Data governance — A data management concept concerning the capability of an organisation to ensure that high data quality exists throughout the complete life cycle of the data, and data controls are implemented that support business objectives. The key focus areas of data governance include data availability, usability, consistency, integrity, and sharing. It also regards establishing processes to ensure effective data management throughout the enterprise such as accountability for the adverse effects of poor data quality and ensuring that the data which an enterprise has can be used by the entire organisation[40].

Data life cycle management — Data life cycle management corresponds to the set of applicants' procedures in place for managing the flow of data used during the life cycle of the AI/ML constituent, from identification and collection of the data to the time when it becomes obsolete and is deleted.

Data protection impact assessment (DPIA) — Evaluation of the effects that the processing of personal data might have on individuals to whom the data relates. A DPIA is necessary in all cases in which the technology creates a high risk of violation of the rights and freedoms of individuals. The law requires a DPIA in case of automated processing, including profiling (i), processing of personal data revealing sensitive information like racial of ethnic origin, political opinions, religious or philosophical beliefs (ii), processing of personal data relating to criminal convictions and offences (iii) and systematic monitoring of a publicly accessible area on a large scale (iv)[41].

Data Protection Officer (DPO) — This denotes an expert on data protection law. The function of a DPO is to internally monitor a public or private organisation's compliance with GDPR. Public or private organisations must appoint DPOs in the following circumstances: (i) data processing activities are carried out by a public authority or body, except for courts acting in their judicial capacity; (ii) the processing of personal data requires regular and systematic monitoring of individuals on a large scale; (iii) the processing of personal data reveals sensitive information like racial of ethnic origin, political opinions, religious or philosophical beliefs, or refers to criminal convictions and offences. A DPO must be independent of the appointing organisation[42].

[40] Source: adapted from (EU High-Level Expert Group on AI, 2020).
[41] Source: adapted from (EU High-Level Expert Group on AI, 2020).
[42] Source: adapted from (EU High-Level Expert Group on AI, 2020).

Data set[43] (in ML in general) — The sample of data used for various development phases of the model, i.e. the model training, the learning process verification, and the inference model verification.

— **Training data set** — Data that is input to an ML model in order to establish its behaviour.

— **Validation data set**— Used to tune a subset of the hyper-parameters of a model (e.g. number of hidden layers, learning rate, etc.).

— **Test data set**— Used to assess the performance of the model, independent of the training data set.

Data for safety (EASA) — Data4Safety (also known as D4S) is a data collection and analysis programme that supports the goal of ensuring the highest common level of safety and environmental protection for the European aviation system.

The programme aims at collecting and gathering all data that may support the management of safety risks at European level. This includes safety reports (or occurrences), flight data (i.e. data collected from the aircraft systems via a non-protected recording system, such as a quick-access recorder), surveillance data (air traffic data), weather data — but those are only a few from a much longer list.

As for the analysis, the programme's ultimate goal is to help to 'know where to look' and to 'see it coming'. In other words, it will support the performance-based environment and set up a more predictive system.

More specifically, the programme will facilitate better knowledge of where the risks are (safety issue identification), determine the nature of these risks (risk assessment) and verify whether the safety actions are delivering the needed level of safety (performance measurement). It aims to develop the capability to discover vulnerabilities in the system across terabytes of data [Source: EASA].

Decision — A conclusion or resolution reached after consideration[44]. A choice that is made about something after thinking about several possibilities[45].

Decision-making – The cognitive process resulting in the selection of a course of action among several possible alternative options[46]**.** Automated or automatic decision-making is the process of making a decision by automated means without any human involvement[47].

Deep learning (DL) — A specific type of machine learning based on the use of large neural networks to learn abstract representations of the input data by composing many layers.

Derived requirements — Requirements produced by the learning assurance processes which (a) are not directly traceable to higher-level requirements, and/or (b) specify behaviour beyond that specified by the requirements allocated to the AI/ML constituent.

Determinism — A system is deterministic if when given identical inputs produces identical outputs.

---

43 Source: adapted from (ER-022 - EUROCAE, 2021).

44 Source: OxfordLanguages.

45 Source: adapted from the Cambridge Dictionary.

46 Source: adapted from Wikipedia.

47 Source: adapted from ico.org.uk.

Development assurance — All those planned and systematic actions used to substantiate, to an adequate level of confidence, that errors in requirements, design, and implementation have been identified and corrected such that the system satisfies the applicable certification basis.

Development error — A mistake in requirements, design, or implementation.

Domain — Operational area in which a system incorporating an ML subsystem could be implemented/used. Examples of domains considered in the scope of this guideline are ATM/ANS, air operations, flight crew training, environmental protection or aerodromes.

Edge case (see also corner case) — Relates to a situation that, considering a given parameter of the AI/ML constituent ODD, occurs rarely (i.e. low representation of the associated value in the distribution for that parameter).

End user — An end user is the person that ultimately uses or is intended to ultimately use the AI-based system. This could either be a consumer or a professional within a public or private organisation. The end user stands in contrast to users who support or maintain the product[48].

Evasion (attack) — A type of attack in which the attacker alters the ML model's inputs to find small perturbations leading to large modification of its outputs (e.g. object detection errors, decision errors, etc.). It is as if the attacker created an optical illusion for the ML model. Such modified inputs are often called adversarial examples (ENISA, December 2021). Example: the projection of images on a runway could lead the AI-based system of a visual landing guidance assistant to alert the pilot on an object on this runway[49].

Failure — An occurrence which affects the operation of a component, part, or element such that it can no longer function as intended (this includes both loss of function and malfunction). Note: Errors may cause failures, but are not considered to be failures.

Fairness — Refers to ensuring equal opportunities and non-discriminatory practices applied to individuals or groups of users (or end users). Definition based on EU guidelines on non-discrimination[50].

Feature (in computer science) — A feature is any piece of information which is relevant for solving the computational task related to a certain application.

— **Feature** (in machine learning in general) — A feature is an individual measurable property or characteristic of a phenomenon being observed.

— **Feature** (in computer vision) — A feature is a piece of information about the content of an image; typically about whether a certain region of the image has certain properties.

General Data Protection Regulation (GDPR) — EU's data protection law, refer to https://gdpr.eu for more details.

48 Source: adapted from (EU High-Level Expert Group on AI, 2020).

49 Source: adapted from (ENISA, December 2021)

50 Article 21 'Non-discrimination' | European Union Agency for Fundamental Rights (europa.eu)

Human agency — Human agency is the capacity of human beings to make choices and to impose those choices on the world.

Hyper-parameter — A parameter that is used to control the algorithm's behaviour during the learning process (e.g. for deep learning with neural networks, the learning rate, the batch size or the initialisation strategy). Hyper-parameters affect the time and memory cost of running the learning algorithm, or the quality of the model obtained at the end of the training process. By contrast, other parameters, such as node weights or biases, are the result of the training process[51].

Independence — in this document, depending on the context, this word has several possible definitions:

— **Safety assessment context –** A concept that minimises the likelihood of common mode errors and cascade failures between aircraft/system functions or items.

— **Assurance context –** Separation of responsibilities that assures the accomplishment of objective evaluation e.g. validation activities not performed solely by the developer of the requirement of a system or item.

— **Data management context –** Two data sets are independent when they do not share common data and have a certain level of statistical independence (also referred to as 'i.i.d'[52] in statistics).

Inference — The process of feeding the machine learning model an input and computing its output. See also related definition of **Training**.

Information security — The preservation of confidentiality, integrity, authenticity and availability of network and information systems.

Inlier — An inlier is a data value that incorrectly lies within the AI/ML constituent ODD following an error during data management. A simple example of an inlier might be a value in a record reported in the wrong units, say degrees Fahrenheit instead of degrees Celsius. Because inliers are difficult to distinguish from good data values, they are sometimes difficult to find and correct[53].

Input space — Given a set of training examples of the form $\{(x_1,y_1) \ldots (x_N,y_N)\}$ such that x_i is the feature vector of the i-th example and y_i is its label (i.e. class), a learning algorithm seeks a function $g : X \rightarrow Y$, where X is the input space and Y is the output space.

Integrity — An attribute of the system or an item indicating that it can be relied upon to work correctly on demand.

— **Integrity** (of data) — A degree of assurance that the data and its value has not been lost or altered since the data collection.

— **Integrity** (of a service) – A property of a service provided by a service provider indicating that it can be relied upon to be delivered correctly on demand.

---

[51] Source: adapted from (Goodfellow-et-al, 2016).

[52] In probability theory and statistics, a collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent. This property is usually abbreviated as i.i.d. or iid or IID.

[53] Glossary of statistical terms, https://stats.oecd.org/glossary/detail.asp?ID=3464.

In sample (data) — Data used during the development phase of the ML model. This data mainly consists of the training, validation and test data sets.

Machine learning (ML) — The branch of AI concerned with the development of learning algorithms that allow computers to evolve behaviours based on observing data and making inferences on this data.

ML strategies include three methods:

— **Supervised learning** — The process of learning in which the learning algorithm processes the input data set, and a cost function measures the difference between the ML model output and the labelled data. The learning algorithm then adjusts the parameters to increase the accuracy of the ML model.

— **Unsupervised learning (or self-learning)** — The process of learning in which the learning algorithm processes the data set, and a cost function indicates whether the ML model has converged to a stable solution. The learning algorithm then adjusts the parameters to increase the accuracy of the ML model.

— **Reinforcement learning** — The process of learning in which the agent(s) is (are) rewarded positively or negatively based on the effect of the actions on the environment. The ML model parameters are updated from this trial-and-error sequence to optimise the outcome.

ML processes can be further characterised as:

— **Offline learning** — The process of learning where the ML model is frozen at the end of the development phase;

— **Online learning** — The process of learning where the ML model parameters can be updated based on data acquired during operation (see also adaptivity).

ML model — A parameterised function that maps inputs to outputs. The parameters are determined during the training process.

— **Trained model —** the ML model which is obtained at the end of the learning/training phase.

— **Inference model —** the ML model obtained after transformation of the trained model, so that the model is adapted to the target platform.

Multicollinearity — Multicollinearity generally occurs when there are high correlations between two or more predictor variables or candidate features.

Natural language processing (NLP) — Refers to the branch of computer science — and more specifically, the branch of AI — concerned with giving computers the ability to understand text and spoken words in much the same way as human beings can (IBM Cloud Education, 2020).

Operational domain (OD) — Operating conditions under which a given AI-based system is specifically designed to function as intended, in line with the defined ConOps, including but not limited to environmental, geographical, and/or time-of-day restrictions[54].

[54] Source: adapted from SAE J3016, Level of driving automation, 2021.

Operational design domain (ODD) — The ODD defines the set of operating parameters, together with the range and distribution within which the AI/ML constituent is designed to operate, and as such, will only operate nominally when the parameters described within the ODD are satisfied. The ODD also considers dependencies between operating parameters in order to refine the ranges between these parameters when appropriate; in other words, the range(s) for one or several operating parameters could depend on the value or range of another parameter.

Oracle (attack) — A type of attack in which the attacker explores a model by providing a series of carefully crafted inputs and observing outputs. These attacks can be previous steps to more harmful types, evasion or poisoning for example. It is as if the attacker made the model talk to then better compromise it or to obtain information about it (e.g. model extraction) or its training data (e.g. membership inferences attacks and inversion attacks). Example: an attacker studies the set of input-output pairs and uses the results to retrieve training data[55].

Outlier — Data which is outside the range of at least one AI/ML constituent ODD parameter.

Out of distribution (data) — Data which is sampled from a different distribution than the one of the training data set. Data collected at a different time, and possibly under different conditions or in a different environment, than the data collected to create the ML model are likely to be out of distribution.

Out of sample (data) — Data which is unseen during the development phase, and that is processed by the ML model during inference in operation.

Overreliance — is the state when the end user is excessively relying on, depending on or trusting in the AI-based system[56].

Poisoning (attack) — A type of attack in which the attacker altered data or the model to modify the learning algorithm's behaviour in a chosen direction (e.g. to sabotage its results, to insert a backdoor). It is as if the attacker conditioned the learning algorithm according to its motivations. Such attacks are also called causative attacks (ENISA, December 2021). Example: massively indicating to an image recognition algorithm that images of helicopters are indeed aircraft to lead it to interpret them this way[57].

Predictability — The degree to which a correct forecast of a system's state can be made quantitatively. Limitations on predictability could be caused by factors such as a lack of information or excessive complexity.

Redress by design — Redress by design relates to the idea of establishing, from the design phase, mechanisms to ensure redundancy, alternative systems, alternative procedures, etc. in order to be able to effectively detect, audit, rectify the wrong decisions taken by a perfectly functioning system and, if possible, improve the system[58].

---

[55] Source: adapted from (ENISA, December 2021)

[56] Source: adapted from Merriam-Webster Inc.

[57] Source: adapted from (ENISA, December 2021).

[58] Source: adapted from (EU High-Level Expert Group on AI, 2020).

Reliability — The probability that an item will perform a required function under specified conditions, without failure, for a specified period of time[59].

Reliance — Is the state of the end user when choosing to depend on or to trust in the AI-based system; this does not prevent the end user from exercising oversight[60].

Representativeness (of a data set) — A data set is representative when the distribution of its key characteristics is similar to the actual input state space for the intended application.

Residual risk — Risk remaining after protective measures have been taken[61]. In the context of this guidance, residual risk designates the amount of risk remaining due to a partial coverage of some objectives. Indeed, it may not be possible in some cases to fully cover the learning assurance building block objectives or the explainability block objectives. In such cases, the applicant should design its AI/ML system to first minimise the residual risk and then mitigate the remaining risk using the SRM concept defined in this guidance.

Resilience — The ability of a system to continue to operate while an error or a fault has occurred (DEEL Certification Workgroup, 2021).

Robustness — Ability of a system to maintain its level of performance under all foreseeable conditions. At model level (trained or inference), the robustness objectives are further split into two groups: the ones pertaining to 'model stability' and the ones pertaining to 'robustness in adverse conditions'.

Safety criteria — This term is specific to the ATM/ANS domain and is defined in point ATS.OR.210 of Regulation (EU) 2017/373. This Regulation does not have the notion of safety objective for non-ATS providers; it instead uses the notion of safety criteria. Although the two notions are not fully identical, they are used in an equivalent manner in this document.

Safety objective — A qualitative and/or quantitative attribute necessary to achieve the required level of safety for the identified failure condition, depending on its classification.

Safety requirement — A requirement that is necessary to achieve either a safety objective or satisfy a constraint established by the safety process.

This term is used in various domains with domain-specific definitions. For the ATM/ANS domain, according to GM1 to AMC2 ATS.OR.205(a)(2), safety requirements are design characteristics/items of the functional system to ensure that the system operates as specified.

Safety science — A broad field that refers to the collective processes, theories, concepts, tools and technologies that support safety management.

Safety support requirement — Safety support requirements are characteristics/items of the functional system to ensure that the system operates as specified. This term is used in the ATM/ANS domain for non-ATS providers and is defined in GM1 to AMC2 ATM/ANS.OR.C.005(a)(2).

59 Source: ARP 4761 Guidelines and Methods for Conducting the Safety Assessment Process on Civil Airborne Systems and Equipment, 1996.

60 Source: Cambridge Dictionary.

61 Source: IEV ref 903-01-11 — https://std.iec.ch/iev/iev.nsf/ID_xref/en:903-01-11.

Stability of the learning algorithm — Refers to ensuring that the produced model does not change a lot under perturbations of the training data set.

Stability of the model — Refers to keeping input-output relations of the model under small perturbations, i.e.:

$\|x' - x\| < \delta \Rightarrow \|\hat{f}(x') - \hat{f}(x)\| < \varepsilon$, where $x,\ x' \in X$ and $\delta,\ \varepsilon \in R_{>0}$.

Subject — A subject is a person, or a group of persons affected by the AI-based system[62].

Surrogate model (or substitute model or emulation model) — is generally a mathematical model that is used to approximate the behaviour of a complex system. In the aviation industry, surrogate models are often used to represent the performance of aircraft, propulsion systems, structural dynamics, flight dynamics, and other complex systems. They can be particularly useful when it is not practical or cost-effective to use physical models or prototypes for testing or evaluation.

Synthetic data —Data that is generated by computer simulation or algorithm as an alternative to real-world data.

System — A combination of inter-related items arranged to perform a specific function(s) [ED-79A/ARP4754A]

Traceability — The ability to track the journey of a data input through all stages of sampling, labelling, processing and decision-making[63].

Training — The process of optimising the parameters (weights) of an ML model given a data set and a task to achieve on that data set. For example, in supervised learning the training data consists of input (e.g. an image) / output (e.g. a class label) pairs and the ML model 'learns' the function that maps the input to the output, by optimising its internal parameters. See also the related definition of **Inference**.

Unmanned aircraft system (UAS) — An unmanned aircraft and the equipment to control it remotely.

User — A user is a person that supports or maintains the product, such as system administrators, database administrators, information technology experts, software professionals and computer technicians[64].

Variance — An error from sensitivity to small fluctuations in the training set. High variance can cause a learning algorithm to model the random noise in the training data, rather than the intended outputs (=overfitting).

[62] Source: adapted from (EU High-Level Expert Group on AI, 2020).
[63] Source: adapted from (EU High-Level Expert Group on AI, 2020).
[64] Source: adapted from (EU High-Level Expert Group on AI, 2020).

2. Acronyms

AI	artificial intelligence
AL	assurance level
ALTAI	Assessment List for Trustworthy AI
ALS	airworthiness limitation section
AMAN	arrival manager
AMC	acceptable means of compliance
AMO	approved maintenance organisation
ANN	artificial neural network
ANS	air navigation services
ANSP	air navigation service provider
ATC	air traffic control service
ATFCM	air traffic flow and capacity management
ATCO	air traffic controller
ATM	air traffic management
ATO	approved training organisation
ATS	air traffic service
CAMO	continuing airworthiness management organisation
CBT	computer-based training
CDM	collaborative decision-making
CHG	change message
CMRs	certification maintenance requirements
CNN	convolutional neural network
CNS	communication navigation and surveillance systems
ConOps	concept of operations
CRI	certification review item
CS	certification specification
D4S	Data for Safety
DAL	development assurance level

DBS	distance-based separation
DevOps	development and operations
DF	deceleration fix
DL	deep learning
DLA	delay(ed) message
DNN	deep neural network
DOA	design organisation approval
DPIA	data protection impact assessment
DPO	Data Protection Officer
DQRs	data quality requirements
EASA	European Union Aviation Safety Agency
ENISA	European Union Agency for Cybersecurity
EOBT	estimated off-block time
EU	European Union
EUROCAE	European Organisation for Civil Aviation Equipment
FL	flight level
FPL	flight plan
FMP	flow management position
FSTD	flight simulation training device
FTD	final target distance
GDPR	General Data Protection Regulation
GM	guidance material
GPU	graphics processing unit
HAI	human-AI
HAII	human-AI interaction
HAIRM	human-AI resource management
HAT	human-AI teaming
HIC	human-in-command
HITL	human-in-the loop

HLEG	AI High-Level Expert Group
HMI	human-machine interface
HOTL	human-on-the-loop
HOOTL	human-out-of-the-loop
ICA	instructions for continuing airworthiness
ICAO	International Civil Aviation Organization
IoU	intersection over union
IDAL	item development assurance level
IFPS	initial flight plan processing system
IR	implementing rule
ISM	independent system monitoring
ISMS	information security management system
ITD	initial target distance
IUEI	intentional unauthorised electronic interaction
JAA	Joint Aviation Authorities
LAS	learning accomplishment summary
LOAT	level of automation
MCP	multicore processor
ML	machine learning
MLEAP	machine learning application approval
MLOps	machine learning operations
MOA	maintenance organisation approval
MOC	means of compliance
NDT	non-destructive testing
NLP	natural language processing
NN	neural network
ODD	operational design domain
OoD	out of distribution
ORD	optimum runway delivery

PAR	place and route
PISRA	product information security risk assessment
PLAC	plan for learning assurance
RNN	recurrent neural network
RPAS	remotely piloted aircraft system
RSUP	room supervisor
RTCA	Radio Technical Commission for Aeronautics
SA	situational awareness
SAL	security assurance level
SLT	statistical learning theory
SMS	safety management system
SPO	single-pilot operation
SRM	safety risk mitigation
SAE	Society of Automotive Engineering
SWAL	software assurance level
TBS	time-based separation
UAS	unmanned aircraft system
VTOL	vertical take-off and landing
WG	working group

H. Annex 4 — References

Daedalean. 2020. *Concepts of Design Assurance for Neural Networks (CoDANN).* Cologne : EASA, 2020.

—. 2021. *Concepts of Design Assurance for Neural networks (CoDANN) II.* Cologne : EASA, 2021.

DEEL Certification Workgroup. 2021. *White Paper - Machine Learning in Certified Systems.* Toulouse : IRT StExupery, 2021.

Ditlevsen, Armen Der Kiureghian and Ove. 2009. Aleatory or epistemic? Does it matter? 2009, Vol. 31, 2.

ECATA Group. 2019. *ECATA Technical Report 2019 - The exploitation of Artificial Intelligence in future Aircraft Systems.* 2019.

Enhancing the reliability of out-of-distribution image detection in neural networks. **Liang-et-al, Shiyu. 2018.** Vancouver : s.n., 2018. ICLR 2018.

ENISA. December 2021. *SECURING MACHINE LEARNING ALGORITHMS.* s.l. : (accessible at https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms), December 2021.

ER-022 - EUROCAE. 2021. *Artificial Intelligence in aeronautical systems: Statement of concern.* s.l. : EUROCAE, 2021. ER-022.

EU Commission. 2018. *Communication AI for Europe.* 2018.

—. 2021. *EU Commission - Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM/2021/206 final.* 2021.

EU High Level Expert Group on AI. 2020. *Assessment List for Trustworthy AI (ALTAI).* s.l. : European Commission, 2020.

—. 2019. *Ethics Guidelines for Trustworthy AI.* s.l. : European Commission, 2019.

EU High-Level Expert Group on AI. 2020. *Assessment List for Trustworthy AI (ALTAI).* s.l. : European Commission, 2020.

—. 2019. *Ethics Guidelines for Trustworthy AI.* s.l. : European Commission, 2019.

EUROCONTROL. 2020. *ATFCM Users Manual, Edition: 24.0 - Validity Date: 23/06/2020.* s.l. : (accessible at: https://www.eurocontrol.int/publication/atfcm-users-manual), 2020.

—. 2021. *Calibration of Optimised Approach Spacing Tool; ED V1.1; Date: 16/04/2021.* s.l. : (accessible at: https://www.eurocontrol.int/publication/eurocontrol-coast-calibration-optimised-approach-spacing-tool-use-machine-learning), 2021.

—. 2020. *IFPS Users Manual, Edition: 24.1 - Validity Date: 01/12/2020.* s.l. : (accessible at: https://www.eurocontrol.int/publication/ifps-users-manual), 2020.

Federal Aviation Administration. May 2022. *Neural Network Based Runway Landing Guidance for General Aviation Autoland.* s.l. : (accessible at: http://www.tc.faa.gov/its/worldpac/techrpt/tc21-48.pdf), May 2022.

Function Allocation Considerations in the Era of Human Autonomy Teaming. **Emilie M. Roth, Christen Sushereba, Laura G. Militello, Julie Diiulio, Katie Ernst. December 2019.** 4 page(s): 199-220, s.l. : Journal of Cognitive Engineering and Decision Making , December 2019, Vol. 12.

Goodfellow-et-al. 2016. *Deep Learning.* s.l. : MIT Press, 2016.

IBM Cloud Education. 2020. Natural Language Processing (NLP). [Online] IBM, 2 July 2020. https://www.ibm.com/cloud/learn/natural-language-processing.

Javier Nuñez et al. 2019. *Harvis D1.1 State of the Art Review.* s.l. : Clean Sky 2 JU, 2019.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. 2018. Generalization in Deep Learning. *Mathematics of Deep Learning.* s.l. : Cambridge Unitveristy Press, 2018, p. Proposition 5.

Liu, Qiang & Li, Pan & Zhao, Wentao & Cai, Wei & Yu, Shui. 2018. *A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View.* s.l. : IEEE Access. 6. 12103-12117. 10.1109/ACCESS.2018.2805680, 2018.

On the Uniform Convergence of Relative Frequencies of Events to their Probabilities. **Chervonenkis, V.N. Vapnik and A.Ya. 1971.** 2, 1971, Theory of Probability and its Applications, Vol. 16, pp. 264-280.

Parasuraman-et-al, Raja. 2000. A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, Vol 30, No. 3.* May 2000, pp. 286-297.

SKYbrary. Crew resource Management. [Online] https://www.skybrary.aero/articles/crew-resource-management-crm.

Stronger generalization bounds for deep nets via a compression approach. **Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. 2018.** s.l. : Andreas Krause and Jennifer Dy. International Machine Learning Society (IMLS), 2018. 35th International Conference on Machine Learning (ICML). Vol. 35th International Conference on Machine Learning (ICML), pp. pp. 390–418.

Thales. September 2022. *ODDandHighLevelProperties_V0.8.* September 2022.

I. Annex 5 — Full list of questions from the ALTAI adapted to aviation

The following questions in this annex are taken from the document of the EU Commission published in 2020 - High Level Expert Group on AI 'Assessment List for Trustworthy AI (ALTAI)' and have been partially adapted and aligned for usage in this guideline document. The tables below contain in the first column the ALTAI question, which, if modified is marked by using italic font. The second column provides a link to AI trustworthiness objectives, including rationale and record of identified challenges.

In the aviation domain and in particular in the present document, the term 'subjects' refers to the general public. Safety of the general public is ensured through the compliance of the aviation system with EU regulations, and in particular for safety-related AI applications through the future compliance with the concept paper guidelines by the applicants. Thus, the term 'subjects' has intentionally not been kept in the ALTAI items, in order to focus the ethics-based assessment on the potential impact on the safety of 'users' or 'end users', which in turn ensures the safety of the general public.

1. Gear #1 — Human agency and oversight

Quote from the ALTAI: 'This subsection deals with the effect AI systems can have on human behaviour in the broadest sense. It deals with the effect of AI systems that are aimed at guiding, influencing or supporting humans in decision making processes, for example, algorithmic decision support systems, risk analysis/prediction systems (recommender systems, predictive policing, financial risk analysis, etc.). It also deals with the effect on human perception and expectation when confronted with AI systems that 'act' like humans. Finally, it deals with the effect of AI systems on human affection, trust and (in)dependence. […] This subsection helps to self-assess necessary oversight measures through governance mechanisms.'

Human agency in aviation applications

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
G1.a. Is the AI-based system designed to interact with, guide or take decisions ~~by~~ *for* end users, that could affect humans or society?	Objective: Provision ORG-07. Rationale: Rely on licensing/training to share the pertinent information about the AI-based system. Slightly reworded for clarity. Impact on society at large is considered to be managed through the existing aviation system and regulations.
i. Could the AI-based system generate confusion for some or all end users ~~and/or subjects~~ on whether a decision, content, advice or outcome is the result of an algorithmic decision?	Objective: EXP-05 to EXP-11, Provision ORG-07. Rationale: The operational explainability guidance addresses the objectiveness of every output of the AI-based system that is relevant to the operations. Rely on licensing/training to share the pertinent information about the AI-based system.
ii. Are end users ~~and/or other subjects~~ adequately made aware that a decision, content, advice or outcome is the result	Objective: EXP-05 to EXP-11, Provision ORG-07. Rationale: The operational explainability guidance addresses the objectiveness of every

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
of an algorithmic decision?	output of the AI-based system that is relevant to the operations. Rely on licensing/training to share the pertinent information about the AI-based system.
G1.b. Could the AI-*based* system generate confusion for some or all end users ~~or subjects~~ on whether they are interacting with a human or AI-*based* system?	Objective: Provision ORG-07. Rationale: Rely on licensing/training to share the pertinent information about the AI-based system.
i. Are end users ~~or subjects~~ informed that they are interacting with an AI-*based* system?	Objective: See item G1.b. Rationale: See item G1.b.
G1.c. Could the AI-based system affect human autonomy by generating over-reliance by end users?	Objective: ORG-04, Provision ORG-07. Rationale: Overreliance is a safety risk which may occur in operations and needs to be monitored through continuous safety assessment (ORG-04) and prevented by effective training activities (Provision ORG-07) with the end users.
i. Did you put in place procedures to avoid that end users over-rely on the AI-*based* system?	Objective: See item G1.c. Rationale: See item G1.c.
G1.d. Could the AI-*based* system affect human autonomy by interfering with the end user's decision-making process in any other unintended and undesirable way?	Objective: ORG-01, EXP-05 to EXP-11 Rationale: The organisation should put in place adequate processes and procedures linked with the introduction of the AI-based systems. The end user should get enough and precise explainability about the AI-based system's output to make an appropriate and correct decision.
i. Did you put in place any procedure to avoid that the AI-based system inadvertently affects human autonomy?	Objective: See item G1.d. Rationale: See item G1.d.
G1.e. Does the AI-*based* system simulate social interaction with or between end users ~~or subjects~~?	Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale). Rationale: Social interaction (*a process of reciprocal stimulation or response between two people*) of an AI-based system with an end user is not considered as requiring additional guidance compared to the objectives for human-AI collaboration developed in the objectives of this document.
G1.f. Does the AI-*based* system risk creating human attachment, stimulating addictive behaviour, or manipulating user	Objective: ET-02. Rationale: In the current state of technology, AI-based systems with the potential of creating

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
behaviour? Depending on which risks are possible or likely, please answer the questions below:	human attachment, stimulating addictive behaviour, or manipulating user behaviour are not considered acceptable for the aviation domain. The organisations shall adapt their processes and procedures to ensure that these risks are strictly avoided.
i. Did you take measures to deal with possible negative consequences for end users ~~or subjects~~ in case they develop a disproportionate attachment to the AI-*based* system?	Objective: See item G1.f. Rationale: See item G1.f.
ii. Did you take measures to minimise the risk of addiction?	Objective: See item G1.f. Rationale: See item G1.f.
iii. Did you take measures to mitigate the risk of manipulation?	Objective: See item G1.f. Rationale: See item G1.f.

Human oversight in aviation applications

G1.g. Please determine whether the AI-*based* system is overseen by a Human-in-the-Loop, Human-on-the-Loop, Human-in-Command, *considering the definitions below*.	Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale). Rationale: The oversight mechanisms proposed in the ALTAI are not used in the current version of the EASA concept paper, and it was not deemed necessary to provide a different set of definitions at this stage. Applicants may find necessary to answer the ALTAI item G1.g with more details and characterise the functions/tasks of the AI-based system(s) with such oversight mechanisms. In such a case, the applicant should clarify the definitions used. The sub-item 'Is a self-learning or autonomous system' is mixing unrelated concepts and is not considered relevant as part of this item (see G1.k).
G1.h. Have the humans overseeing the AI-based system (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise human oversight?	Objective: Provision ORG-07. Rationale: Rely on licensing to ensure adequate training of the end users overseeing the AI-based systems' operations.
G1.i. Did you establish any detection and response mechanisms for undesirable adverse effects of the AI-*based* system for the *end user* ~~*or subject*~~?	Objective: SA-01, ICSA-01 to ICSA-02, IS-01, EXP-14 to EXP-17, DA-01 to DA-04 Rationale: The question is answered through the safety (SA-01), continuous safety (ICSA-01 to ICSA-02) and security assessment (IS-01) and monitoring for the adherence of operational

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
	boundaries (EXP-14 to EXP-17), which result is finally fed back into the learning assurance process requirements (DA-01 to DA-04).

G1.j. Did you ensure a ~~*'stop button' or*~~ procedure to safely ~~*abort*~~ *override* an operation *by a human end-user* when needed?	Objective: SA-01, ICSA-01 to ICSA-02, IS-01, EXP-07, DA-01 to DA-04 Rationale: The override-procedure should be assessed for compliance with safety objectives (SA-01, ICSA-01 to ICSA-02) and security objective (IS-01), safeguarded by the relevant explainability (EXP-07) and specified through the learning assurance process requirements (DA-01 to DA-04). The use of a 'stop button' to 'abort' an operation is a prescriptive design choice which may not be appropriate for all systems. EASA prefers to focus on a the notion of 'safely override an operation' which is more generic and encompasses the use of a 'stop button' where appropriate.
G1.k. Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI-based system?	Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale). Rationale: The two notions of 'self-learning' and 'autonomous nature' are very distinct considerations that should not be mixed. 'Self-learning' AI/ML items refer to a particular learning technique, unsupervised learning, which is not covered in the scope of the current document and will be addressed in a subsequent version of this EASA concept paper. It is anticipated that the adaptation of the learning assurance building block to unsupervised learning techniques, as well as the development of operational explainability guidance will fully address the question of oversight and control measures for 'self-learning' applications. More autonomous systems are considered to be covered under Level 3 AI applications and will be addressed in a future revision of these guidelines.

2. Gear #2 — Technical robustness and safety

Quote from the ALTAI: 'A crucial requirement for achieving Trustworthy AI systems is their dependability (the ability to deliver services that can justifiably be trusted) and resilience (robustness when facing changes). Technical robustness requires that AI systems are developed with a preventative approach to risks and that they behave reliably and as intended while minimising unintentional and unexpected harm as well as preventing it where possible. This should also apply in the event of potential changes in their operating environment or the presence of other agents (human or artificial) that may interact with the AI system in an adversarial manner. The questions in this section address four main issues: 1) security; 2) safety; 3) accuracy; and 4) reliability, fall-back plans and reproducibility.'

Resilience to attack and security in aviation applications

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
G2.a. Could the AI-*based* system have adversarial, critical or damaging effects (e.g. to human or societal safety) in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use?	Objective: SA-01, IS-01 Rationale: The answer is 'YES' for any system falling within the scope of this EASA guidance document. The associated risk is assessed through objective SA-01 (for safety) and IS-01 (for security). The AI-based system should be assessed for security vulnerabilities with impact on safety and general safety risks.
G2.b. Is the AI-*based* system *compliant with* ~~*certified*~~ ~~*for*~~ *information* security *requirements* (e.g. the certification scheme created by the Cybersecurity Act in Europe) ~~*or is it compliant*~~ *and* with ~~*specific*~~ *the applicable* security standards?	Objective: IS-01 to IS-02 Rationale: Information security risks should be identified and a mitigation approach planned and implemented, in line with current information security risk assessment guidance and, as of 16 October 2025 with Regulation (EU) 2022/1645 (Part-IS). The ALTAI question G2.b was reformulated to reflect the EASA system.
G2.c. How exposed is the AI-*based* system to cyberattacks?	Objective: IS-01 Rationale: Information security risks of the AI-based system should be assessed for their impact on safety
i. Did you assess potential forms of attacks to which the AI-*based* system could be vulnerable?	Objective: See item G2.c. Rationale: See item G2.c.
ii. Did you consider different types of vulnerabilities and potential entry points for attacks such as: ▪ data poisoning (i.e. manipulation of training data), ▪ model evasion (i.e. classifying the data according to the attacker's will), ▪ model inversion (i.e. infer the model parameters).	Objective: IS-01, ORG-02 Rationale: The different types of threats and their risks should be identified and assessed by the organisations responsible for design, production and operation phases.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
G2.d. Did you put measures in place to ensure the integrity, robustness and overall security of the AI-*based* system against potential attacks over its life cycle?	Objective: IS-02 Rationale: The applicant is asked to implement procedures and processes to avoid or mitigate the reduction of safety levels due to information security risks of the AI-based system.
G2.e. Did you red-team/pentest the system?	Objective: [to be developed] Rationale: [to be developed]
G2.f. Did you inform end users of the duration of security coverage and updates?	Objective: AMC20-42 or Part-IS Rationale: The organisation (design or operation) should monitor the evolution of security risks as defined in e.g., AMC20-42 or Part-IS and communicate the information accordingly to e.g. a maintenance organisation.
i. What length is the expected time frame within which you provide security updates for the AI-based system?	Objective: See item G2.f. Rationale: See item G2.f.

General safety in aviation applications

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
G2.g. Did you define risks, risk metrics and risk levels of the AI-*based* system in each specific use case?	Objective: SA-01 Rationale: Risks of the AI-based system should be identified (SA-01) and assessed.
i. Did you put in place a process to continuously measure and assess risks?	Objective: ICSA-01 to ICSA-02. Rationale: A process for continuous risk monitoring, using defined metrics and levels (ICSA-01 to ICSA-02) should be implemented and the residual risk communicated to the end user through training activities.
ii. Did you inform end users ~~*and/or subjects*~~ of existing or potential risks?	Objective: Provision ORG-07. Rationale: Rely on training to communicate on the potential risks.
G2.h. Did you identify the possible threats to the AI-*based* system ~~(design faults, technical faults, environmental threats)~~ and the possible consequences?	Objective: SA-01, IS-01 Rationale: This question covers the assessment of the risk from the perspective of safety (SA-01) and security (IS-01). The text 'design faults, technical faults, environmental threats' was removed as being too specific.
i. Did you assess the risk of possible malicious use, misuse or inappropriate use of the AI-*based* system?	Objective: SA-01, IS-01 Rationale: Safety and information security assessments address malicious use, misuse and inappropriate use.

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
ii. Did you define safety-criticality levels ~~*(e.g. related to human integrity)*~~ of the possible consequences of faults or misuse of the AI-*based* system?	Objective: SA-01 Rationale: The safety assessment includes the assignment of assurance levels to the AI-based system. Example removed as safety-criticality levels are not defined for human integrity in the aviation domain.
G2.i. Did you assess the dependency of a critical AI-*based* system's decisions on its stable and reliable behaviour?	Objective: SA-01, LM-02 Rationale: The safety (support) assessment (SA-01) should, amongst others, define safety objectives on reliability metrics for the AI-based system. The learning management requirements (LM-02) should capture the stability metrics for the AI-based system constituents.
i. Did you align the ~~*reliability/testing*~~ requirements with the appropriate levels of stability and reliability?	Objective: See item G2.i. Rationale: See item G2.i.
G2.j. Did you plan fault tolerance via, e.g. a duplicated system or another parallel system (AI-based or 'conventional')?	Objective: SA-01, DA-03 Rationale: The safety assessment should account for necessary architectural mitigation strategies to meet the safety requirements.
G2.k. Did you develop a mechanism to evaluate when the AI-*based* system has been changed to merit a new review of its technical robustness and safety?	Objective: CM-01 Rationale: The change management process (CM-01) should trigger a change impact analysis, and on this basis define the need for re-performing activities to maintain safety and technical robustness.

Accuracy in aviation applications

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
G2.l. Could a low level of accuracy of the AI-*based* system result in critical, adversarial or damaging consequences?	Objective: SA-01, ICSA-02, EXP-15 Rationale: The level of performance/accuracy of the AI-based system is defined and assessed in the safety assessment (SA-01) and continuously monitored (EXP-15) and assessed (ICSA-02)
G2.m. Did you put in place measures to ensure that the data (including training data) used to develop the AI-*based* system is up to date, of high quality, complete and representative of the environment the system will be deployed in?	Objective: DM-02 to DM-11 Rationale: All data management process objectives are linked to ensure that the data used to plan, design and implement, train and operate the AI-based system is appropriate, i.e. up to date, complete, representative and verified to be compliant with the DQRs.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
G2.n. Did you put in place a series of steps to monitor, and document the AI-*based* system's accuracy?	Objective: LM-09, IMP-06, DA-07, EXP-15, ICSA-02 Rationale: During the development phase (LM-09, IMP-06) the performance of the trained and inference models should be evaluated and documented. The accuracy of the system is then verified and documented at system level (DA-07). During the operational phase (EXP-15, ICSA-02), continuous monitoring, recording and accuracy assessment should be performed and documented.
G2.o. Did you consider whether the AI-*based* system's operation can invalidate the data or assumptions it was trained on, and how this might lead to adversarial effects?	Objective: ICSA-02 Rationale: The continuous safety assessment (ICSA-02) aims at identifying invalid assumptions on the data used to train the system and on the system's operation, to prevent possible adversarial effects.
G2.p. Did you put processes in place to ensure that the level of accuracy of the AI-*based* system to be expected by end users ~~*and/or subjects*~~ is properly communicated?	Objective: EXP-15 Rationale: Relevant information concerning deviations of the AI-based system's output from the specified performance need to be indicated (EXP-15) to the end users.

Reliability, fallback plans and reproducibility in aviation applications

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
G2.q. Could the AI-*based* system cause critical, adversarial, or damaging consequences (e.g. pertaining to human safety) in case of low reliability and/or reproducibility?	Objective: DA-07, LM-07 Rationale: The answer to this question is 'Yes' for the type of systems covered by this EASA guidance document. The learning assurance process should address both the verification of intended function (DA-07) and the reproducibility of the learning process (LM-07).
i. Did you put in place a well-defined process to ~~*monitor if*~~ *verify that* the AI-based system is meeting the intended goals?	Objective: DA-07 Rationale: Objective DA-07 verifies that all system requirements are met.
ii. Did you test whether specific contexts or conditions need to be taken into account to ensure reproducibility?	Objective: LM-07 Rationale: The bias-variance trade-off should be accounted for in the model family selection in order to provide evidence of the reproducibility of the training process.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
G2.r. Did you put in place verification and validation methods and documentation (e.g. logging) to evaluate and ensure different aspects of the AI-*based* system's reliability ~~*and reproducibility*~~?	Objective: SA-01, DA-04, DA-07 Rationale: The safety assessment (SA-01) should ensure different aspects of the AI-based system's reliability in the requirements. All requirements are being validated (DA-04) and verified (DA-07). The reference to reproducibility has been removed from the ALTAI objective G2.r because reproducibility is covered in the objective G2.q.
i. Did you clearly document and operationalise processes for the testing and verification of the reliability ~~*and reproducibility*~~ of the AI-based system?	Objective: See item G2.r. Rationale: See item G2.r.
G2.s. Did you define tested fail-safe fallback plans to address AI-*based* system errors of whatever origin and put ~~*governance*~~ procedures in place to trigger them?	Objective: SA-01, SRM-01 to SRM-02 Rationale: The safety assessment (SA-01) should validate the safety architecture of the AI-based system including necessary fail-safe fallback provisions. Additionally, fail-safe fallback plans may be identified by safety risk management (SRM-01 to SRM-02) processes and adequate procedures defined. The word 'governance' is proposed to be removed to avoid limiting the scope of procedures that are meant.
G2.t. Did you put in place a proper procedure for handling the cases where the AI-*based* system yields results with a low confidence score?	Objective: EXP-15 to EXP-16 Rationale: The AI-based system output performance should be monitored (EXP-15) and procedures put in place to act on the possible output of the AI-based system's monitoring (EXP-16).
G2.u. Is your AI-*based* system using (online) continual learning?	Objective: ET-03. Rationale: Continuous/online learning is outside the scope of this guidance document; therefore, such applications will not be accepted by EASA at this stage.
i. Did you consider potential negative consequences from the AI-based system learning novel or unusual methods to score well on its objective function?	Objective: See item G2.u. Rationale: See item G2.u.

3. Gear #3 — Privacy and data governance

Quote from the ALTAI: 'Closely linked to the principle of prevention of harm is privacy, a fundamental right particularly affected by AI systems. Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.'

Privacy in aviation applications

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
G3.a. Did you consider the impact of the AI-*based* system on the right to privacy ~~*the right to physical, mental and/or moral integrity*~~ and the right to data protection?	Objective: ET-04 Rationale: The AI-based system should comply with applicable data protection requirements to protect data and preserve the privacy of data. The phrase 'the right to physical, mental and/or moral integrity' is proposed to be removed to prevent distraction from the scope of this document section 'privacy' of the use of data. The struck-through text was relocated in the MOC of G6.a to better highlight the possible effects on human health.
G3.b. Depending on the use case, did you establish mechanisms that allow flagging issues related to *data* privacy concerning the AI-*based* system?	Objective: ET-04 Rationale: See item G3.a.

Data governance in aviation applications

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
G3.c. Is your AI-*based* system being trained, or was it developed, by using or processing personal data (including special categories of personal data)?	Objective: ET-04 Rationale: See item G3.a.
G3.d. Did you put in place any of the following measures some of which are mandatory under the General Data Protection Regulation (GDPR), or a non-European equivalent?	
i. Data protection impact assessment (DPIA);	
ii. Designate a Data Protection Officer (DPO) and include them at an early state in the development,	

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
procurement or use phase of the AI-*based* system;	
iii. Oversight mechanisms for data processing (including limiting access to qualified personnel, mechanisms for logging data access and making modifications);	
iv. Measures to achieve privacy-by-design and default (e.g. encryption, pseudonymisation, aggregation, anonymisation);	
v. Data minimisation, in particular personal data (including special categories of data).	
G3.e. Did you implement the right to withdraw consent, the right to object and the right to be forgotten into the development of the AI-*based* system?	
G3.f. Did you consider the privacy and data protection implications of data collected, generated or processed over the course of the AI-*based* system's life cycle?	
G3.g. Did you consider the privacy and data protection implications of the AI-*based* system's non-personal training data or other processed non-personal data?	Objective: IS-01 to IS-02 Rationale: Non-personal data, which is processed by an AI-based system should be protected for data security by assessing the security risks and installing security controls.
G3.h. Did you align the AI-*based* system with relevant standards (e.g. ISO, IEEE) or widely adopted protocols for (daily) data management and governance?	Objective: [to be developed] Rationale: [to be developed]

4. Gear #4 — Transparency

Quote from the ALTAI: 'A crucial component of achieving Trustworthy AI is transparency which encompasses three elements: 1) traceability, 2) explainability and 3) open communication about the limitations of the AI system.'

Traceability

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
G4.a. Did you put in place measures that address the traceability of the AI-based system during its entire life cycle?	Objective: QA-01, CM-01. Rationale: During the development and change management process of the AI-based system, all configuration items should be traceable to other configuration items, from which they derive (QA-01, CM-01).
G4.b. Did you put in place measures to continuously assess the quality of the input data to the AI-*based* system?	Objective: ICSA-01, ICSA-02, EXP 04 and EXP-14 Rationale: A process for data recording (EXP-04, ICSA-01) and continuous safety assessment (ICSA-02) should be implemented and enable the capability to continuously assess the quality of the input data to the AI-based system. In addition, the ODD monitoring (EXP-14) should support analysis of the cases where the AI-based system input did not match the expected ODD.
G4.c. Can you trace back which data was used by the AI-*based* system to make a certain decision(s) or recommendation(s)?	Objective: ICSA-01, ICSA-02, EXP 04. Rationale: A process for data recording (EXP-04, ICSA-01) and continuous safety assessment (ICSA-02) should be implemented and enable the capability to trace back which data was used by the AI-based system to make a certain decision(s) or recommendation(s).
G4.d. Can you trace back which AI model ~~or rules~~ led to the decision(s) or recommendation(s) of the AI-*based* system?	Objective: ICSA-01, ICSA-02, EXP 04. Rationale: A process for data recording (EXP-04, ICSA-01) and continuous safety assessment (ICSA-02) should be implemented and enable the capability to trace back which AI model led to the decision(s) or recommendation(s) of the AI-based system. Reformulation of the question by removing the misleading phrase 'or rules'.
G4.e. Did you put in place measures to continuously assess the quality of the output(s) of the AI-*based* system?	Objective: ICSA-01, ICSA-02, EXP 04 and EXP-15. Rationale: A process for data recording (EXP-04, ICSA-01) and continuous safety assessment (ICSA-02) should be implemented and enable the capability to continuously assess the quality of the output(s) of the AI-based system. In addition, the monitoring of the system performance (EXP-15) supports the analysis of

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
	events where the AI-based system performed below the expected level of performance.
G4.f. Did you put adequate logging practices in place to record the decision(s) or recommendation(s) of the AI-*based* system?	Objective: ICSA-01, ICSA-02, EXP 04. Rationale: A process for data recording (EXP-04, ICSA-01) should be implemented.

Explainability in aviation applications

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
G4.g. Did you explain the decision(s) of the AI-*based* system to the *end users*?	Objective: EXP-05 to EXP-11 Rationale: The end user should get appropriately detailed, timely delivered explanations in a clear and unambiguous format, whose content meets operational and end users' needs.
G4.h. Do you continuously survey the *end users* if they understand the decision(s) of the AI-*based* system?	Objective: <to be developed or removed>. Rationale: There is currently no objective to cover the surveying of the end user in the operational phase.

Communication in aviation applications

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
G4.i. In cases of interactive AI-*based* systems, do you communicate to users that they are interacting with an AI-*based* system instead of a human?	Objective: Same as for G1.b Rationale: Same as for G1.b
G4.j. Did you establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI-*based* system?	Objective: EXP-01 to EXP-03, Provision ORG-06 Rationale: Identified users should be provided with explanations through training needed for the development and learning assurance processes on methods used at AI/ML item or output level (EXP-01 to EXP-03). Relying as well on training put in place by the organisations to address this topic.
~~i. *Did you communicate the benefits of the AI-based system to users?*~~	Objective: not applicable. Rationale: sub-item removed, as it is not safety-risk-related.
ii. Did you communicate the technical limitations and potential risks of the AI-*based* system to users, such as its	Objective: See item G4.j. Rationale: See item G4.j.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
level of accuracy and/ or error rates?	
iii. Did you provide appropriate training material and disclaimers to users on how to adequately use the AI-*based* system?	Objective: See item G4.j. Rationale: See item G4.j.

5. Gear #5 — Diversity, non-discrimination and fairness

Quote from the ALTAI: 'In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system's life cycle. AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness, and bad governance models. The continuation of such biases could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation. Harm can also result from the intentional exploitation of (consumer) biases or by engaging in unfair competition, such as the homogenisation of prices by means of collusion or a non-transparent market. Identifiable and discriminatory bias should be removed in the collection phase where possible. AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance.'

This gear may not be applicable to all aviation use cases. Therefore, in a first analysis, applicants should check whether the AI-based system could have any impact on diversity, non-discrimination and fairness. Diversity, non-discrimination and fairness, in the context of Gear #5, have to be interpreted as applying to people or groups of humans, not to data sources (which are addressed through the learning assurance guidance). These people are the users, meaning the ones designing, developing implementing, monitoring and/or decommissioning (involved in any other part of the life cycle of the AI-based system) plus the end users that will use directly the AI-based systems during their work practice.

If no impact exists, the record of this analysis should be added to the ethical assessment documentation, with a clear declaration of non-applicability.

In case of an impact, please consider the following questions from the ALTAI (EU High-Level Expert Group on AI, 2020) related to Gear #5.

It is understood that some of the ALTAI Gear #5 questions should be analysed through the perspective of the organisations (enterprise, company) that develop or use the AI-based system, and not so much focused on the AI-based system itself.

Avoidance of unfair bias

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
G5.a. Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI-*based* system, both regarding the use of input data as well as for the ML model ~~*algorithm*~~ design?	Objective: ET-05, MOC DM-13-2; MOC DM-13-3; EXP-02; LM-07 and LM-08. Rationale: The avoidance of potential unfair bias is addressed through the systematic mitigation of any potential biases in all phases of the AI-based system development and operations. All objectives mentioned above contribute to this goal: • Learning assurance aims at detecting potential biases in the data, through data representativeness (MOC DM 13-2) and data accuracy and correctness (MOC DM 13-3). • Objectives LM-07 and LM-08 contribute to ensuring that biases have been detected and mitigated in the trained model as a result from the learning process. • The development explainability objectives (driven by EXP-02) support detection of bias that may not have been detected in previous W-shaped process steps. The Continuous Safety Assessment (ICSA-02) aims at identifying bias or poor performance in the systems operation.
G5.b. Did you consider diversity and representativeness of end users ~~and/or subjects~~ in the data?	Objective: Anticipated MOC DM-02 Rationale: The guidance on data representativeness of the data sets covers the diversity of end users, when they are included in the ConOps (objective CO-01).
i. Did you test for specific target groups or problematic use cases?	Objective: Objective LM-13 and IMP-09 Rationale: robustness on adverse cases covering specific target groups and problematic use cases
~~ii. Did you research and use publicly available technical tools, that are state-of-the-art, to improve your understanding of the data, model and performance?~~	Objective: Not applicable. Rationale: Proposed to remove the ALTAI question as enforcement of a selection in publicly available and state-of-the-art tools is considered as too prescriptive for the aviation domain. Tools are selected by the applicants and are managed through the learning assurance process.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
iii. Did you assess and put in place processes to test and monitor for potential bias during the entire life cycle of the AI-*based* system (e.g. bias due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness)?	Objective: MOC DM-13-2; MOC DM-13-3; EXP-02; LM-07 and LM-08, ICSA-02 Rationale: The avoidance of potential unfair bias is addressed through the systematic mitigation of any potential biases in all phases of the AI-based system development and operations is addressed through the systematic mitigation of any potential biases in all phases of the AI-based system development and operations. All objectives mentioned above contribute to this goal: • Learning assurance aims at detecting potential biases in the data, through data representativeness (MOC DM 13-2) and data accuracy and correctness (MOC DM 13-3). • Objectives LM-07 and LM-08 contribute to ensuring that biases have been detected and mitigated in the trained model as a result from the learning process. • The development explainability objectives (driven by EXP-02) support detection of bias that may not have been detected in previous W-shaped process steps. The Continuous Safety Assessment (ICSA-02) aims at identifying bias or poor performance in the systems operation. Learning assurance in particular related to ensure that data representativeness (MOC DM 13-2) and data accuracy and correctness (MOC DM 13-3) are not bias impacted. Also MOC EXP-02 ensuring the goals and in particular LM-07 and LM-08 ensuring that the learning management are not biased. The continuous safety assessment (ICSA-02) aims at identifying invalid assumptions on the data used to train the system and on the system's operation, to prevent possible lack of diversity and/or non-representativeness.
iv. Where relevant, did you consider diversity and representativeness of end users ~~and/or subjects~~ in the data?	Objective: See item G5.b. Rationale: See item G5.b.
G5.c. Did you put in place educational and awareness initiatives to help AI designers	Objective: Provision ORG-06. Rationale: The organisation should put in place

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
and AI developers be more aware of the possible bias they can inject in designing and developing the AI-*based* system?	training initiatives that would support the development of bias awareness and other AI-specific competencies for the users.
G5.d. Did you ensure a mechanism that allows for the flagging of issues related to bias, ~~*discrimination*~~ or poor performance of the AI-*based* system *that may cause discrimination?*	Objective: MOC DM-13-2; MOC DM-13-3; EXP-02 and EXP-15; LM-07 and LM-08, ICSA-02 Rationale: In this item, the word 'discrimination' has been shifted to the end of the sentence, to present it as a consequence of the issues related to bias or poor performance. The mitigation of potential for discrimination is addressed through the monitoring of system performance (EXP-15) and through the systematic mitigation of any potential biases in all phases of the AI-based system development and operations. All other objectives mentioned above contribute to the latter goal: • Learning assurance aims at detecting potential biases in the data, through data representativeness (MOC DM 13-2) and data accuracy and correctness (MOC DM 13-3). • Objectives LM-07 and LM-08 contribute to ensuring that biases have been detected and mitigated in the trained model as a result from the learning process. • The development explainability objectives (driven by EXP-02) support detection of bias that may not have been detected in previous W-shaped process steps. The Continuous Safety Assessment (ICSA-02) aims at identifying bias or poor performance in the systems operation.
i. Did you establish clear steps and ways of communicating on how and to whom such issues can be raised?	Objective: ORG-03 Rationale: The data-driven AI continuous safety assessment system ensures that steps and ways of communicating detected issues to the applicant are put in place.
ii. Did you identify the subjects that could potentially be (in)directly affected by the AI-*based* system, in addition to the (end) users and/or subjects?	Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale). Rationale: In the aviation domain and in particular in the present document, the term 'subjects' refers to the general public. Safety of the general public is ensured through the

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
	compliance of the aviation system with EU regulations, and in particular for safety-related AI applications through the future compliance with the concept paper guidelines by the applicants. This explains why this item is not required to be addressed by EASA applicants..
G5.e. Is your definition of fairness commonly used and implemented in any phase of the process of setting up the AI-*based* system?	Objective: MOC DM-13-2; MOC DM-13-3; EXP-02; LM-07 and LM-08, ICSA-02. Rationale: The applicable definition of fairness is defined in the glossary of the present document. Regarding the mitigation of potential unfairness, the removal of potential for discrimination is addressed through the systematic mitigation of any potential biases in all phases of the AI-based system development and operations. All objectives mentioned above contribute to this goal: • Learning assurance aims at detecting potential biases in the data, through data representativeness (MOC DM 13-2) and data accuracy and correctness (MOC DM 13-3). • Objectives LM-07 and LM-08 contribute to ensuring that biases have been detected and mitigated in the trained model as a result from the learning process. • The development explainability objectives (driven by EXP-02) support detection of bias that may not have been detected in previous W-shaped process steps. The Continuous Safety Assessment (ICSA-02) aims at identifying bias or poor performance in the systems operation.
i. Did you consider other definitions of fairness before choosing this one?	Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale). Rationale: Several sources of information defining the concept of fairness were consulted including the ALTAI. The definition of fairness in this document is based on the EU non-discrimination guidelines.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
ii. Did you consult with the impacted communities about the correct definition of fairness, i.e. representatives of elderly persons or persons with disabilities?	Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale). Rationale: Having the definition based on the EU non-discrimination guidelines, all the impacted communities are understood as considered.
iii. Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness?	Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale). Rationale: If the EASA applicant uses the definition based on the EU non-discrimination guidelines, it is considered that there is no need for quantitative analysis or metrics to go under testing.
iv. Did you establish mechanisms to ensure fairness in your AI-*based* system?	Objective: MOC DM-13-2; MOC DM-13-3; EXP-02; LM-07 and LM-08, ICSA-02 Rationale: The mitigation of potential unfairness is addressed through the systematic mitigation of any potential biases in all phases of the AI-based system development and operations. The objectives mentioned above all contribute to this goal: • Learning assurance aims at detecting potential biases in the data, through data representativeness (MOC DM 13-2) and data accuracy and correctness (MOC DM 13-3). • Objectives LM-07 and LM-08 contribute to ensuring that biases have been detected and mitigated in the trained model as a result from the learning process. • The development explainability objectives (driven by EXP-02) support detection of bias that may not have been detected in previous W-shaped process steps. The Continuous Safety Assessment (ICSA-02) aims at identifying bias or poor performance in the systems operation.

Accessibility and universal design

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
G5.f. Did you ensure that the AI-*based* system corresponds to the variety of preferences and abilities in society?	Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale). Rationale: Aviation products are typically designed for end users with specific licensing and skills. It is therefore expected that the questions related to 'Accessibility and universal design' do not impose additional requirements to the applicant.
G5.g. Did you assess whether the AI-*based* system's user interface is usable by those with special needs or disabilities or those at risk of exclusion?	Objective: See item G5.f Rationale: See item G5.f
i. *Did you ensure that information about the AI-based system is also accessible to users of assistive technologies (such as screen readers)*?	Objective: See item G5.f Rationale: See item G5.f
ii. *Did you ensure that the user interface of the AI-based system is also usable by users of assistive technologies (such as screen readers)*?	Objective: See item G5.f Rationale: See item G5.f
iii. Did you involve or consult with end users ~~*and/or subjects*~~ in need for assistive technology during the planning and development phase of the AI-*based* system?	Objective: See item G5.f Rationale: See item G5.f
G5.h. Did you ensure that universal design principles are taken into account during every step of the planning and development process, if applicable?	Objective: See item G5.f Rationale: See item G5.f
G5.i. Did you take the impact of the AI-*based* system on the potential end users ~~*and/or subjects*~~ *into account?*	Objective: CO-05. Rationale: Through objective CO-05, the applicant should take the impact on identified end users into account by involving representative members of end users in the development lifecycle of the AI-based system.
i. Did you assess whether the team involved in building the AI-*based* system engaged with the possible target end users ~~*and/or subjects*~~?	Objective: CO-05. Rationale: The consultation of end users is a common practice/requirement in aviation when developing, certifying and approving any system, covered in this document in the objective CO-05.

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
ii. Did you assess whether there could be groups who might be disproportionately affected by the outcomes of the AI-*based* system?	Objective: SA-01, ICSA-02 Rationale: The safety assessment (SA-01) should ensure by design that no disproportionate effect is to be anticipated. Should any be identified in operations, occurrence reporting as well as the continuous safety assessment strategy developed in this document (ICSA-02) support removal of any potential remaining disproportionate effect.
iii. Did you assess the risk of the possible unfairness of the system onto the end users' ~~and/or subjects'~~ communities?	Objective: MOC DM-13-2; MOC DM-13-3; EXP-02; LM-07 and LM-08, ICSA-02 Rationale: Regarding the mitigation of potential unfairness, it is addressed through the systematic mitigation of any potential biases in all phases of the AI-based system development and operations. The objectives mentioned above all contribute to this goal: • Learning assurance aims at detecting potential biases in the data, through data representativeness (MOC DM 13-2) and data accuracy and correctness (MOC DM 13-3). • Objectives LM-07 and LM-08 contribute to ensuring that biases have been detected and mitigated in the trained model as a result from the learning process. • The development explainability objectives (driven by EXP-02) support detection of bias that may not have been detected in previous W-shaped process steps. The Continuous Safety Assessment (ICSA-02) aims at identifying bias or poor performance in the systems operation.

Stakeholder participation

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
G5.j. Did you consider a mechanism to include the participation of the widest range of possible stakeholders in the *AI-based* system's design and development?	Objective: ORG-01 Rationale: The deployment of organisation processes adapted to the introduction of AI should account for the participation and level of involvement of stakeholders such as but not limited to: data scientists, software experts, system architects, safety experts, operational experts, UX/UI experts, management decision-makers, inside and outside the organisation. Also, the linkage in particular with academic organisations, innovation research centres, and national and European authorities for aviation regulation should be accounted for.

6. Gear #6 — Societal and environmental well-being

Environmental well-being

Quote from the ALTAI: 'This subsection helps to self-assess the (potential) positive and negative impacts of the AI system on the environment. AI systems, even if they promise to help tackle some of the most pressing societal concerns, e.g. climate change, must work in the most environmentally friendly way possible. The AI system's development, deployment and use process, as well as its entire supply chain, should be assessed in this regard (e.g. via a critical examination of the resource usage and energy consumption during training, opting for less net negative choices). Measures to secure the environmental friendliness of an AI system's entire supply chain should be encouraged.'

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
G6.a. Did you identify and assess potential negative impacts of the AI-based system on the environment *and on human health throughout its life cycle (development, deployment, use, end of life)?.*	Objective: ET-06 Rationale: This ALTAI question has been reworked: — to imply that the negative impact analysis should be driven by an identification and assessment step (this has the effect of merging the sub-item that was under this ALTAI question with the main question); — so that the impact assessment also takes into account the consequences on human health, including the right to physical, mental and moral integrity; and — so that the analysis covers all of the phases of the life cycle of a product.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
G6.b. Did you establish mechanisms to evaluate the environmental impact of the AI-*based* system's development, deployment and/or use (for example, the amount of energy used and carbon emissions)?	Objective: ET-06. Rationale: This item is covered by the MOC ET-06.
G6.c. Did you define measures to reduce *or mitigate these impacts*?	Objective: ET-07. Rationale: The mitigation of identified impacts is a key objective.

Work and skills, and impact on society at large or democracy

Quote from ALTAI: 'AI systems may fundamentally alter the work sphere. They should support humans in the working environment, and aim for the creation of meaningful work. This subsection helps self-assess the impact of the AI system and its use in a working environment on workers, the relationship between workers and employers, and on skills. This subsection [i.e. regarding society at large or Democracy] helps to self-assess the impact of an AI system from a societal perspective, taking into account its effect on institutions, democracy and society at large. The use of AI systems should be given careful consideration, particularly in situations relating to the democratic processes, including not only political decision-making but also electoral contexts (e.g. when AI systems amplify fake news, segregate the electorate, facilitate totalitarian behaviour, etc.).'

*Except for topics related to **Objective ET-08 and Objective ET-09**, this sub-gear may not be applicable to all aviation use cases. Therefore, in a first analysis, applicants should check whether the AI-based system could have any impact on work and skills.*

*If no impact exists, the record of this analysis should be added to the ethical assessment documentation (per **Objective ET-01**).*

In case of an impact, please consider the questions from the ALTAI related to Gear #6 'Work and skills' and 'Impact on society at large or democracy'. Those questions can be found in the table below.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
G6.d. Does the AI-*based* system impact human work and work arrangements?	Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale). Rationale: In case of an impact, the assessment of the answer to these questions does not fall under the remit of EASA and would be performed by a competent authority for 'Work and skills'' matters, at European level or at national level as applicable. .
G6.e. Did you pave the way for the introduction of the AI-*based* system in your organisation by informing and consulting	Objective: [to be developed]. Rationale: [to be developed].

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
with impacted workers and their representatives (trade unions, (European) work councils) in advance?	
G6.f. Did you adopt measures to ensure that the impacts of the AI-*based* system on human work are well-understood?	Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale). Rationale: In case of an impact, the assessment of the answer to these questions does not fall under the remit of EASA and would be performed by a competent authority for 'Work and skills, and impact on society at large or democracy' matters, at European level or at national level as applicable.
i. Did you ensure that workers understand how the AI-*based* system operates, which capabilities it has and which it does not have?	Objective: See item G6.f. Rationale: See item G6.f.
G6.g. Could the AI-*based* system create the risk of de-skilling of the workforce?	Objective: ET-09. Rationale: When introducing new working practices, there is a risk of de-skilling meaning that the staff will no longer make use of their competence, they will no longer be ready for performance, or not accurate in terms of timing and effectiveness.
i. Did you take measures to counteract de-skilling risks?	Objective: ET-09. Rationale: This risk should be mitigated though refresher training.
G6.h. Does the system promote or require new (digital) skills?	Objective: ET-08 Rationale: As any innovation, new skills would most probably be a need. Competence building will be ensured through the provision of training objectives identified through objective ET-08.
i. Did you provide training opportunities and materials for re- and up-skilling?	Objective: See item G6.h. Rationale: See item G6.h.
G6.i. Could the AI-*based* system have a negative impact on society at large or democracy?	Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale). Rationale: In case of an impact, the assessment of the answer to these questions does not fall under the remit of EASA and would be performed by a competent authority, at European level or at national level as applicable.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
i. Did you assess the societal impact of the AI-*based* system's use beyond the (end) user ~~and/or subject~~, such as potentially indirectly affected stakeholders or society at large?	Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale). Rationale: In case of an impact, the assessment of the answer to these questions does not fall under the remit of EASA and would be performed by a competent authority, at European level or at national level as applicable.
ii. Did you take action to minimise potential societal harm of the AI-*based* system?	Objective: Not addressed through the objectives of this Concept Paper please consider the rationale). Rationale: In case of an impact, the assessment of the answer to these questions does not fall under the remit of EASA and would be performed by a competent authority, at European level or at national level as applicable.
iii. Did you take measures that ensure that the AI-*based* system does not negatively impact democracy?	Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale). Rationale: In case of an impact, the assessment of the answer to these questions does not fall under the remit of EASA and would be performed by a competent authority, at European level or at national level as applicable.

7. Gear #7 — Accountability

Quote from the ALTAI: 'The principle of accountability necessitates that mechanisms be put in place to ensure responsibility for the development, deployment and/or use of AI systems. This topic is closely related to risk management, identifying and mitigating risks in a transparent way that can be explained to and audited by third parties. When unjust or adverse impacts occur, accessible mechanisms for accountability should be in place that ensure an adequate possibility of redress.'

Auditability

Quote from the ALTAI: 'This subsection helps to self-assess the existing or necessary level that would be required for an evaluation of the AI system by internal and external auditors. The possibility to conduct evaluations as well as to access records on said evaluations can contribute to Trustworthy AI. In applications affecting fundamental rights, including safety-critical applications, AI systems should be able to be independently audited. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available.'

The AI system should be auditable by internal and external parties, including the approving authorities.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
G7.a. Did you establish mechanisms that facilitate the AI-*based* system's auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)?	Objective: DA-01, CM-01, DM-02, QA-01 Rationale: All development processes are planned (DA-01), all life cycle data managed in configuration (CM-01). In particular, sourcing of training data is performed as specified in the data management step (DM-02). The process monitoring, including negative and positive outcome, is performed through process and quality assurance (QA-01).
G7.b. Did you ensure that the AI-*based* system can be audited by independent third parties?	Objective: ORG-04. Rationale: The AI-based system should be auditable by internal and external entities.

Risk management

Some of the accountability gear items may not be applicable to all aviation use cases. Therefore, in a first analysis, applicants should check whether the AI-based system could have any impact on the monitoring of ethical concerns from an organisation's perspective.

*If no impact exists, the record of this analysis should be added to the ethical assessment documentation (per **Objective ET-01**).*

In case of an impact, please consider the following questions from the ALTAI related to Gear #7 'Accountability'.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
G7.c. Did you foresee any kind of external guidance or third-party auditing processes to oversee ethical concerns and accountability measures?	Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale). Rationale: In case of an impact, the assessment of the answer to these questions does not fall under the remit of EASA and would be performed by a competent authority, at European level or at national level as applicable.
i. Does the involvement of these third parties go beyond the development phase?	Objective: See item G7.c. Rationale: See item G7.c.
G7.d. Did you organise risk training and, if so, does this also inform about the potential legal framework applicable to the AI-*based* system?	Objective: ET-08 Rationale: The organisation's training requirements should cover the full scope of necessary skills and should also encompass the ethics-based aspects related to the legal framework, as far as safety is concerned.

ALTAI items	**Link to EASA concept paper objectives** Objective(s)/Rationale for the link
G7.e. Did you consider establishing an AI ethics review board or a similar mechanism to discuss the overall accountability and ethics practices, including potential unclear grey areas?	Objective: Provision ORG-08 Rationale: The organisation should establish means (e.g. processes) to continuously assess ethics-based aspects and in case of conflict between different ethical principles, warrant an explanation on the decision-making. This includes the consideration of establishing an AI ethics review board.
G7.f. Did you establish a process to discuss and continuously monitor and assess the AI-*based* system's adherence to the ethics-based assessment guidance?	Objective: Provision ORG-08 Rationale: The organisation should establish means (e.g. processes) to continuously assess ethics-based aspects and in case of conflict between different ethical principles, warrant an explanation on the decision-making.
i. Does this process include identification and documentation of conflicts between the six aforementioned *gears* or between different ethical principles and explanation of the 'trade-off' decisions made?	Objective: See item G7.f Rationale: See item G7.f.
ii. Did you provide appropriate training to those involved in such a process and does this also cover the legal framework applicable to the AI-*based* system?	Objective: ET-08 Rationale: The organisation's training requirements should cover the full scope of necessary skills and should also encompass the ethics-based aspects related to the legal framework, as far as safety is concerned.
G7.g. Did you establish a process for third parties (e.g. suppliers, end users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or bias in the AI-*based* system?	Objective: Provisions ORG-02 to ORG-03, Provision ORG-05 Rationale: The AI-based system should continuously be assessed for potential security (ORG-02) and safety vulnerabilities (ORG-03), bias, and the risk management process should be implemented (ORG-05).
i. Does this process foster revision of the risk management process?	Objective: See item G7.g Rationale: See item G7.g
G7.h. For applications that can adversely affect individuals, have redress-by-design mechanisms been put in place?	Objective: SRM-02, ICSA-02 Rationale: Continuously assess the safety in operation (ICSA-02) and mitigate the risk by architectural design, e.g. safety net (SRM-02).

European Union Aviation Safety Agency

Stay informed:
easa.europa.eu/ai

European Union Aviation Safety Agency

Postal address
Postfach 10 12 53
50452 Cologne
Germany

Visiting address
Konrad-Adenauer-Ufer 3
50668 Cologne
Germany

Tel. +49 221 89990-000
Fax +49 221 89990-999
Web www.easa.europa.eu