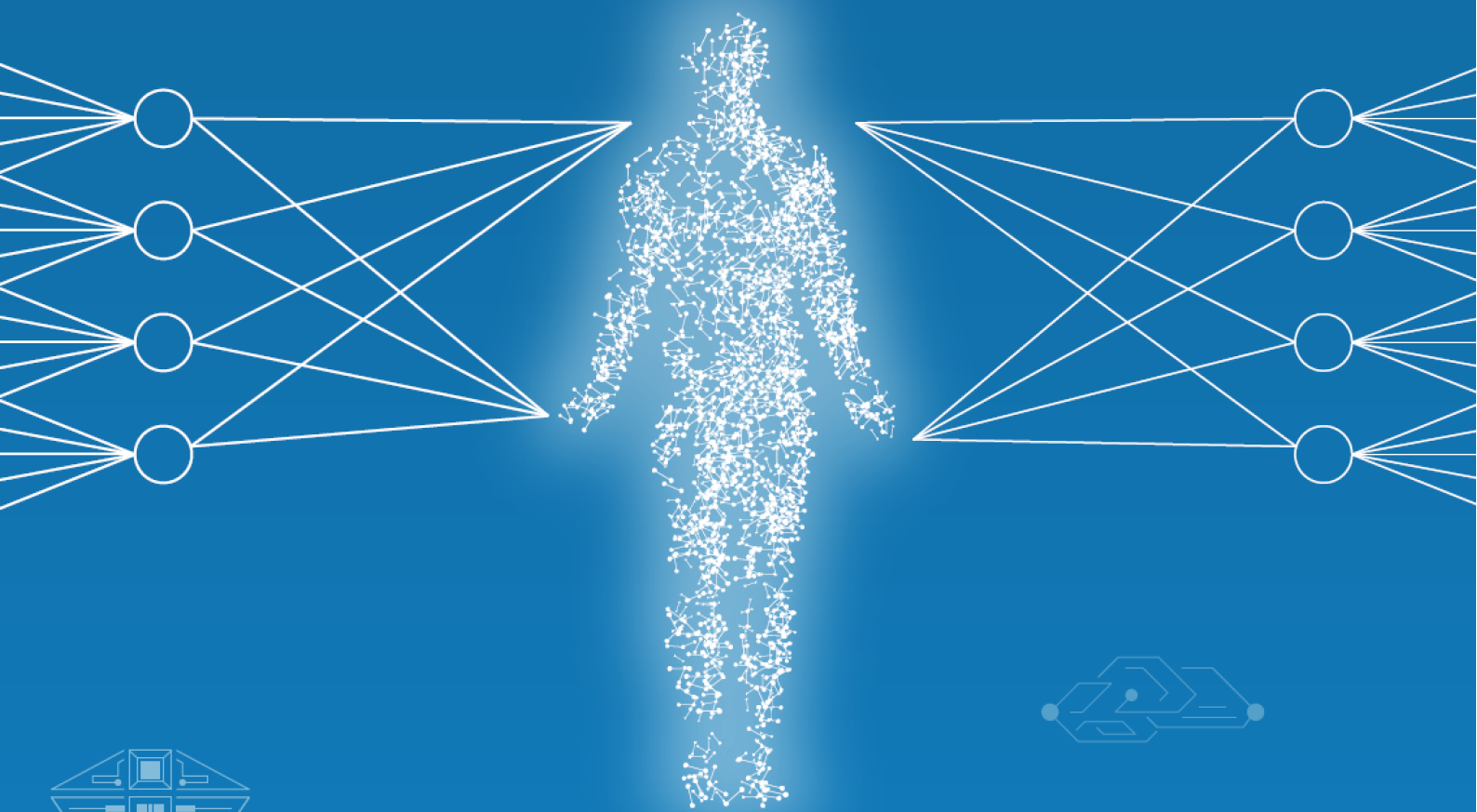




EASA Concept Paper : Guidance for Level 1 & 2 machine learning applications

A deliverable of the EASA AI Roadmap



March 2024
Issue 02

easa.europa.eu/ai

Table of Contents

A. Foreword	4
B. Introduction	6
1. Statement of issue	6
2. AI trustworthiness framework overview	8
3. Terminology and scope of the document	10
4. Criticality of AI applications	12
5. Classification of AI applications — overview	12
6. Novel concepts developed for data-driven AI	13
6.1. Learning assurance.....	13
6.2. AI explainability.....	14
6.3. Operational domain (OD) and operational design domain (ODD).....	16
6.4. Human-AI teaming	17
C. AI trustworthiness guidelines.....	20
1. Purpose and applicability.....	20
2. Trustworthiness analysis.....	22
2.1. Characterisation of the AI application	22
2.2. Safety assessment of ML applications	29
2.3. Information security considerations for ML applications	40
2.4. Ethics-based assessment.....	43
3. AI assurance	50
3.1. Learning assurance.....	50
3.2. Development & post-ops AI explainability	88
4. Human factors for AI.....	96
4.1. AI operational explainability	98
4.2. Human-AI teaming	106
4.3. Modality of interaction and style of interface	114
4.4. Error management.....	124
4.5. Failure management	129
5. AI safety risk mitigation	131
5.1. AI safety risk mitigation concept.....	131
5.2. AI safety risk mitigation top-level objectives.....	132
6. Organisations	133
6.1. High-level provisions and anticipated AMC.....	133
6.2. Competence considerations	135

6.3. Design organisation case.....	136
D. Proportionality of the guidance.....	138
1. Concept for modulation of objectives	138
2. Risk-based levelling of objectives	139
3. Additional risk-based levelling of information-security-related objectives.....	150
E. Annex 1 — Anticipated impact on regulations and MOC for major domains	151
1. Product design and operations	151
2. ATM/ANS.....	153
3. Aircraft production and maintenance	154
4. Training / FSTD	154
5. Aerodromes	156
6. Environmental protection.....	157
F. Annex 2 — Use cases for major aviation domains	158
1. Introduction	158
2. Use cases — aircraft design and operations.....	159
2.1. Visual landing guidance system — IPC CoDANN with Daedalean AG.....	160
2.2. Pilot assistance — radio frequency suggestion	174
2.3. Auto-Taxi system — IPC with Boeing.....	175
2.4. Pilot AI teaming — Proxima virtual use case	181
3. Use cases — ATM/ANS	185
3.1. AI-based augmented 4D trajectory prediction — climb and descent rates	185
3.2. Time-based separation (TBS) and optimised runway delivery (ORD) solutions	211
4. Use cases — aircraft production and maintenance.....	227
4.1. Controlling corrosion by usage-driven inspections.....	228
4.2. Damage detection in images (X-Ray, ultrasonic, thermography)	232
5. Use cases — training / FSTD	237
5.1. Assessment of training performance.....	237
6. Use cases — aerodromes.....	237
6.1. Detection of foreign object debris (FOD) on the runway	237
6.2. Avian radars	237
6.3. UAS detection systems.....	238
7. Use cases — environmental protection.....	238
7.1. Engine thrust and flight emissions estimation.....	238
8. Use cases — safety management	238
8.1. Quality management of the European Central Repository (ECR).....	238
8.2. Support to automatic safety report data capture	238

8.3.	Support to automatic risk classification.....	238
G.	Annex 3 — Definitions and acronyms	239
1.	Definitions.....	239
2.	Acronyms	251
H.	Annex 4 — References	255
I.	Annex 5 — Full list of questions from the ALTAI adapted to aviation	257
1.	Gear #1 — Human agency and oversight	257
2.	Gear #2 — Technical robustness and safety.....	261
3.	Gear #3 — Privacy, data protection and data governance.....	266
4.	Gear #4 — Transparency.....	268
5.	Gear #5 — Diversity, non-discrimination and fairness	270
6.	Gear #6 — Societal and environmental well-being	278
7.	Gear #7 — Accountability	281

Author	Guillaume Soudain — EASA AI Programme Manager
Reviewer	François 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 2.0 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-AI teaming'), 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 2.0 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 2021¹**. While, according to that Commission proposal², 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.2);
- **AI assurance** (Section C.3);
- **human-factors for AI** (Section C.4); and
- **AI safety risk mitigation** (Section C.5).

Section C.6 addresses the provisions that are anticipated to apply to the **organisations** developing or deploying AI-based systems.

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 2.0. 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-AI 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 (‘advanced automation’)**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 is evolving as technology develops. In the EASA AI Roadmap 1.0, it was chosen to use a wide-spectrum definition that is ‘any technology that appears to emulate the performance of a human’.

For its version 2.0 of the AI Roadmap, EASA has moved to the even wider-spectrum definition from the ‘Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence’ (EU Artificial Intelligence Act) (EU Commission, 2021), that is ‘technology

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.

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’.

In line with Annex I to the Proposal for an EU AI Act, AI techniques and approaches can be divided in **machine learning approaches** (also known as data-driven AI), **logic- and knowledge-based approaches** (also known as symbolic AI) and **statistical approaches**.

Even if the use of learning solutions remains predominant in the applications and use cases received from the aviation industry, it turns out that meeting the high safety standards brought by current aviation regulations pushes certain applicants towards a renewed set of **knowledge-based AI** approaches.

Moreover, it is important to note that those different AI approaches may be used in combination (also known as **hybrid AI**), which is also considered to fall within the scope of this Roadmap. Generative AI using large language models (LLMs) also will be considered under this category.

Consequently, the EASA AI Roadmap has been extended to encompass all techniques and approaches described in the following figure.

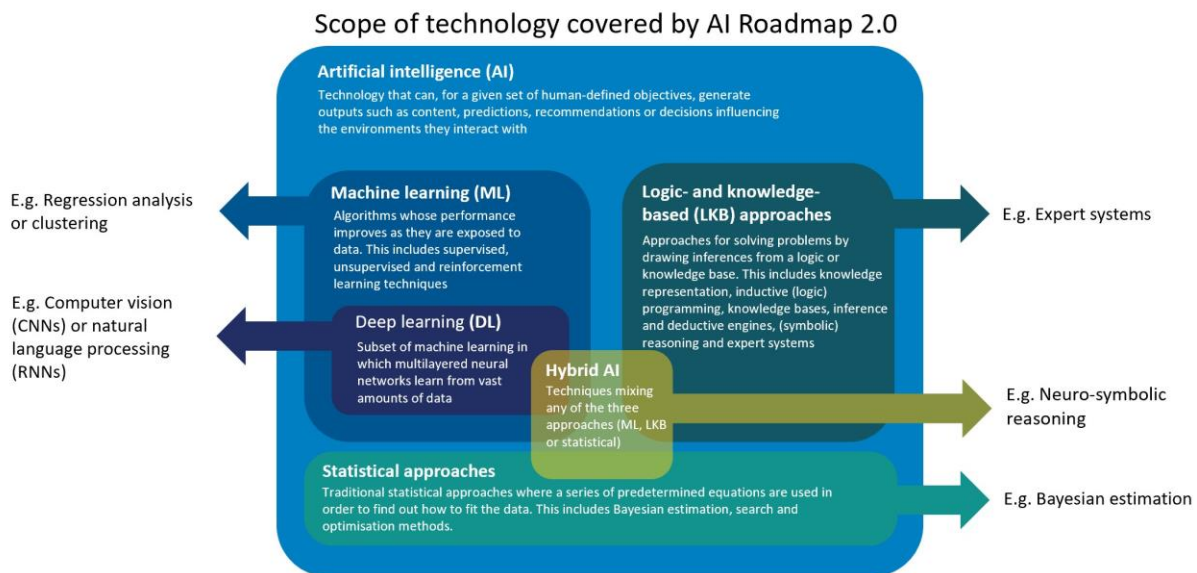


Figure 1 — AI taxonomy in this Roadmap

The technical scope of the Concept Paper will be augmented in subsequent Issues, to progressively encompass the whole scope of Figure 1. For now, the present document still applies only to a reduced scope encompassing **machine learning (ML)** and its **deep learning (DL)** subset.

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;
- Ensuring robustness and absence of ‘unintended behaviour’ 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 the ‘AI black box’. The expression ‘black box’ is a general concern raised with AI/ML techniques, as the complexity and nature of ML models bring a level of opaqueness that renders them more difficult to verify (unlike rule-based software); and
- Enabling trust by end users.

2. AI trustworthiness framework overview

To address the challenges of data-driven learning approaches, EASA AI Roadmap 1.0 identified 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 line with EASA AI Roadmap 2.0, two of the original building blocks required 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’:

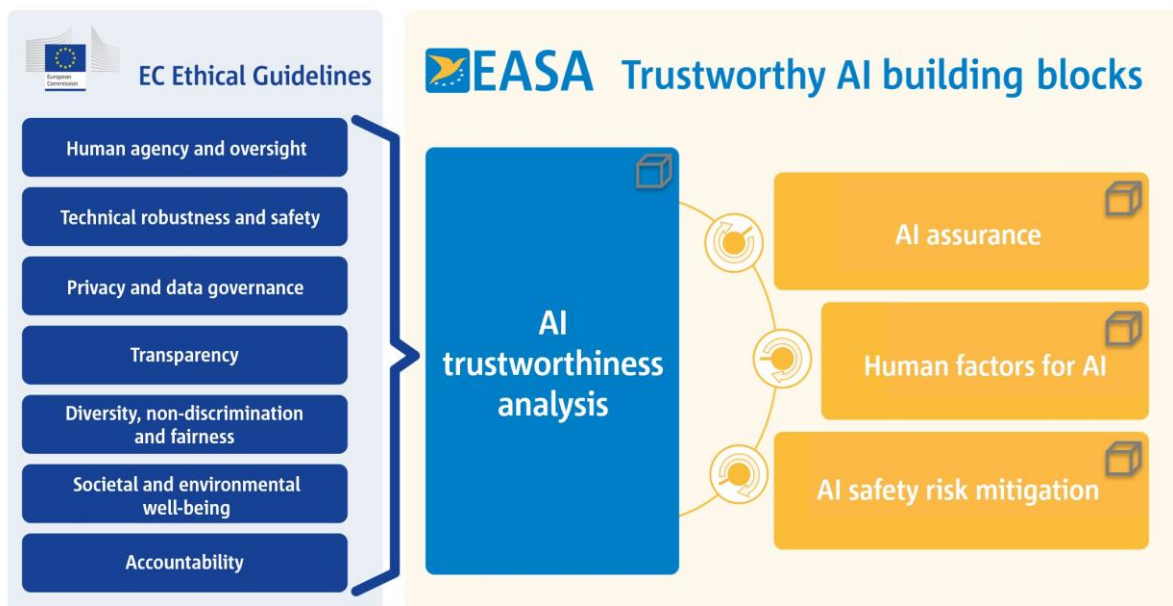


Figure 2 — EASA AI trustworthiness roadmap building blocks

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 the 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. It is important to remind that the **safety and security assessments** correspond to existing mandatory practices in the aviation industry; however, they are affected by the introduction of AI. These are not modified as regards their principles but require complementary guidance to address the specificities of AI techniques.

The **AI assurance** building block is intended to address the AI-specific guidance pertaining to the AI-based system. It encompasses three 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 & post-ops explainability** deals with the capability to provide users with understandable, reliable and relevant information with the appropriate level of detail on how an AI/ML application produces its results. Finally, 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 end users with understandable, reliable and relevant information with the appropriate level of detail and with appropriate timing on how an AI/ML application produces its results. This block also introduces the concept of **human-AI teaming** to ensure adequate cooperation or collaboration between 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 satisfy the whole set of objectives defined for the **AI assurance** and the **human factors for AI** building blocks, 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.

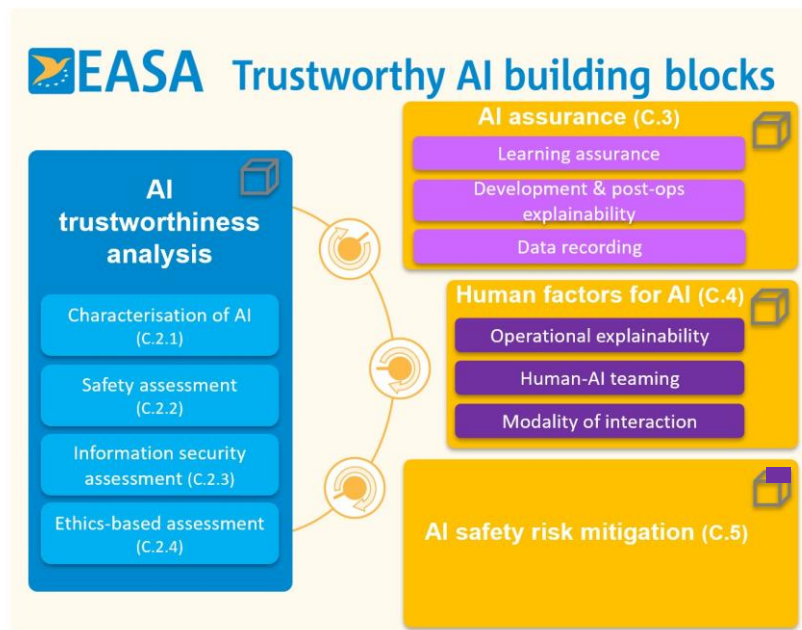


Figure 3 — EASA AI trustworthiness building blocks

The trustworthiness analysis is always required 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 classification and the criticality of the application, as described in Chapter D.

3. Terminology and scope of the document

The intent with the EASA Roadmap 2.0 technical scope extension is to identify AI techniques (beyond learning) requiring additional guidance compared to the existing development assurance. This will be done in subsequent Issues of the EASA Concept Paper. Already certified rule-based algorithms are not the target but rather a boundary to precise the applicability scope will be defined for each approach.

In this Issue of the document, the focus remains on **data-driven AI** approaches. Those can be further divided considering the 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 expected output (labels in the data sets). The learning algorithm then adjusts the parameters to increase the accuracy of the ML model.
- **Unsupervised learning** — this strategy is used in cases where the available data set is not labelled. 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.

Issue 2 of this document delves deeper into **supervised learning** approaches, while incorporating an initial set of objectives for **unsupervised learning** approaches. **Reinforcement learning** approaches will be further addressed in a next update of this Concept Paper.

Considering this scope, the W-shaped learning assurance process developed under the AI assurance building block has highlighted the need for an intermediate level between system and item, called ‘AI/ML constituent’. 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. In this figure, the elements identified as ‘traditional’ are meant to be addressed by the existing applicable system and item guidance.

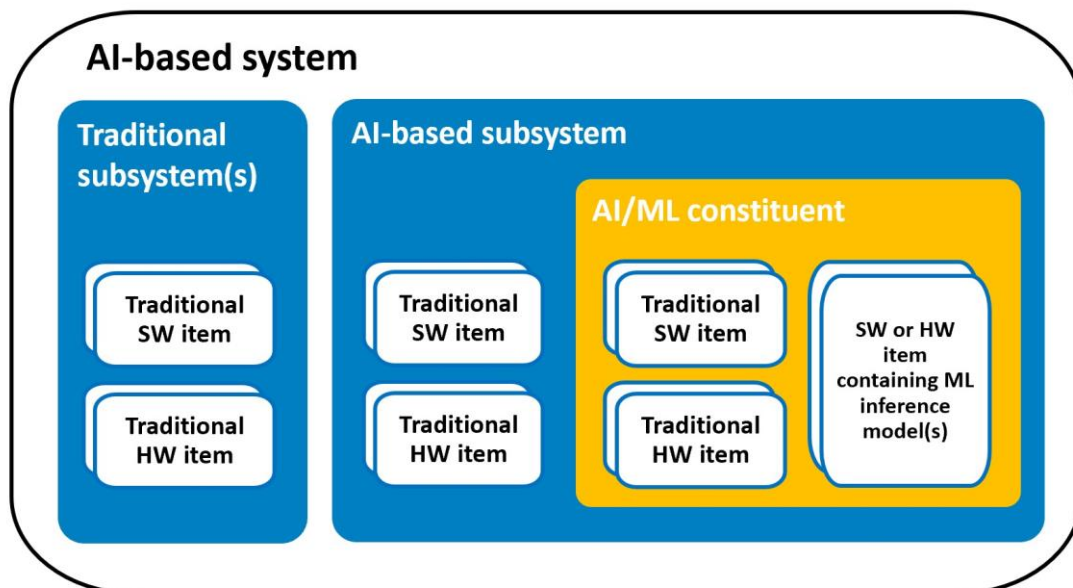


Figure 4 — Decomposition of an AI-based system

In this Figure 4:

- 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 defined and bounded collection of hardware and/or software item(s) which are grouped for integration purpose to support one AI-based subsystem function, including:
 - 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 necessary pre- and post-processing traditional items;
- 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 **supervised learning**, there is still limited experience gained from operations on the guidance proposed in this document and some anticipated MOC for a number of challenging objectives applicable to the highest levels of criticality are not yet available. Consequently, EASA will initially accept only applications where AI/ML constituents do not include IDAL A or B / SWAL 1 or 2 / AL 1, 2 or 3 items.

For **unsupervised learning**, some of the anticipated anticipated MOC are even less mature for a number of challenging objectives, such as the generalisation bounds expressed under Objective LM-04. Consequently, EASA will initially accept only applications where AI/ML constituents include IDAL D / SWAL 4 / AL 5 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-AI 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:

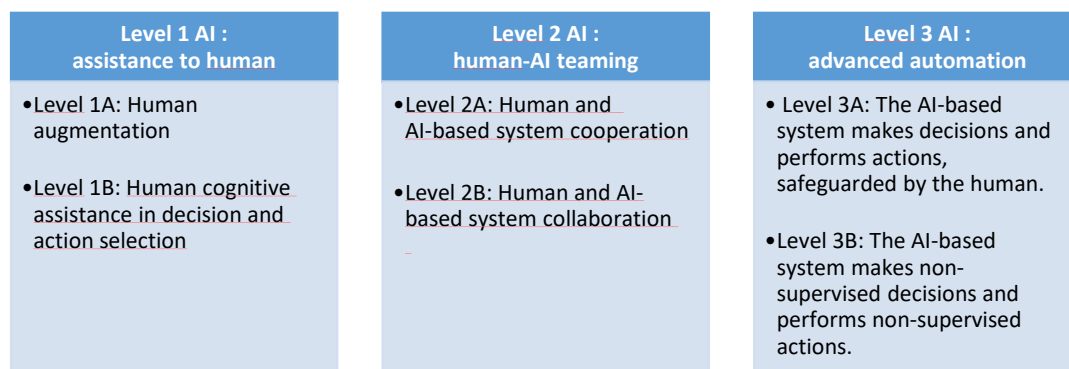


Figure 5 — Classification of AI applications

The key distinction between Level 1 and Level 2 AI applications lies in how decisions are implemented. For Level 1 AI, decisions are taken by the end user based on support by the AI-based system, and all actions are implemented by the end user. In contrast, Level 2 AI-based systems can perform automatic selection and implementation of actions, the end user still maintaining full oversight and override capability of the AI-based systems actions at any time. At Level 2, decisions can be taken either by the end user or automatically by the AI-based system under the direction and oversight of the end user.

The difference between Level 2 and Level 3 AI revolves around the extent of authority given to the AI-based system. At Level 2, the human-AI teaming concept foresees a partial release of authority to the system however under full oversight of the end user, who consistently remains accountable for the operations. On the contrary, at Level 3, the AI-based system is given full authority to make and implement decisions, under remote monitoring of the end user (Level 3A AI) or without end user involvement (Level 3B AI).

Note: Considering this distinction, the development of future Level 3 guidance will require specific considerations on the impact of this transfer of authority to the AI-based systems on the accountability scheme. In contrast, for Level 1 and 2 (scope of the present document) it is considered unaffected compared to current practices.

Detailed guidance on how to classify an AI-based system is provided in Section C.2.1.4.

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 the 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 need significant adaptations or changes to the existing ‘development assurance’ methods.

Intuitively, the assurance process should be oriented towards the correctness and completeness/representativeness of the data or scenarios used during the development and on the learning and its verification. Most importantly, the main challenges lie in providing an adequate level

of confidence that the trained models can generalise with an adequate performance on unseen operational data, and that the ML models are robust in all foreseeable conditions.

To this purpose, a new concept of ‘learning assurance’ is proposed to extend the traditional ‘development assurance’. 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.

6.2. AI explainability

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 AI/ML constituent satisfies the applicable requirements at a specified level of performance, and provides sufficient generalisation and robustness capabilities.

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 applications which include decision-making capabilities, questions arise as to how the end user will interpret the results and reasoning of AI-based systems. The development of advanced and complex AI techniques, for example, deep neural networks (DNNs), leads to major transparency issues for the end user.

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

- The information required to make an ML model interpretable for the users; and
- Understandable information for the end 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 abstraction of an explanation 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 detail 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 four 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 the approval/certification of (functional) systems embedding AI applications: certification authorities, NSAs, 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
<ul style="list-style-type: none"> ▪ 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 stakeholders to identify errors in the model or poor performance in some areas of the input space, and explain them ▪ Support certification and oversight ▪ Support continuous analysis of the AI-based system behaviour ▪ Support the safety investigation of accidents and incidents where an AI-based system was involved 	<ul style="list-style-type: none"> ▪ Contribute to building trust for the end user ▪ Contribute to anticipating 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: the capture of operating conditions is already a practice in the aviation domain, which corresponds to the conditions under which a given product or 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 detail captured at system level is not commensurate with the level of detail 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. A correct and complete definition of the ODD is a prerequisite to an adequate level of quality of the data sets involved in the learning assurance process.

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.3.1.2.1).

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, may involve a shared understanding of goals, roles and processes (decision-making/problem solving) between the HAT 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 MOC (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 situation 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 end user and the AI-based system work together and jointly to achieve a predefined shared goal and solve a problem through co-constructive approach. Collaboration implies the capability to share situation awareness and to readjust strategies and task allocation in real time. Communication is paramount to share valuable information needed to achieve the goal.

Note: While it is understood that AI-based systems do not have situation awareness but situation representation, for ease of reading the term ‘shared situation awareness’ is used to denote this specific element of collaboration.

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.

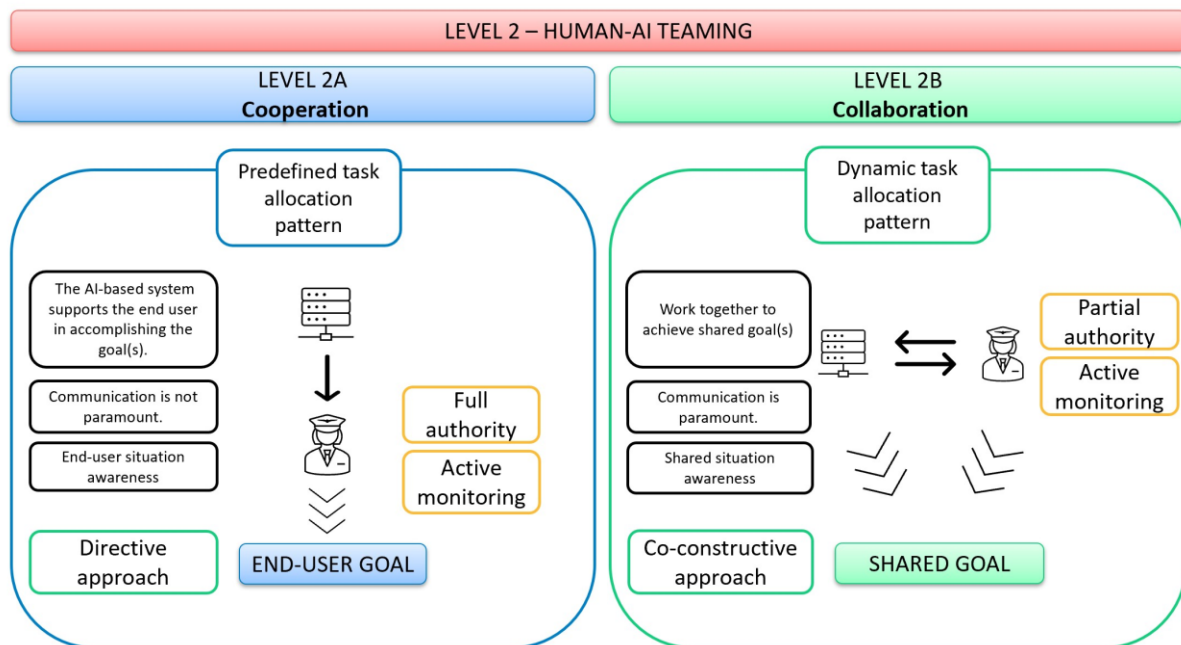


Figure 6 — HAT concept overview

Within the context of Figure 6, it is important to consider two constructs that are key to describing the roles and responsibilities that will be assigned to AI-based systems:

- Goals and tasks.
- Allocation schemes and patterns.

Goals and tasks

Goals and tasks describe the organisation and breakdown of what is expected of the human AI-team. A 'goal' is a predefined higher-level purpose towards which the teaming effort is directed. A 'high-level task' is a cluster of tasks contributing to achieving a goal, at the highest level of interaction between the human and the AI-based system.

A 'task' is any discrete and complete activity contributing to the achievement a high-level task.

An example of a Level 2B 'goal' might be 'manage flight profile', and an associated high-level task can be 'descend the aircraft'; the AI-based system and pilot collaborate on achieving the goal. The AI-based system takes responsibility for the speed and the pilot takes responsibility for the aircraft attitude and trim. The tasks related to speed (for example, airbrakes and throttle) are managed by the AI-based system. The pilot does not interfere with the management of the throttles. General principles in considering tasks include:

- A single goal can be achieved through one or more tasks.
- The same task can be allocated to either the end-user or the AI-based system, but not both at the same time.

- A task that was previously allocated to an end-user can be allocated to an AI-based system at a different time.
- Multiple different tasks may be allocated to the end-user or AI-based system simultaneously.
- Each task should be discrete and complete.

Allocation scheme and allocation pattern

When addressing the migration of a system between AI Level 2A, Level 2B and Level 3A, the ‘allocation scheme’ and the ‘allocation pattern’ must be considered. Allocation schemes refer to the overall envelope of tasks which can be allocated to either the end user or the AI-based system. For instance, an AI-based system that is charged with managing the aircraft altitude could have access to tasks such as: move horizontal surfaces, extend or retract flaps, set reference altitude and engage autopilot, set thrust, etc. The same AI-based system may not have access to any other function of the aircraft. In this case, the allocation scheme of such an AI-based system is limited to those functions that affect the aircraft’s altitude.

Using the same example, an allocation pattern would refer to the set of tasks that are allocated to the AI-based system at a specific time. During cruise, the AI-based system allocation pattern may only monitor and manage speed to maintain altitude. During initial climb, the AI-based system allocation pattern may include controlling flaps, ailerons, thrust aircraft attitude and speed. In each case, the allocation pattern is different, but both scenarios fall within a single allocation scheme.

Both for Level 2A and 2B the allocation scheme is fixed. The allocation pattern within 2A is predefined, whereas within Level 2B it is dynamic.

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 capabilities (more specifically for Level 2B), situation representation, transparency and adaptivity.

The human factors guidance developed in Section C.4.2 provides detailed objectives and anticipated MOC 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 are 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 explainability; 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 MOC.

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, 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. New regulations have been adopted in support of the conformity assessment framework in the ATM/ANS domain: Delegated Regulation (EU) 2023/1768 lays down detailed rules for the certification and declaration of air traffic management/air navigation services systems and air traffic management/air navigation services constituents, while Implementing Regulation (EU) 2023/1769 establishes technical requirements and administrative procedures for the approval of organisations involved in the design or production of air traffic management/air navigation services systems and constituents. The conformity assessment framework now benefits from AMC, GM, and DSs for the certification or declaration of conformity, or statement of design compliance of the ATM/ANS equipment.

- **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), design or production organisations (DPOs) of ATM/ANS systems and ATN/ANS constituents (hereafter ‘ATM/ANS equipment’), 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** or **unsupervised learning**, but not other types of learning such as **reinforcement learning**;
- covering **offline learning** processes where the model is ‘frozen’ at the time of approval, but not **online learning** processes.

2. Trustworthiness analysis

The **trustworthiness analysis** building block encompasses different assessments including ethical aspects, safety and security. The security and safety assessments are not modified as regards their principles but require complementary guidance to address the specificities of AI techniques. Additional guidance is necessary to cover the recent development in ethical aspects.

This assumes that there is no distinction depending on the type of learning selected, i.e. **supervised**, **unsupervised** or **reinforcement learning**. Therefore, the objectives of this chapter apply indistinctively to all learning approaches.

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 (including indication of the level of teaming with the AI-based system, i.e. none, cooperation, collaboration) and 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 goals and associated high-level tasks are intended to be performed in interaction with the AI-based system.

Anticipated MOC CO-02: The high-level tasks should be identified at the highest level of interaction between the human and the AI-based system, not going down to the level of each single task 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 product or 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 goals and associated 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 scheme between the end user(s) and the AI-based system, further dividing the high-level tasks identified under **Objective CO-02** in as many 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 scheme;
- the definition of the OD, including the specific operating limitations and conditions appropriate to the proposed operation(s) and considering the product as a whole; for instance, in the airworthiness domain, the AI/ML (sub)system should perform as intended under the aeroplane operating and environmental conditions);

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

As mentioned above, there exists a relationship between an operational scenario and the operational domain in the sense that an operational scenario sustaining the ConOps is executed in a given operational domain that needs to be characterised as well. Figure 7 shows the interrelationship between the operational scenarios for the ConOps and the operating parameters for the OD:

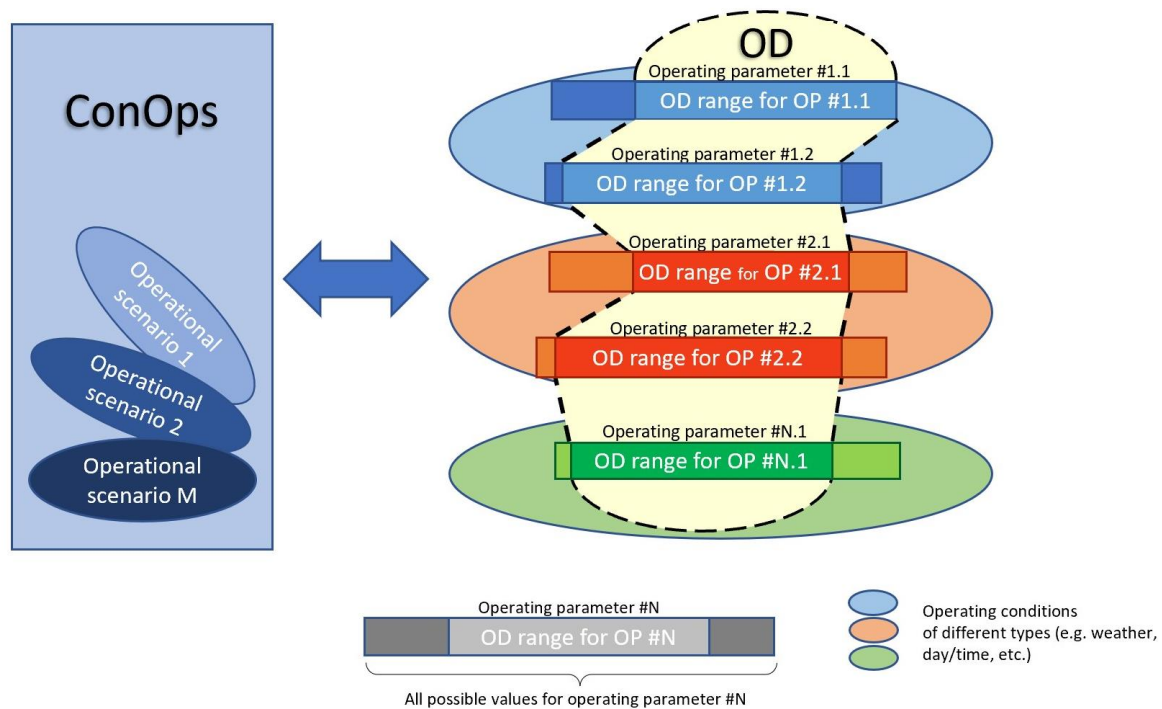


Figure 7 — Interrelationship between ConOps and OD

Notes:

- The OD takes into consideration the environmental conditions, including geographical aspects or weather conditions, under which the AI-based system is intended to operate.
- 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.3.1.2.1).
- 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.
- Operational scenarios should not be limited to nominal cases but also consider degraded modes where the AI-based system is not performing as expected.

- 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 end-user representatives in planning, design, validation, verification and certification/approval of an AI-based system. The end-user representatives' 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, as well as a functional decomposition and allocation down to the lowest level.

Anticipated MOC-CO-06: The functional analysis and decomposition consist in identifying a set of high-level function(s), and their breakdown 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 and decomposition is an enabler to meet the objectives in Section C.3.1.2 'Requirements and architecture management' of the learning assurance.
- The functional analysis and decomposition is a means supporting the functional hazard assessment (FHA) as per Section C.2.2.3 'Initial safety (support) assessment'.
- The following standards with adaptation may be used for embedded systems: ED-79B/ARP4754B.

2.1.4. Classification of the AI application

This usable guidance document at Issue 02 focuses on **Level 1 and Level 2 AI applications**. It therefore provides classification guidelines for these levels, including boundaries between them, 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 ‘**directed**’, ‘**supervised**’, ‘**safeguarded**’ or ‘**non-supervised**’ by the end user. Moreover, 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 defined as the ability to make decisions without the need for approval from the other member.

- **Directed:** *capability of the end user to actively monitor the tasks allocated to the AI-based system, with the ability to cross-check every decision-making and intervene in every action implementation of the AI-based system. This corresponds to **full authority for the end user**.*
- **Supervised:** *capability of the end user to actively monitor the tasks allocated to the AI-based system, with the ability to intervene in every action implementation of the AI-based system, with some decisions being taken and actions being implemented by the AI-based system in a relative independence, while maintaining a shared situation awareness between both members. This corresponds to **partial authority for the end user**.*
- **Safeguarded:** *capability of the end user to oversee the operations of the AI-based system, with the ability to override the authority of the AI-based system (for selected decisions and actions) when it is necessary to ensure safety and security of the operations (upon alerting). This corresponds to **limited authority for the end user upon alerting**. The end user may revert to ‘full’ or ‘partial’ authority depending on the ConOps and on the nature of events occurring in the operations.*
- **Non-supervised:** *no end user is involved in the operations and therefore there is no capability to override the AI-based system’s 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	Directed decision and automatic action implementation	Full
Level 2B Human-AI collaboration	Supervised automatic decision and action implementation	Partial
Level 3A Safeguarded advanced automation	Safeguarded automatic decision and action implementation	Limited, upon alerting
Level 3B Non-supervised advanced automation	Non-supervised automatic decision and 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 2.
- 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 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, 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. With the latter, the levels 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 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 support to 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 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). Finally, the notion of support implies that the selected action implementation is solely taken by the end user and not by the AI-based system.

Note: Level 1A or 1B imply no capability of decision-making for the AI-based system. There may be automatic action implementation by the AI-based system at Level 1 depending on the ConOps; however, this is not relevant to the AI Level classification.

The boundary between level 1B and level 2A is based on the distinction between support to decision-making and automatic decision and action implementation (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 or actions implementation are fully monitored and overridable by the 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 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 decision-making and action 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 situation awareness and to readjust task allocation in real time (e.g. virtual co-pilot in a reduced-crew 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 HAT members to readjust strategies and decisions).

The boundary between level 2B and level 3A lies in the high level of authority of the AI-based system and the limited oversight that is performed by the 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 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 (Level 3A) or absence (Level 3B) of a end user in the loop of operations.

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** process 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.

For non-AI-based (sub)systems, 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 (see Figure 4) operate at least as safely as traditional systems developed using existing *development assurance* processes and *safety assessment* methodologies⁶: the acceptable level of risk to persons, personal properties or critical infrastructure incurred by an AI technology introduction, should be no higher than of an equivalent traditional system. 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 the development 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.

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

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.

Aviation domains	'Initial' safety assessment	'Continuous' safety assessment
Initial and continuing airworthiness	As per Section C.2.2.2.1	As per Section C.2.2.4 'continuous safety assessment' and Provision ORG-03
Air operations	See Note A	As per Section C.2.2.4 'continuous safety assessment' and Provision ORG-03
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 Section C.2.2.4 'continuous safety assessment' and Provision ORG-03 – see Note F
Maintenance	See Notes A and C	As per Section C.2.2.4 'continuous safety assessment' and Provision ORG-03
Training	See Notes A and D	Managed from an organisation, operations and negative training, as per Section C.2.2.4 'continuous safety assessment' and Provision ORG-03
Aerodromes	See Note A	As per Section C.2.2.4 'continuous safety assessment' and Provision ORG-03
Environmental protection	See Note E	Currently not applicable

Table 3 — Safety assessment concept for the major aviation domains

Note A: For domains not having guidance on initial safety assessment, an AI-specific risk assessment process is intended to be developed through RMT.0742 to support **Objective SA-01** and anticipated MOC developed in Section C.2.2.3.

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

Note D: 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 E: 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 F: 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 development phase.

- Perform functional hazard assessment in the context of the ConOps
- Safety assessment activities supporting design and validation phases
 - Define safety objectives⁷, 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 performance⁸ 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.**
 - **Identify AI/ML item failure modes**
 - Derive safety requirements including independence requirements to meet the safety objective and support the architecture
 - Define and validate assumptions
- **Verification phase**
 - **Perform final safety assessment**

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 DA-03.

- **Consolidate the safety assessment to verify that the implementation satisfies the safety objectives¹⁰.**

2.2.2.2. Impact on safety support assessment

The analyses below describe the typical safety support assessment activities performed during the development 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 performance¹¹ 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**

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.

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 DA-03.

The following anticipated MOC 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:
 - ED-79B/ARP4754B 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.

Starting from the AI-based system and functional analysis, the DAL/SWAL allocation should be done down to the AI/ML constituent level.

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

Considering the limited experience from operations on the guidance proposed in this document and the unavailability of some MOC for a number of challenging objectives applicable to the highest levels of criticality, EASA will initially accept only applications where AI/ML constituents do not 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.

However, should an AI-based (sub)system be composed of different AI/ML constituents, the safety analysis could however allocate different assurance levels to these different AI/ML constituents.

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).

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).

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 limitations or 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 safe operation.

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

To mitigate the exposure to data outside the OD or ODD, these means or a combination of them are expected to be necessary to deliver the intended behaviour:

- 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 functions for the AI/ML constituent to continue to deliver the intended behaviour when input data is outside the ODD;
- Put in place functions for the AI-based (sub)system to ensure safe operation when input data is outside the OD.

For low-dimensional input space (e.g. sensors producing 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 (such as in computer vision with images or videos, or in NLP). In such use cases, techniques such as the out of distribution (OoD) discriminator (EASA and Daedalean, 2020) could be envisaged.

When input data is outside the OD, the intended function cannot be fulfilled. In such a situation, it is expected that monitoring combined with alerting functions and procedures are implemented to ensure safe operation.

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 deficiencies due to lack of knowledge or information. In the context of ML, epistemic uncertainty corresponds to the situation where the model has not been exposed to data adequately covering the whole ODD or where the ODD definition needs to be refined or completed.
- Aleatory uncertainty refers to the intrinsic randomness in the data. This can derive from data collection errors, sensor noise, or noisy labels. In this case, the model has learnt based on data suffering from such uncertainties..

Notes:

- It is to be noted that these notions of epistemic and aleatory uncertainties are not new; however, they require a specific refinement and disposition in the context of this AI/ML guidance.

- 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 two following anticipated MOC.

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

Sources of uncertainties affecting the AI/ML constituent 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.

Epistemic uncertainty is addressed through the learning assurance objectives.

Anticipated MOC-SA-01-6: Establishment of AI/ML constituent failure modes¹³:

- Establish a taxonomy of AI/ML constituent failures;
- Evaluate possible failure modes and associated detection means.

Anticipated MOC-SA-01-7: Link between performance metrics and safety assessment

When a quantitative safety (support) assessment is required to demonstrate that the safety requirements are met, performance metrics should provide a conservative estimation of the probability of occurrence of the AI/ML constituent failures modes.

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

Anticipated MOC-SA-01-8: Link between generalisation bounds and safety assessment

Based on the generalisation gap evaluated per **Objective LM-04**, the applicant should assess the impact on the safety (support) assessment. This should be supported by specifying margins on performance requirements as part of the safety (support) assessment.

When a quantitative safety (support) assessment is required to demonstrate that the safety requirements are met, the probability of occurrence of the AI/ML constituent failure modes 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

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.

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-01-8 or additional anticipated MOC is expected to benefit from the MLEAP project deliverables.

As an example, in support of anticipated MOC SA-01-2, anticipated MOC SA-01-6, anticipated MOC SA-01-7 and anticipated MOC SA-01-8, the following approach may be used to establish safety requirements associated with AI/ML constituent 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. Establish AI/ML constituent failure modes (as per anticipated MOC-SA-01-6).
3. Identify appropriate metrics to evaluate the model performance and initiate an early specification of the thresholds necessary to meet the safety objectives (as per anticipated MOC SA-01-2).
4. Identify how performance metrics translate into a probability of occurrence of the ML model failure mode (as per anticipated MOC-SA-01-7).

Note: This step is done through **Objective LM-09** and **Objective IMP-06** in the learning assurance chapter.

5. Assess and quantify, when applicable, generalisation bounds either through the model complexity approach or through the validation/evaluation approach. This leads to bounds for almost all data sets on average over all inputs. Based on those bounds, specify margins on performance metrics.

Note: This step is done through **Objective LM-04** in the learning assurance chapter. The output of this objective may then be used to specify margins on performance metrics. There may be some iterations between **Objective LM-04** and **Objective SA-01** in case the generalisation bound would force to account for too high margins. In such a case, either a stronger generalisation bound may be achieved by constraining further the learning process or changes to the system (e.g. system architecture consideration) may be considered.

6. Identify how performance metrics with associated margins translate into a probability of occurrence of the ML model failure mode (as per anticipated MOC SA-01-8).
7. Analyse the post-processing system to show how it modifies the latter failure probabilities. Usually, the post-processing results in improved performance (with respect to the chosen metrics) and/or reduction of the impact of the ML model failures on the AI/ML constituent performance metrics.
8. Study the elevated values of the error metrics for the model on the training/validation (eventually testing) data sets, and develop adequate mitigations, for example by:
 - Characterising regions of the ODD where elevated values of the error metrics are gathered
 - Proposing architectural means or limitations
 - Proposing other mitigations discussed in Section C.5.

9. Based on all the previous steps, derive the necessary safety requirements.

Anticipated MOC SA-01-9: 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:
 - ED-79B/ARP4754B 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 ‘AI safety risk mitigation’ building block is defined. The AI safety risk mitigation 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 ‘AI safety risk mitigation’, which is not aimed at compensating for 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. Continuous safety assessment¹⁴

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;

¹⁴ 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.

- 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)).

As highlighted in these examples in relation to the airworthiness and ATM/ANS domains, some existing aviation regulations already anticipate the need to record certain information to ensure that the level of safety is maintained throughout the life of a certain aviation product.

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

Objective SA-02: The applicant should identify which data needs to be recorded for the purpose of supporting the continuous safety assessment.

Anticipated MOC SA-02

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.

The applicant should make sure that the data identified per this objective are listed in the data to be recorded per **MOC EXP-04-2**.

Objective SA-03: In preparation of the continuous safety assessment, the applicant should define metrics, target values, thresholds and evaluation periods to guarantee that design assumptions hold.

Anticipated MOC SA-03

For the airworthiness domain, this objective could be implemented through a CCMR approach. For other domains, specific guidance will need to be developed.

When defining the metrics, the data set and gathering methodology should ensure:

- the acquisition of safety-relevant data related to accidents and incidents (e.g. near-miss events);
- the monitoring of in-service data to detect potential issues or suboptimal performance trends that might contribute to safety margin erosion;
- 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.

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, another source of information security risk may be the organisation processes such as design, maintenance or production processes, which should be adequately managed.

In light of this, Commission Delegated Regulation (EU) 2022/1645 (applicable as of 16 October 2025) and Commission Implementing Regulation (EU) 2023/203 have introduced a set of information security requirements for approved organisations, that should be also taken into account. For further considerations related to AI-based system, refer to 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.

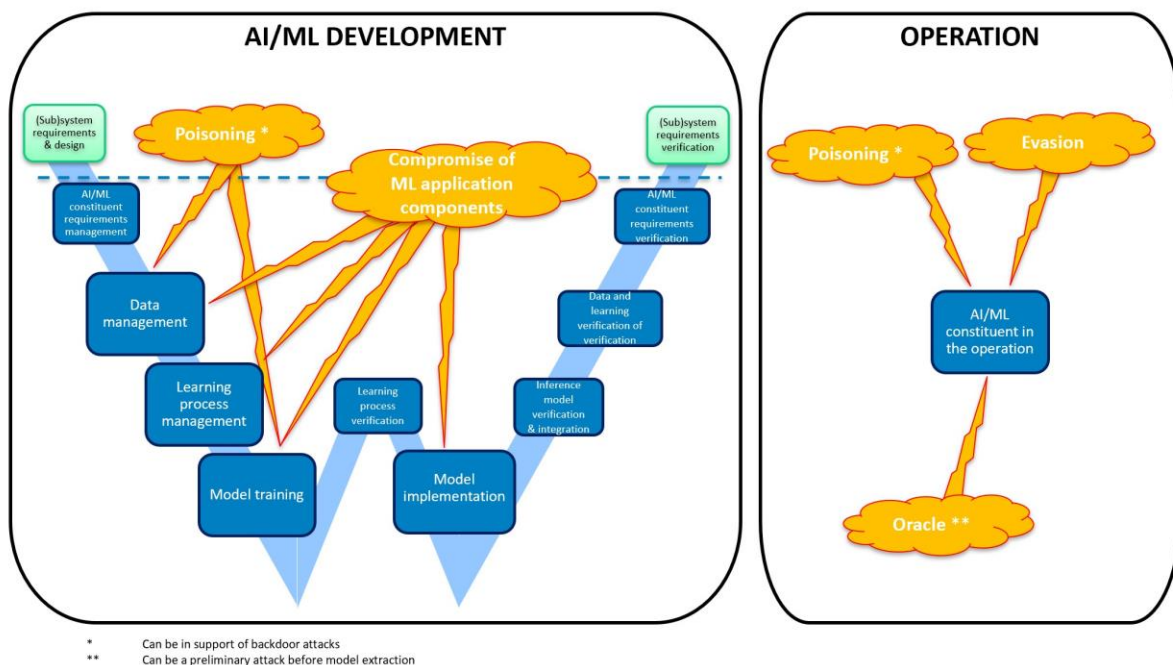


Figure 8 — Threats during the life cycle of the AI/ML constituent

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 ‘poisoning’ attacks (in addition to already identified considerations at organisational level (see Anticipated AMC ORG-02)) in which the attacker alters data to modify the behaviour of the algorithm in a chosen direction.
- 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. ‘Oracle’ types of attack should not only concern an immediate loss of IP, as they provide the attacker with useful insights on the model, enabling the design of more harmful attacks.

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, besides the security control already in place. 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.3.1).

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.

The verification of the effectiveness of the security controls typically takes place as part of any verification step during the development cycle, taking into account the specific threat under consideration.

As an example, the STRIP technique could be applied for verifying robustness against ‘poisoning’ before the model enters into service. Also Anticipated MOC LM-13-1 refers to verification aspects regarding adversarial cases in the context of ‘evasion’.

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. lawfulness¹⁵, 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)**¹⁶, containing several questions and explanation.

Conscious of this 2019/2020 Commission approach being a non-binding suggestion, the present EASA-guidelines builds on this concept and strives to 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 shown 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.

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).

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**.

¹⁵ 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.’

¹⁶ https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

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

Anticipated MOC ET-02: AI-based systems with the potential of creating overreliance, attachment, stimulating addictive behaviour, or manipulating the user's or end user's 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: 'reliance', 'overreliance' and 'attachment' which have been added in Annex 3 — Definitions and acronyms. A notable difference is that attachment is related to an emotional link or bond whereas over-reliance is more pragmatically related to trust and dependence on support. The organisation processes and procedures should 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), the difference lying in the capacity of the end user to perform oversight.

To ensure avoidance of overreliance, attachment, dependency and manipulation, requirements-based tests (per **Objective IMP-09**) should include the verification that the end users interacting with the AI-based system can perform oversight.

Note: Risks related to 'manipulation' are further mitigated through the guidance on operational explainability.

2.4.2. Gear #2 — Technical robustness and safety

All ALTAI 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.

2.4.3. Gear #3 — Privacy, data protection 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 Data Protection Officer of the organisation, 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.

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

Objective ET-03: The applicant should comply with national and EU data protection regulations (e.g. GDPR), i.e. involve their Data Protection Officer, consult with their National Data Protection Authority, etc.

Anticipated MOC ET-03: 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).

For requirements and objectives linked to the governance (ownership and usage) data that is used for the training of the AI/ML models or resulting from the interaction between the end user and the AI-based system, additional guidelines will be developed in the future Issue 03 of this document.

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. EASA nevertheless 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 focusing on individual persons or groups of people, not to consideration on bias in data sets used to train an AI-based system (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 with safety relevance, 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-04: The applicant should ensure that the creation or reinforcement of unfair bias in the AI-based system, regarding both the data sets and the trained models, is avoided, as far as such unfair bias could have a negative impact on performance and safety.

Anticipated MOC ET-04: 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 performance and safety.

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

Societal well-being

Objective ET-05: The applicant should ensure that end users are made aware of the fact that they interact with an AI-based system, and, if applicable, whether some personal data is recorded by the system.

Anticipated MOC ET-05: The applicant should issue clear and transparent information to the end user on the AI-based nature of the system and on any end-user-related data that is recorded due to his or her interaction with the system. The information could be provided through user manuals or through the AI-based system itself.

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), and define measures to reduce or mitigate these impacts.

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. For instance, 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 autonomy¹⁷.
- Does the AI-based system have adverse effects on the regulated aircraft/engine noise and emissions or aircraft fuel venting?

¹⁷ Kern et al., [Sustainable software products – Towards assessment criteria for resource and energy efficiency](#), Elsevier B.V., 2018.

- 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 (CO₂ emissions) and aircraft noise around airports.
- Could the use of the AI-based system have rebound effects, e.g. lead to an increase in traffic, which in turn could become harmful for the environment or human health?
- Could the use of the AI-based system have direct effects on the human health, including the right to physical, mental and moral integrity?

Regarding the reduction or mitigation measures, the applicant could 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 could 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-07** and **Objective ET-08**, this sub-gear may not be applicable to all aviation use cases. EASA nevertheless 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 with safety relevance, 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-07: 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-07, Provision ORG-08**).

Anticipated MOC ET-07: The applicant should identify the new skills for the users and end users through means of comparison with on one hand the past working practice and on the other hand what is expected from the AI-based system. The set of new skills should be developed through means of training (theoretical and practical) complemented by a mentoring phase on the job in order to be sure that the skills resulted in a successful and safe work performance.

Objective ET-08: 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-07**, **Provision ORG-08**).

Anticipated MOC ET-08: The applicant after a skills needs analysis should provide means to retain these set of skills, for example through training and practice in a controlled environment. At the end of the training programme, skills should be evaluated in order to ensure that they remain at an adequate proficiency level. The analysis and the means to retain the skills should be commensurate with the AI Level of the AI-based system (i.e. expected to be more stringent for Level 2B versus Level 2A).

2.4.7. Gear #7 — Accountability

Parts of the ‘Accountability’ questions, i.e. regarding ‘auditability’ and ‘risk management’, are considered by EASA to be addressed through compliance with other objectives of the AI trustworthiness framework contained in this document; other parts fall outside the EASA remit.

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:

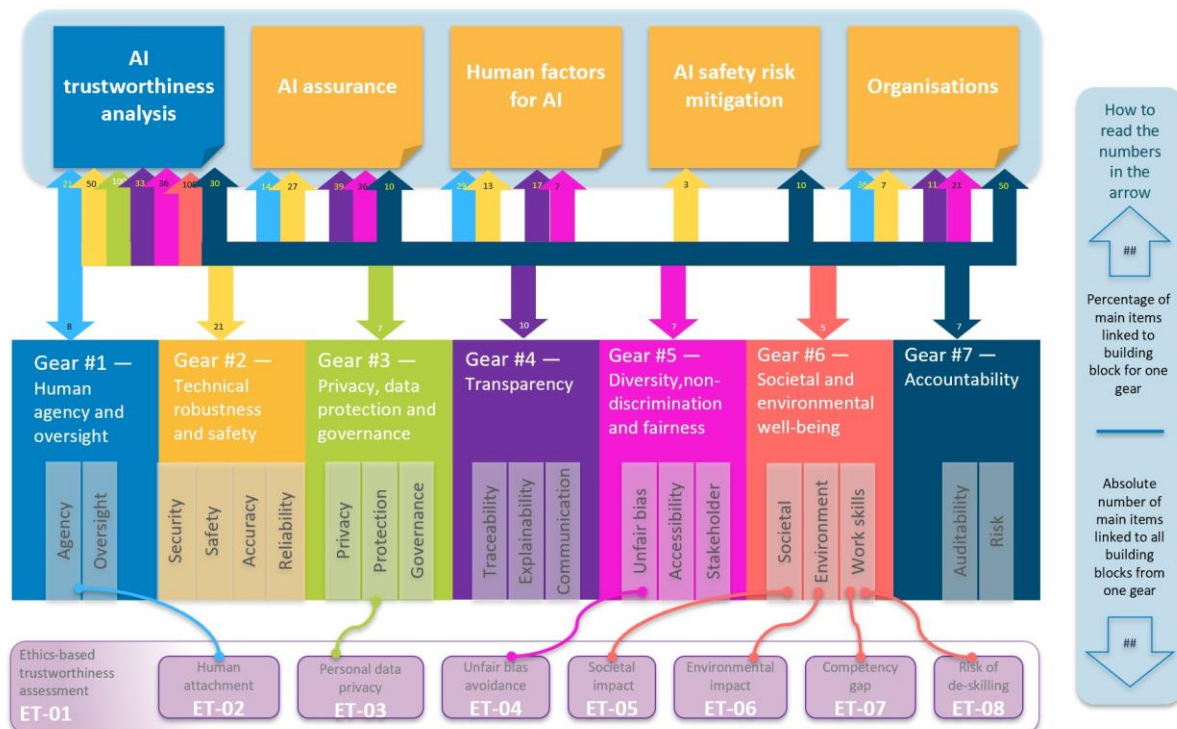


Figure 9 — Mapping of the 7 gears to the AI trustworthiness building blocks

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

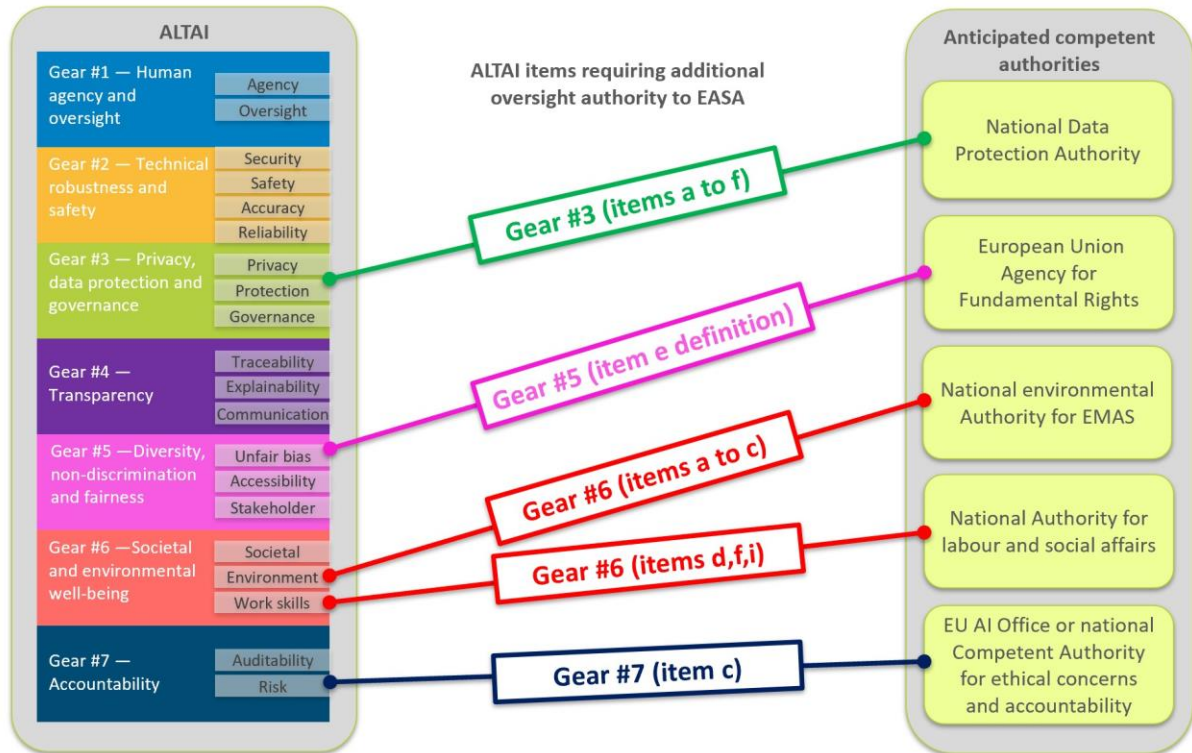


Figure 10 — Anticipation of mapping of outstanding items to other competent authorities

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 the focus on **human factors for 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 and post-ops 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.3.2.7 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 behaviour of the AI-based system at an appropriate level of performance, and at ensuring that the resulting trained models possess sufficient generalisation and robustness capabilities.

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

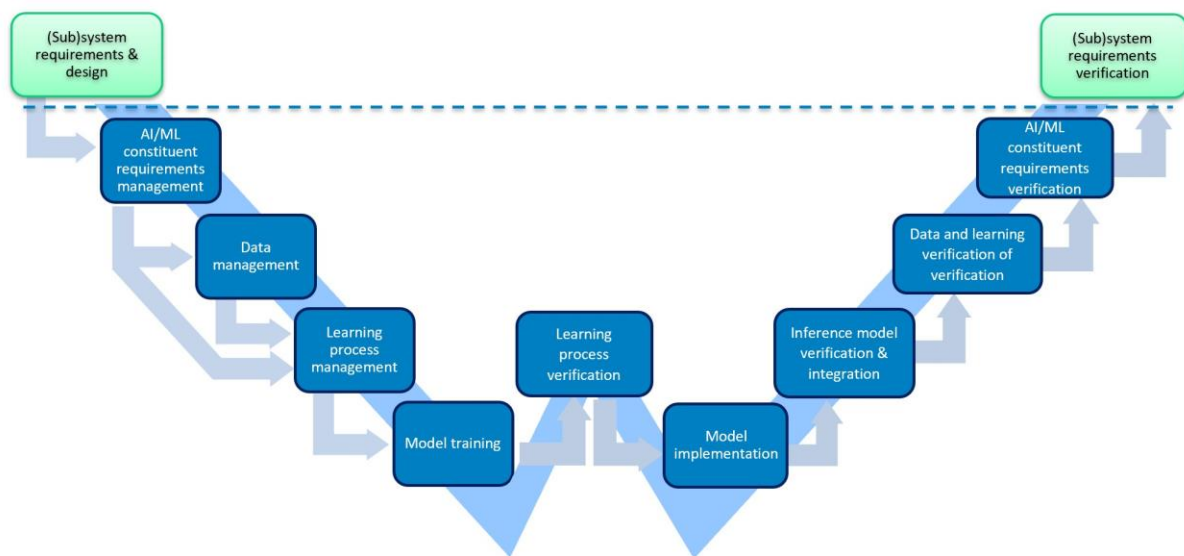


Figure 11 — Learning assurance W-shaped process

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.

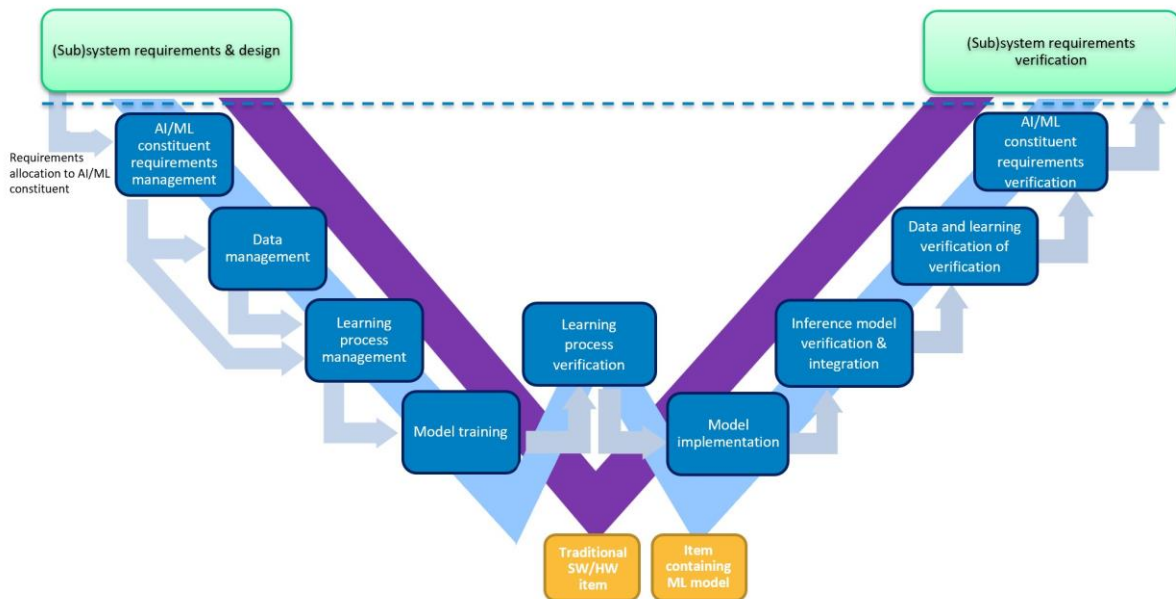


Figure 12 — Global view of learning assurance W-shaped process, non-AI/ML constituent V-cycle process

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 depicted in Figure 13 — Iterative nature of the learning assurance process (purple and green arrows).

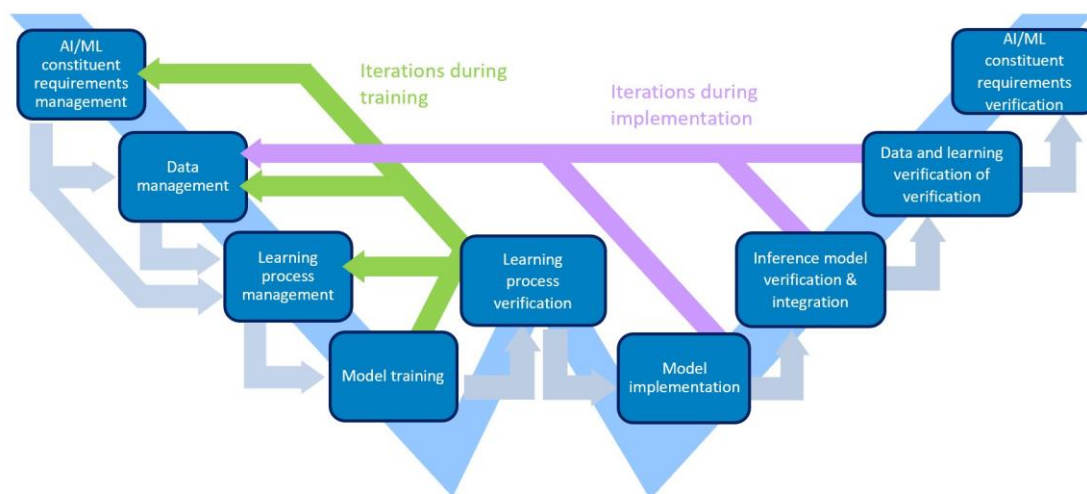


Figure 13 — Iterative nature of the learning assurance process

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 process¹⁸;
 - 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 environment represented by the OD or ODD, and the associated data on which the ML model was trained, 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.3.1.14, as well as the interface and compatibility with development assurance processes.

Anticipated MOC DA-01: 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).

¹⁸ While it is not possible to completely automate all the process steps (e.g. feature engineering or data labelling), 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/ML constituent architecture development¹⁹;
- requirements validation.

3.1.2.1. Capture of the AI/ML constituent requirements

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

Objective DA-02: Based on (sub)system requirements allocated to the AI/ML constituent, the applicant should capture the following minimum for the AI/ML constituent requirements:

- safety requirements allocated to the AI/ML constituent (e.g. performance, reliability, resilience);
- 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 AI/ML constituent ODD monitoring and performance monitoring (to support related objectives in Section C.3.2.6), detection of OoD input data and data-recording requirements (to support objectives in Section C.3.2.7);
- other non-functional requirements allocated to the AI/ML constituent (e.g. scalability); and
- interface requirements.

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

Objective DA-03: The applicant should define the set of parameters pertaining to the AI/ML constituent ODD, and trace them to the corresponding parameters pertaining to the OD when applicable.

¹⁹ This step is different from the model architecture described in Section C.3.1.4.

Figure 14 shows the refinement of the OD into the AI/ML constituent ODD.

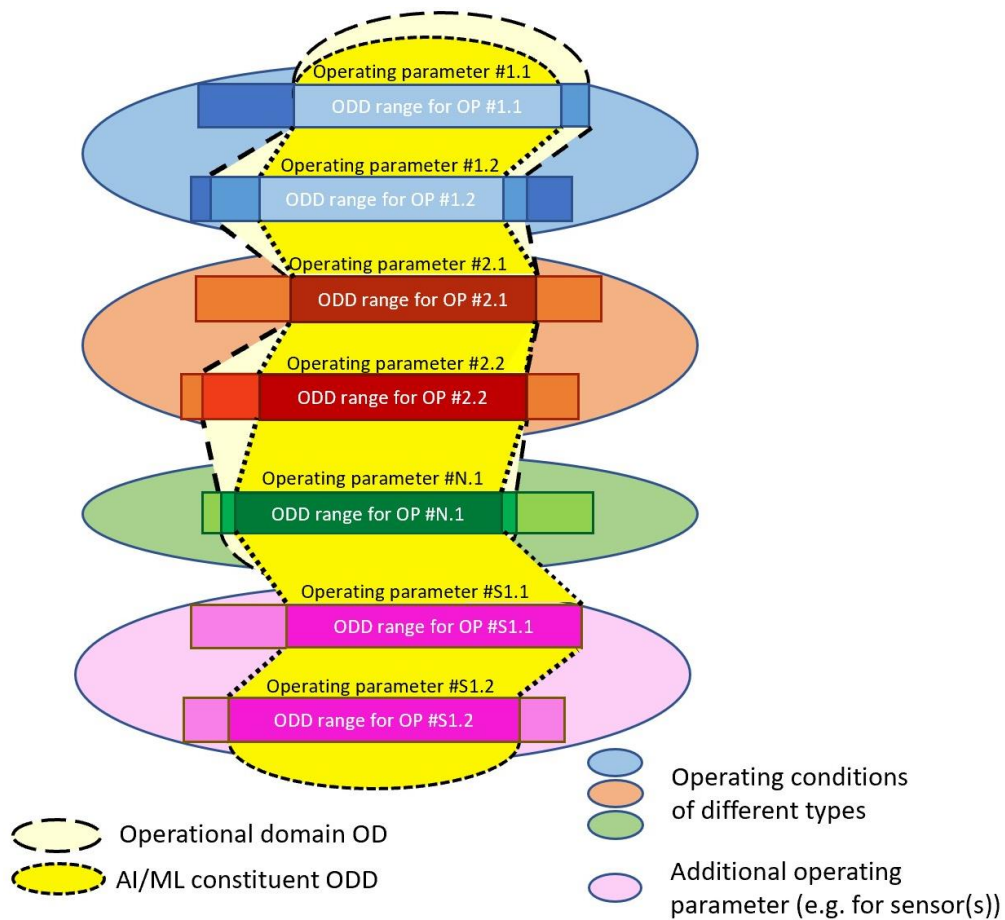


Figure 14 — AI/ML constituent ODD

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.).
- Some operating parameters will need a semantic approach for their definition, especially in high-dimension use cases such as computer vision.
- Ranges for the parameters in the AI/ML constituent ODD can be a subset of the ranges at the level of the operation domain (OD) (see Figure 16 below), limiting the design to an area of the OD where the ML model performance is aligned with the captured requirements (e.g. more stringent weather conditions for the ODD than for the OD of the corresponding (sub)system).
- Exceptionally, one or a 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).

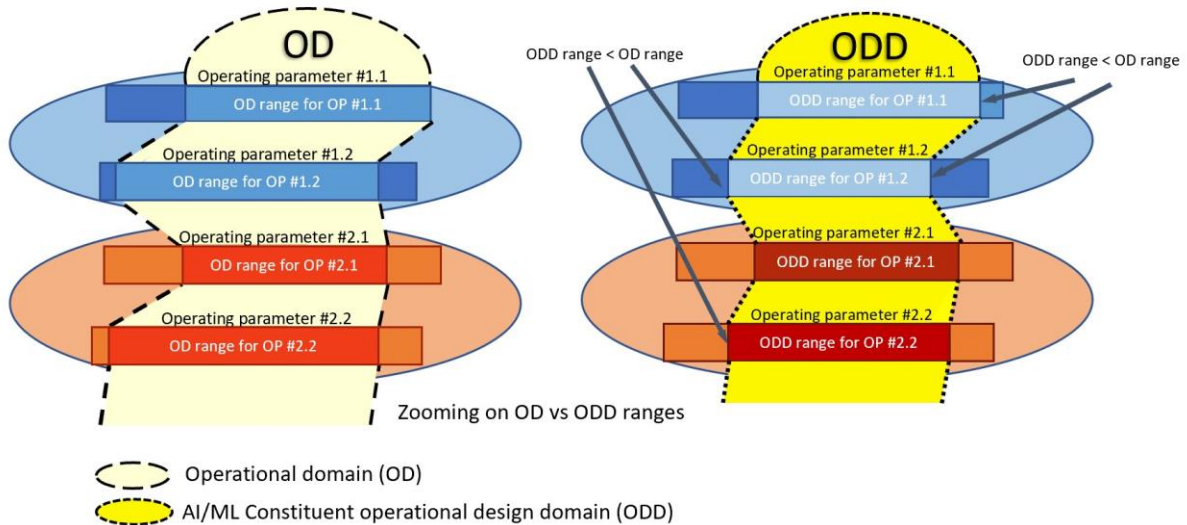
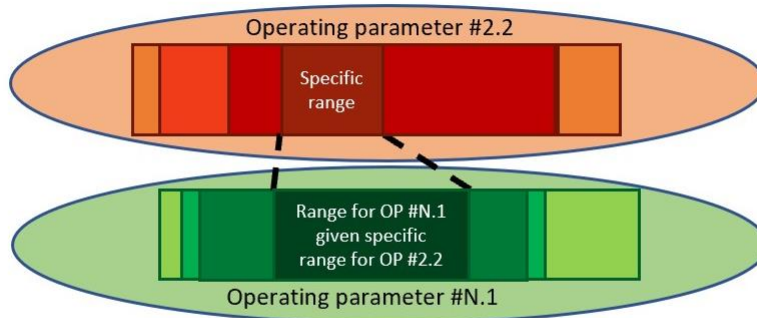


Figure 15 — Definition of ODD ranges versus OD ranges

- 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 16:



Zooming on dependencies between operating parameters

Figure 16 — Relations between operating parameters in ODD

- In relation to the iterative nature of the process aiming at characterising the ODD, stop criteria could be established based on the achievement of some performance requirements of the ML model or the AI/ML constituent.
- In the case of unsupervised learning, characterising the ODD appears to be possibly more challenging (e.g. there is no a priori labelled data to support the identification of any ODD parameter, or the identification of outliers should be carefully studied). Characterising the ODD will likely involve an even more iterative approach than in supervised learning.

Anticipated MOC DA-03: The definition of the parameters pertaining to the AI/ML constituent ODD should be the outcome of relevant industry standards.

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 DA-04: The applicant should capture the DQRs for all data required for training, testing, and verification of the AI/ML constituent, including but not limited to:

- the data relevance 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, processed, or transmitted over a communication network;
- the completeness and representativeness of the data sets; and
- the level of independence between the training, validation and test data sets.

Anticipated MOC DA-04: 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 integrity of the data, i.e. the assurance that it has not been corrupted while stored, processed or transmitted over a communication network (e.g. during data collection);
- the necessary manipulations of the data (e.g. anonymisation);

- 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 relevance 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 the progress in the industry standards development (e.g. EUROCAE/SAE WG-114/G-34) and other best practices (e.g. reference: (DEEL Certification Workgroup, 2021)).

Notes:

- It is anticipated that the DQRs could be more stringent for an AI/ML constituent with higher assurance level. This is, for example, the case for the requirement on the independence of the data sets. Whereas a strict application of the definition is expected (see definition of independence in the context of data management in Section G.1) for an AI/ML constituent at higher-criticality level, this requirement could be relaxed for low-criticality applications (e.g. acceptable ratio of common data between the training/validation data sets and the test data set).
- For what concerns data corruption aspects, specific objectives related to the intentional data corruption (unauthorised alterations of the data sets commonly referred to as ‘data set poisoning’) are provided in the document under Section C.6.1.
- When the origin of the data is external to the applicant (e.g. open-source data or data sourced via a contract established between the applicant and a data provider), the applicant could restrict the stage in the pipeline considered as the origin and clarify how the source has been managed from the origin of the data to the new restricted origin.
- In the case of supervised or unsupervised learning, the ‘learning assurance’ will be based on the use of three separate and independent data sets, also referred to as the training, validation and test data sets.

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

Objective DA-05: The applicant should capture the requirements on data to be pre-processed and engineered for the inference model in development and for the operations.

3.1.2.2. Preliminary AI/ML constituent architecture development

On the basis of the **AI-based (sub)system architecture** and the **AI/ML constituent requirements**, a candidate **AI/ML constituent architecture** development needs to take place. This is not a novel step compared to traditional systems development approaches; it is however an essential step in detailing the AI-based system (or subsystem if applicable) architecture at the layer of the AI/ML constituent.

Objective DA-06: The applicant should describe a preliminary AI/ML constituent 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-79B/ARP-4754B for product certification).

Objective DA-07: The applicant should validate each of the requirements captured under **Objectives DA-02, DA-03, DA-04, DA-05** and the architecture captured under **Objective DA-06**.

Anticipated MOC DA-07: The correctness and completeness of the operating parameters of the AI/ML constituent ODD, as well as their ranges and interdependencies should be reviewed by appropriate subject matter experts in the integration of the affected system and in the development of the AI/ML constituent. The review should also consider elements of the AI/ML constituent ODD that are semantically defined (e.g. weather conditions, period of the year, airspace structure in computer vision use cases). Moreover the review should ensure and maintain consistency between operational domain (including OD and ODD) and the AI/ML constituents functional requirements.

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

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 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 DA-03, Objective DA-04, Objective DA-05, Objective LM-01, Objective LM-02, Objective LM-04, and Objective IMP-01**, and need special attention by the applicant.

Objective DA-08: 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-09: 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

One of the main objectives of the **data management** process is to manage the data and deliver the data sets that will be used during the W-shaped process.

The **data management** process is the first step of the data life cycle management. It identifies objectives to be satisfied at the level of each individual data point, as well as at the level of the data sets that will be used later in the project development lifecycle. Figure 17 below depicts the main activities covered.

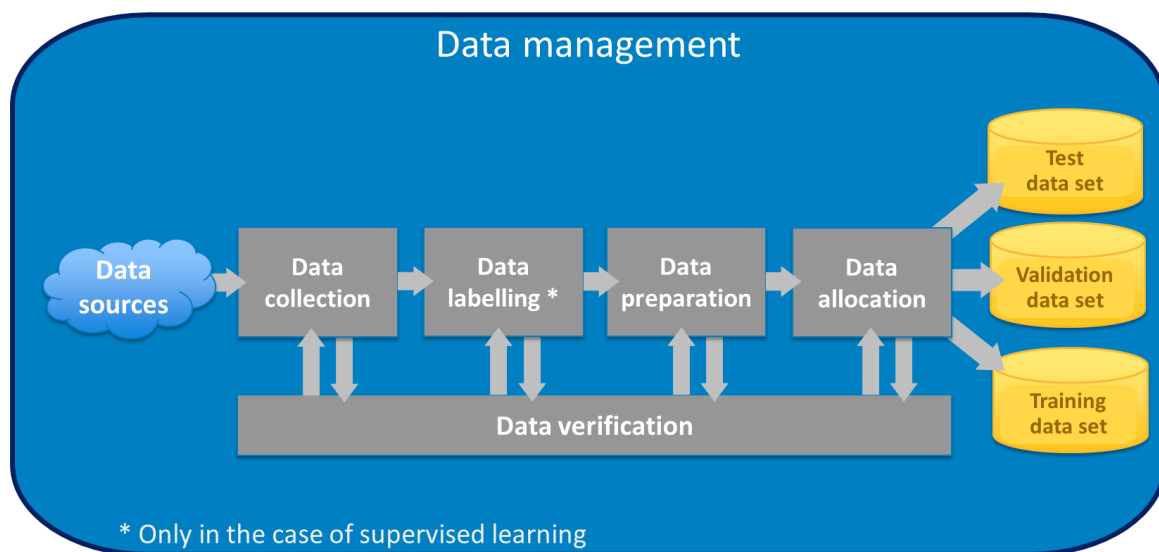


Figure 17 — Data management process

The **data management** process covers:

- data management requirements capture;
- data collection;
- data labelling (only in supervised learning);
- 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 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 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-01: 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 gathered 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).

3.1.3.2. Data labelling in supervised learning

In the context of supervised learning techniques, the data set will need to be annotated or labelled. Please refer to section C.3.1.3.5 as regards unsupervised learning.

Objective DM-02-SL: Once data sources are collected and labelled, the applicant should ensure that the annotated or labelled data in the data set satisfies the DQRs captured under **Objective DA-04**.

All data points 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.

3.1.3.3. 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-03: 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-04: The applicant should define and document pre-processing operations on the collected data in preparation of the model training.

Anticipated MOC DM-04: 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, which are based on **supervised learning**, use the feature learning/extraction capabilities of a convolutional neural network, and do not apply any **feature engineering** step. In the context of **unsupervised learning**, **feature engineering** can also be a valuable tool; however, caution should be exercised in order to avoid introducing any bias into the results.

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-05: 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, depending on the data in the input space, and based on the understanding of the physics of the problem, 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 or combining 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-05-1: In relation with the objective, the applicant should perform a quantitative analysis of the candidate features, applying a dimensionality reduction step when identified as valuable. This step aims at limiting the dimension of the feature space.

Anticipated MOC DM-05-2: In relation with the objective, the applicant should aim at removing multicollinearity between candidate features.

Anticipated MOC DM-05-3: In relation with the objective, if the learning algorithm is sensitive to the scale of the input data, the applicant should ensure that the data is scaled so as to ensure the stability of the learning process.

Data normalisation is one possible MOC 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:

X and X' are the candidate features before and after normalisation,

X_{min} and X_{max} 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.

It is to be noted that data normalisation does not apply to all **supervised learning** 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. In the context of **unsupervised learning**, data normalisation can also be considered. Data normalisation may be applied later during the learning process, when outliers have been managed.

3.1.3.4. Data allocation

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

Objective DM-06: The applicant should distribute the data into three separate data sets which meet the specified DQRs in terms of independence (as per **Objective DA-04**):

- 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 DA-04 and DM-07**).

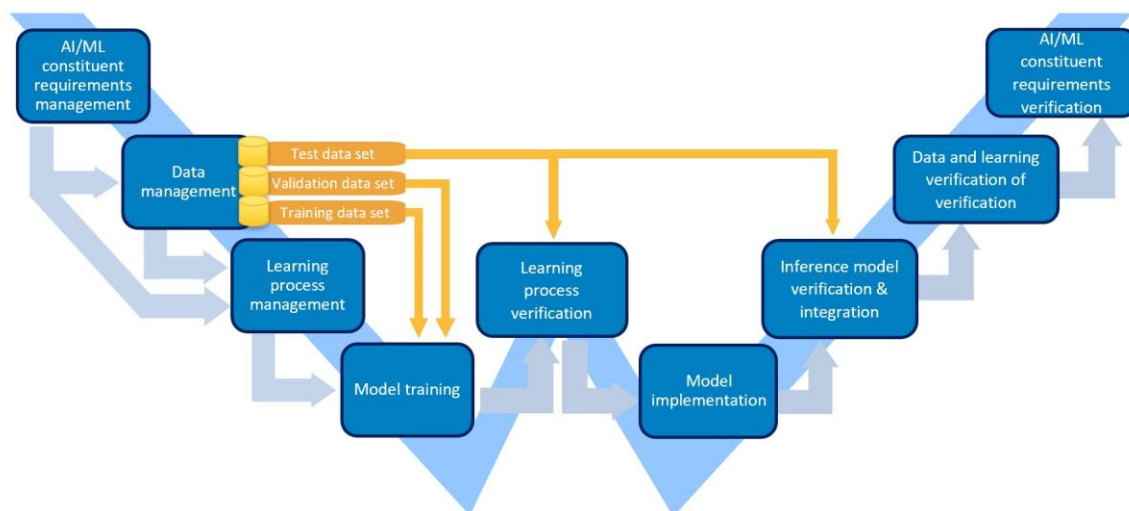


Figure 18 — Training, validation and test data sets usage in W-shaped cycle

3.1.3.5. Data labelling in unsupervised learning

In the context of unsupervised learning techniques, only the test data set will need to be annotated or labelled in order to conduct the verification steps of the learned model.

Objective DM-02-UL: Once data sources are collected and the test data set labelled, the applicant should ensure that the annotated or labelled data in this test data set satisfies the DQRs captured under **Objective DA-04**.

All data points in the test data set 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 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-07: The applicant should ensure verification of the data, as appropriate, throughout 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:

Assessment of the completeness and representativeness of the data sets is a prerequisite to ensure performance on unseen data and to derive generalisation bounds for the trained model.

Anticipated MOC DM-07-1: Data completeness

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

One of the major difficulties in assessing completeness of a data set is to have reliable information about the distributions of phenomena of the intended behaviour in the ODD. Based on the outcomes of the MLEAP Horizon Europe research project, such assessment must be performed on a case-by-case basis, using multiple methods and tools, and in most cases requires extensive expert work and expert judgement. Multiple methods are envisaged 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.

In its first public deliverable, the MLEAP Horizon Europe research project identifies a set of methods and tools in support of the assessment of completeness.

Principal Component Analysis (PCA) (see Section 3.7.2.1 of (EASA, 2023) MLEAP-D2 Interim Public Report – Issue 01) for prior data set analysis is used to gain visual insight on the completeness of a data set by plotting its projection in the low-dimensional space (usually two or three, as it is difficult for humans to interpret visual information in more than three dimensions) computed by the PCA. The data points are expected to homogeneously occupy the entire plot. Any cluster or empty space might be indicative of some form of lack of completeness (i.e. cluster density should be reduced or the data set should be enriched to reach a similar density or, conversely, examples should be added to fill the empty spaces).

For low-dimensionality use cases where feature engineering applies, the ‘graph-based analysis’ (see Section 3.7.2.2 of (EASA, 2023) MLEAP-D2 Interim Public Report – Issue 01) aims at traversing the tree-like graph of feature combination of each sample of the data set. A possible strategy would be to automatically identify the thresholds that would encompass 25 %, 50 %, 75% and 100 % of the data under a given pattern, to offer a synthetic visual tool of the imbalances in a data set, and provide insight on potential completeness shortcomings. Another would be to run the algorithm with dynamic thresholds, to ensure that the data set complies with completeness constraints (that would have to be defined upstream).

In addition to methods that will focus on one data set, other methods could allow the comparison between data sets, ensuring that the characteristics of the data are preserved across the different data sets. The ‘sample-wise similarity analysis’ (see Section 3.7.2.4 of (EASA, 2023) MLEAP-D2 Interim Public Report – Issue 01) could be used to assess the relative representativeness of a data set with regard to another; for example, between a training and a test data set.

It is expected that the final deliverable of the MLEAP Horizon Europe research project will provide additional MOC on completeness of the data set(s), as well as guidelines and criteria on how and when to use these MOC.

Anticipated MOC DM-07-2: 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 (EASA and 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.

In its first public deliverable, the MLEAP Horizon Europe research project identifies a set of methods and tools in support of the assessment of representativeness.

For low-dimensionality use cases where feature engineering applies, the ‘Graph-based analysis’ (see Section 3.7.2.2 of (EASA, 2023) MLEAP-D2 Interim Public Report – Issue 01) aims at traversing the tree-like graph of feature combination of each sample of the data set. A possible strategy would be to automatically identify the thresholds that would encompass 25 %, 50 %, 75% and 100 % of the data under a given pattern, to offer a synthetic visual tool of the imbalances in a data set, provide insight on potential representativeness shortcomings. Another would be to run the algorithm with dynamic thresholds, to ensure that the data set complies with representativeness constraints (that would have to be defined upstream).

Adaptable to any type of data, an ‘entropy-based analysis’ (see Section 3.7.2.3 of (EASA, 2023) MLEAP-D2 Interim Public Report – Issue 01) can be used to characterise the samples in a data set. The main point of attention when using entropy is the type of elements in the data set from which the entropy will be computed, to ensure that the metric provides useful information regarding the overall analysis process. When used appropriately, the ‘entropy-based analysis’ could reveal regions in the data set with such a complexity that it is probably insufficiently represented in the data set.

In addition to methods that will focus on one data set, other methods could allow the comparison between data sets, ensuring that the characteristics of the data are preserved across the different data sets. The ‘sample-wise similarity analysis’ (see Section 3.7.2.4 of (EASA, 2023) MLEAP-D2 Interim Public Report – Issue 01) could be used to assess the relative representativeness of a data set with regard to another; for example, between a training and a test data set.

It is expected that the final deliverable of the MLEAP Horizon Europe research project will provide additional MOC on representativeness of the data set(s), as well as guidelines and criteria on how and when to use these MOC.

Anticipated MOC DM-07-3: 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. Special attention should be paid to the verification of the data labelling that corresponds to the 'ground truth' for the ML model.

Anticipated MOC DM-07-4: 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-07-5: 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 safety-related 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). One exception is already identified with surrogate models where some data sets including the test data set could be exclusively composed of data produced from a high-fidelity model.

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 ML model architecture.

Anticipated MOC LM-01: The applicant should describe the ML model (computational graph) architecture in the planning documentation, including the use of sub-models if any.

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;
- explainability capabilities of the selected model;
- activation functions;
- cost/loss function selection describing the link to the performance metrics;
- model bias and variance metrics and acceptable levels (only in supervised learning);
- 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.

In the context of *unsupervised learning*, establishing some of these requirements beforehand might prove even more challenging than in *supervised learning*.

Anticipated MOC LM-02: 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.

In addition, as part of the learning management requirements, the applicant should confirm that the AI-based system presents no capability of online learning.

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

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 bounds.

Anticipated MOC LM-04: The field of statistical learning theory (SLT) offers means to provide bounds on the capability of generalisation of ML models. As introduced in the CoDANN report (EASA and 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 VC-generalisation bounds have been derived since then.

A good generalisation bound means that the ‘in-sample errors’ (i.e. the errors computed during the development 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:

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

where:

χ 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 χ ,

\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 ε :

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

where:

ε is the generalisation gap tolerance,

δ 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):

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

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 (EASA and 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.

The satisfaction of this objective might prove more challenging with *unsupervised learning*, and this is one of the driving factors for EASA to limit in a first step the applicability of the guidance with this learning approach to applications where AI/ML constituents include IDAL D / SWAL 4 / AL 5 items (see Section B.4).

3.1.5. Model training and learning process validation

The *model 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 *model training phase and its validation* can be repeated iteratively until the trained model reaches the expected performance.

In the context of *unsupervised learning*, the applicant may consider establishing a supervised model that mirrors the one obtained during the learning phase. In such a scenario, the initial step would consist into labelling the different data sets before proceeding with the training of the supervised learning. The resulting data sets and trained model would then be used in support of the satisfaction of some objectives outlined below.

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

Anticipated MOC LM-05: 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: 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-SL: The applicant should account for the bias-variance trade-off in the model family selection and should provide evidence of the reproducibility of the model training process.

Anticipated MOC LM-07-SL: The model family bias and variance should be evaluated using the validation data set. 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 model that minimises error on the validation data set is usually selected as the one that best balances bias and variance. However, simpler models that have higher bias may be selected as long as performance metrics including accuracy are met, since simpler models are usually easier to train, may generalise better (have lower variance), and are easier to explain.

The applicant should identify methods to provide the best possible estimates of the bias and variance of the selected model family; for instance, using holdout data, k-fold sampling or 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, especially when the amount of data is small. Regularisation may not be needed when the amount of data is much larger than the number of model parameters.

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.

In the context of **unsupervised learning**, the **learning process verification** makes extensive use of the test data set labelled in line with **Objective DM-02-UL** (see Section C.3.1.3.5).

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: 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.

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: 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 requirements-based verification of the trained model behaviour.

Anticipated MOC LM-10: 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 (EASA and Daedalean, 2020) Section 6.4.1, perturbations in the development 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, covering the whole AI/ML constituent ODD.

Anticipated MOC LM-12: 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;
- singular points, edge and corner cases.

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: The activity should be supported by test cases, including singular points and 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 MOC 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: Evidence of the validity of the anticipated generalisation bounds proposed to fulfil **Objective LM-04** should be recorded.

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

As already discussed in the context of **Objective LM-04**, the satisfaction of this objective might prove more challenging with *unsupervised learning*, and this is one of the driving factors for EASA to limit in a first step the applicability of the guidance with this learning approach to applications where AI/ML constituents include IDAL D / SWAL 4 / AL 5 items (see Section B.4).

Once the learning process verification is complete, an important step consists in the capture of a final ML model description in preparation of the ML model implementation step.

Objective LM-15: the applicant should capture the description of the resulting ML model.

3.1.7. Model implementation

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

Objective IMP-01: The applicant should capture the requirements pertaining to the ML model 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.3.1.2.1);

- 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 expected tolerances for comparison of the inference model outputs with the trained model outputs;
- requirements pertaining to the development of the inference model into software and/or hardware items, such as processing power, parallelisation, latency, worst-case execution time (WCET), and memory.

Objective IMP-02: The applicant should validate the model description captured under **Objective LM-15** as well as each of the requirements captured under **Objective IMP-01**.

Objective IMP-03: The applicant should document evidence that all derived requirements generated through the model implementation process have been provided to the (sub)system processes, including the safety (support) assessment.

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/place and route (PAR) steps). This implementation follows different steps:

- Model conversion
- Model optimisation
- Inference model development

Objective IMP-04: 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 another format (e.g. 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-04-1: 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, should 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-04-2: Identification of the different optimisation steps performed during implementation and confirmation that no impact on the model behaviour is foreseen, taking into account the expected tolerances (identified per **Objective IMP-01**). 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-05: The applicant should plan and execute appropriate development assurance processes to develop the inference model into software and/or hardware items.

Anticipated MOC IMP-05:

- 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-05** 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-05** is fulfilled. FPGAs, ASICs and COTS architectures are covered by the existing guidance; however, other ML architectures, such as graphics processing units (GPUs and other AI accelerators), have specificities that are not accounted for in the existing guidance (e.g. very complex interference mechanisms or non-deterministic pipelining). Specific hardware architectures like GPUs will require specific guidance to be developed.
- 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-05** 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 **learning process 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, by using the same test data set and obtaining the same results, 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-06: 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-06: As a preliminary step, a set of model properties that are expected to be preserved should be captured. The use of specific verification methods (e.g. formal methods) is expected to be necessary to comply with this objective, taking into account the performance metrics and the expected tolerances (identified per **Objective IMP-01**).

3.1.8.2. Platform verification

Objective IMP-07: The differences between the software and hardware of the platform used for model 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-07: 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-08: 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-08: 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-09: The applicant should perform and document the verification of the stability of the inference model.

Anticipated MOC IMP-09: 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;
- singular points, edge and corner cases.

Objective IMP-10: The applicant should perform and document the verification of the robustness of the inference model in adverse conditions.

Anticipated MOC IMP-10: 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-11: The applicant should perform requirements-based verification of the inference model behaviour when integrated into the AI/ML constituent.

Anticipated MOC IMP-11: 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).

Note: In the context of *unsupervised learning*, the objectives covered under Section C.3.1.8.3 and Section C.3.1.8.4 make extensive use of the test data set labelled in line with **Objective DM-02-UL** (see Section C.3.1.3.5).

3.1.9. Data and learning verification of verification

This process step aims at performing verification that all the data management and learning process steps have been performed, considering all necessary coverage activities. It supports the confidence in the absence of unintended behaviour in the resulting ML model and associated AI/ML constituent.

Note: the absence of unintended function is also reinforced by the development explainability (refer to **Objective EXP-02 and EXP-03**).

The *data management verification of verification* step is meant to close the data management life cycle, by verifying 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-08: The applicant should perform a data verification step to confirm the appropriateness of the defined ODD and of the data sets used for the training, validation and verification of the ML model.

Anticipated MOC DM-08: The associated activities include verification of:

- the correct identification of the input space;
- reassessment of the defined ODD;
- compliance of the data sets (training, validation, test) with the data management requirements;
- coverage of the whole ODD by the test data set, with the necessary level of completeness and representativeness.

The **learning process verification of verification** step is meant to verify that the trained model has been satisfactorily verified, including the necessary coverage analyses. 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 LM-16: The applicant should confirm that the trained model verification activities are complete.

Anticipated MOC LM-16: The associated activities include verification of:

- correctness of requirements-based verification procedures;
- correctness of requirements-based verification results, and justification of discrepancies;
- coverage of the AI/ML constituent requirements by verification methods;
- evaluation of the trained model performance, with a coverage of all pairs of ODD parameters;
- coverage of the whole AI/ML constituent ODD when ensuring stability of the trained model.

The **learning process verification of verification** step is meant to verify that the trained model has been satisfactorily verified, including the necessary coverage analyses. 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 IMP-12: The applicant should confirm that the AI/ML constituent verification activities are complete.

Anticipated MOC IMP-12: The associated activities include verification of:

- correctness of the requirements-based verification procedures;
- correctness of requirements-based verification results, and justification of discrepancies;
- coverage of the AI/ML constituent requirements by verification methods;
- evaluation of the AI/ML constituent performance, with a coverage of all pairs of ODD parameters.

3.1.10. Verification of the AI/ML constituent requirements

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-79B/ARP4754B).

Objective DA-10: Each of the captured AI/ML constituent requirements 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: 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.

3.1.13. Reuse of AI/ML models

Once an initial set of applications will have been approved or certified and introduced into the operations, the applicant may consider the replacement or modification of a previously developed AI/ML constituent or ML model.

Also some applicants may consider incorporating already trained ML models (open source models or COTS ML models) in their design of an AI/ML constituent.

While reusing ML models can offer benefits in terms of efficiency and reduced development time and effort, it also presents challenges, including but not limited to the following:

- OD and/or ODD adaptation;
- data quality;
- change management including open problem reports;
- development explainability;
- scalability;

- documentation and version control.

These challenges require careful consideration and planning when incorporating reused ML models into an AI/ML constituent.

ML models are generally trained for specific tasks on specific data sets aligned with a given ODD. Reusing a model designed for a task and one ODD for a different task or in a different ODD can lead to inaccurate results, weak performance (e.g. generalisation, stability, robustness, etc.).

Objective RU-01: The applicant should perform an impact assessment of the reuse of a trained ML model before incorporating the model into an AI/ML constituent. The impact assessment should consider:

- alignment and compatibility of the intended behaviours of the ML models;
- alignment and compatibility of the ODDs;
- compatibility of the performance of the reused ML model with the performance requirements expected for the new application;
- availability of adequate technical documentation (e.g. equivalent documentation depending on the required assurance level);
- possible licensing or legal restrictions on the reused ML model (more particularly in the case of COTS ML models); and
- evaluation of the required development level.

The outcome of such an impact assessment will provide the applicant with valuable information for the plans to be prepared per **Objective DA-01**. In particular, the plan for learning assurance (e.g. plan for learning aspects of certification) should be tailored to the objectives of the AI/ML constituent.

3.1.13.1. Use of COTS ML models

It is assumed that there are two categories of COTS ML models:

- Category 1: COTS ML models which require to be ‘transformed’ into an inference model;
- Category 2: COTS ML models which include the inference model (e.g. as a library) verified on a target platform.

The following discusses the applicability of the learning assurance objectives, while introducing some new objectives for the specificities of the use of COTS ML models.

Regarding requirements and architecture management (see Section C.3.1.2), **Objective DA-02** is applicable.

Objective RU-02: The applicant should perform a functional analysis of the COTS ML model to confirm its adequacy to the requirements and architecture of the AI/ML constituent.

Objective RU-03: The applicant should perform an analysis of the unused functions of the COTS ML model, and prepare the deactivation of these unused functions.

For example, a complex COTS ML model with a complex model architecture could be used for a limited set of its functionalities. In such a case, all unused functionalities should be deactivated.

Objective DA-03 is applicable and special attention should be paid to ensure that the COTS ML model operating conditions (as part of the COTS documentation) are accounted for in the ODD definition at constituent level. **Objective DA-04** is applicable when establishing requirements applying to the test data set. **Objective DA-05** up to **Objective DA-10** are applicable.

The data management process (see Section C.3.1.3) should be executed in order to deliver a test data set to be used during the trained model verification and the inference model verification and integration. **Objective DM-01** up to **Objective DM-05** apply on the future test data set. **Objective DM-06** does not apply as only a test data set is to be considered. **Objective DM-07** is applicable.

Regarding the learning management (see Sections C.3.1.4, C.3.1.5, C.3.1.6), **Objective LM-01** is applicable. **Objective LM-02** is not applicable. From **Objective LM-03** up to **Objective LM-08**, as well as for **Objective LM-11**, credit should be taken from the COTS model documentation; should the COTS model documentation be insufficient to demonstrate satisfaction of these objectives, then they should be applicable. **Objective LM-09**, **Objective LM-10** and **Objective LM-12** up to **Objective LM-15** are applicable, except for COTS ML models of category 2.

Applicability of **Objective IMP-01** up to **Objective IMP-05** is limited to the category 1 COTS ML models, requiring transformations of the COTS ML model to adapt to the target platform. **Objective IMP-06** up to **Objective IMP-11** are applicable.

Objective CM-01 applies. Specific attention should be paid to the versioning of the ML model itself and its supporting documentation.

Objective QA-01 is applicable. Specific attention should be paid to the quality assurance aspects related to the COTS ML model.

Development explainability objectives apply as well. Regarding **Objective EXP-03**, credit should be taken from the COTS ML model documentation. Should the COTS model documentation be insufficient to demonstrate satisfaction of the objective, then other means should be envisaged for its satisfaction.

3.1.13.2. Transfer learning

In supervised learning, *transfer learning* refers to the process of adapting an ML model that has already been trained on one task to perform a new but often related task. Transfer learning can be very useful when the new or adapted task has a limited amount of data available, or when the ML model that has been trained for the original task has demonstrated to be performing well, with the expectation that the performance will continue to be met for the new or adapted task. An ML model that has been trained for the original task is used as a starting point and is re-trained with additional training data to fine-tune the ML model for the new or adapted task.

Nowadays, *transfer learning* is commonly used in computer vision and in natural language processing. In particular, DL architectures facilitate the deployment of *transfer learning* as the early layers of the network can learn low-level features, like detecting edges, colours, variations of intensities, etc. These kind of features might not be specific to a particular data set or a task. It is then the role of the final layers of the network to learn the specificities of the task. In general, the hyperparameters of the early layers are frozen (e.g. by setting the gradients to zero before the training starts) when training for the new task.

Good data management practices remain crucial for the success of a project deploying *transfer learning* and incorporating the resulting ML model into an AI/ML constituent.

The following discusses the applicability of the learning assurance objectives, while confirming the applicability of some new objectives for the specificities of *transfer learning*.

Regarding requirements and architecture management (see Section C.3.1.2), **Objective DA-02** is applicable.

Objectives RU-02 and **RU-03** are applicable to the model that is used as a basis for the *transfer learning*.

Objective DA-03 is applicable and special attention should be paid to adapted or new ODD parameters compared to the ODD used for the original ML model. **Objective DA-04** up to **Objective DA-10** are applicable.

The data management process (see Section C.3.1.3) should be executed in order to deliver the data sets to be used during the rest of the learning assurance phases. In this respect, **Objective DM-01** up to **Objective DM-07** apply with possible exemption for **Objective DM-05**.

Similarly the learning management process (see Sections C.3.1.4, C.3.1.5, C.3.1.6) should be executed and **Objective LM-01** up to **Objective LM-15** are applicable.

All objectives of the model implementation process (see Section C.3.1.7) are applicable, i.e. from **Objective IMP-01** up to **Objective IMP-11**.

Objective CM-01 applies. Specific attention should be paid to the management of the documentation and of the versioning of the ML model used as an input to the application of the *transfer learning* approach.

Objective QA-01 is applicable. Specific attention should be paid to the quality assurance aspects related to the ML model used as an input to the application of the *transfer learning* approach.

Development explainability objectives apply as well. Regarding **Objective EXP-03**, credit should be taken from the explainability aspects of the ML model used as an input to the application of the *transfer learning* approach. These should be complemented to account for the learned ML model.

3.1.13.3. Previously developed models

Preliminary note: modifications to previously developed models is not warranted.

Change of installation

An ML model incorporated into an AI/ML constituent and AI-based (sub)system that has been approved or certified in a specified OD and/or ODD may be used in a changing domain. For instance, a function approved for a specified airport may need to be approved for another airport. Or a conflict resolution solution approved for a specified ATC centre may need to be approved for another ATC centre (even inside the same ANSP).

Change of application

An ML model incorporated into an AI/ML constituent and AI-based (sub)system that has been approved or certified for a specified function may be reused for a changing function. For instance, an

ML model capable of detecting only aircraft in a detect and avoid function could be reused to detect drones.

Change of development environment

An ML model incorporated into an AI/ML constituent and AI-based (sub)system that has been approved or certified based on a given development environment may be subject to a change of development environment. For instance, an ML model developed under a certain learning framework version may be updated using a more recent version of the framework.

Change of inference platform

An ML model incorporated into an AI/ML constituent and AI-based (sub)system that has been approved or certified for a specified inference platform may be reused on a new inference platform, including a possible evolution of the hardware/software architecture of the inference model.

Analysis of the impacts on the applicability of the learning assurance objectives

The table below identifies the applicability of the learning assurance objectives for each type of change described above.

Objectives	Applicability and specificities in objectives			
	Change of installation	Change of application	Change of development environment	Change of inference platform
Objective DA-02	DA-02 is applicable	DA-02 is applicable	DA-02 is applicable	DA-02 is applicable
Objective DA-03	DA-03 is applicable. The focus should be on the variations in the ODD boundaries (a priori no new parameter is expected).	DA-03 is applicable. Special attention should be paid to adapted or new ODD parameters compared to the ODD used for the original application.	No change on ODD	No change on ODD
Objective DA-04	No change in DQRs	DA-04 applicable with a focus on adapted or new ODD parameters.	No change in DQRs	No change in DQRs
Objective DA-05	No change expected	DA-05 is applicable	No change expected	No change expected
Objective DA-06	No change expected	DA-06 is applicable	No change expected	DA-06 is applicable based on the new architecture envisaged for the inference platform.
Objective DA-07	No requirement validation expected	DA-07 is applicable	No requirement validation expected	No requirement validation expected
Objective DA-08	No derived requirements expected	DA-08 is applicable	No derived requirements expected	DA-08 is applicable
Objective DA-09	No derived requirements expected	DA-09 is applicable	No derived requirements expected	DA-09 is applicable
Objective DA-10	DA-10 is applicable	DA-10 is applicable	DA-10 is applicable	DA-10 is applicable
Objective DM-01	DM-01 is applicable to cope with the variations in the ODD.	DM-01 is applicable	No additional data collection expected	No additional data collection expected
Objective DM-02-SL	DM-02-SL is applicable	DM-02-SL is applicable	No data labelling expected	No data labelling expected
Objective DM-03	DM-03 is applicable	DM-03 is applicable	No data preparation operations expected	No data preparation operations expected
Objective DM-04	DM-04 is applicable	DM-04 is applicable	No data preparation operations expected	No data preparation operations expected

Objectives	Applicability and specificities in objectives			
	Change of installation	Change of application	Change of development environment	Change of inference platform
Objective DM-05	DM-05 applicability is unchanged	DM-05 applicability is unchanged	No data preparation operations expected	No data preparation operations expected
Objective DM-06	DM-06 is applicable	DM-06 is applicable	No distribution of the data into separate data sets expected	No distribution of the data into separate data sets expected
Objective DM-02-UL	DM-02-UL is applicable	DM-02-UL is applicable	No data labelling expected	No data labelling expected
Objective DM-07	DM-07 is applicable	DM-07 is applicable	No data verification expected	No data verification expected
Objective DM-08	DM-08 is applicable	DM-08 is applicable	DM-08 is applicable	DM-08 is applicable
Objective LM-01	No change expected	LM-01 is applicable	LM-01 is applicable	No change expected
Objective LM-02	No change expected	LM-02 is applicable	LM-02 is applicable	No change expected
Objective LM-03	No change expected	LM-03 is applicable	LM-03 is applicable	No change expected
Objective LM-04	LM-04 is applicable	LM-04 is applicable	LM-04 is applicable	No change expected
Objective LM-05	LM-05 is applicable	LM-05 is applicable	LM-05 is applicable	No change expected
Objective LM-06	LM-06 is applicable	LM-06 is applicable	LM-06 is applicable	No change expected
Objective LM-07	LM-07 is applicable	LM-07 is applicable	LM-07 is applicable	No change expected
Objective LM-08	LM-08 is applicable	LM-08 is applicable	LM-08 is applicable	No change expected
Objective LM-09	LM-09 is applicable	LM-09 is applicable	LM-09 is applicable	No change expected
Objective LM-10	LM-10 is applicable	LM-10 is applicable	LM-10 is applicable	No change expected
Objective LM-11	LM-11 is applicable	LM-11 is applicable	LM-11 is applicable	No change expected
Objective LM-12	LM-12 is applicable	LM-12 is applicable	LM-12 is applicable	No change expected
Objective LM-13	LM-13 is applicable	LM-13 is applicable	LM-13 is applicable	No change expected
Objective LM-14	LM-14 is applicable	LM-14 is applicable	LM-14 is applicable	No change expected
Objective LM-15	LM-15 is applicable	LM-15 is applicable	LM-15 is applicable	No change expected
Objective IMP-01	No change expected	IMP-01 is applicable	IMP-01 is applicable	IMP-01 is applicable
Objective IMP-02	IMP-02 is applicable	IMP-02 is applicable	IMP-02 is applicable	IMP-02 is applicable
Objective IMP-03	IMP-03 is applicable	IMP-03 is applicable	IMP-03 is applicable	IMP-03 is applicable
Objective IMP-04	IMP-04 is applicable	IMP-04 is applicable	IMP-04 is applicable	IMP-04 is applicable
Objective IMP-05	IMP-05 is applicable	IMP-05 is applicable	IMP-05 is applicable	IMP-05 is applicable
Objective IMP-06	IMP-06 is applicable	IMP-06 is applicable	IMP-06 is applicable	IMP-06 is applicable
Objective IMP-07	IMP-07 is applicable	IMP-07 is applicable	IMP-07 is applicable	IMP-07 is applicable
Objective IMP-08	IMP-08 is applicable	IMP-08 is applicable	IMP-08 is applicable	IMP-08 is applicable
Objective IMP-09	IMP-09 is applicable	IMP-09 is applicable	IMP-09 is applicable	IMP-09 is applicable
Objective IMP-10	IMP-10 is applicable	IMP-10 is applicable	IMP-10 is applicable	IMP-10 is applicable
Objective IMP-11	IMP-11 is applicable	IMP-11 is applicable	IMP-11 is applicable	IMP-11 is applicable
Objective CM-01	CM-01 is applicable. Specific attention should be paid to the change control related to the original ML model.	CM-01 is applicable. Specific attention should be paid to the change control related to the original ML model.	CM-01 is applicable	CM-01 is applicable
Objective QA-01	QA-01 is applicable. Specific attention should be paid to the quality assurance aspects related to the original ML model.	QA-01 is applicable. Specific attention should be paid to the quality assurance aspects related to the original ML model.	QA-01 is applicable	QA-01 is applicable
Objective EXP-01	No change expected	EXP-01 is applicable	No change expected	No change expected
Objective EXP-02	No change expected	EXP-02 is applicable	No change expected	No change expected
Objective EXP-03	EXP-03 is applicable	EXP-03 is applicable	EXP-03 is applicable	No change expected
Objective EXP-04 to EXP-09	EXP-04 to EXP-09 are applicable	EXP-04 to EXP-09 are applicable	EXP-04 to EXP-09 are applicable	EXP-04 to EXP-09 are applicable

Table 4 — Applicability of learning assurance objectives to changes

Notes:

- All the cells with ‘No ...’ mean that credit could be taken from the satisfaction of the objective via lifecycle data on the previous project;
- **Objectives DM-08, LM-16 and IMP-12** are systematically applicable, as they are ‘verification of verification’ objectives.

3.1.13.4. Upgrading a development level

This situation is mainly triggered when an analysis at system level, including a safety (support) assessment, requests that the AI/ML constituent be developed at a higher development level. For instance, in the airworthiness domain an AI/ML constituent previously developed at DAL D should now satisfy the learning assurance objectives at DAL C; or in the ATM/ANS domain, an AI/ML constituent previously developed at SWAL 4 should now satisfy the learning assurance objectives at SWAL 3. As part of the impact assessment (**Objective RU-01**), the applicant should evaluate which additional objectives compared to the ones that applied to the previous development become applicable with the new development level (information in Chapter D could serve as an input for such an evaluation).

The applicant should take advantage of the lifecycle data of the previous development when it already satisfies certain objectives requested by the new development level.

Also lifecycle data from the previous development should be evaluated to ensure that the learning process verification objectives, the inference model verification and the AI/ML integration objectives are satisfied for the new development level and the required level of rigour.

Reverse engineering may be used to generate lifecycle data that is inadequate or missing to satisfy the additional objectives requested by the new development level.

3.1.14. Surrogate modelling

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.

The objectives contained in Sections C.3.1.1 to C.3.1.13 of this document fully apply to surrogate ML models. Beyond this, specific aspects to be considered are developed in the rest of this section.

3.1.14.1. Considerations on definitions of validation and verification

In this document, the terms ‘validation’ and ‘verification’ have been considered in the context of system development assurance (e.g. ED-79B/ARP4754B for IAW). As surrogate modelling is used also for structural applications, it is important to consider commonly agreed definitions in this specific context (contained in ASME VVUQ 1²⁰).

²⁰ ASME Verification, Validation, and Uncertainty Quantification Terminology in Computational Modeling and Simulation, VVUQ 1 – 2022.

Verification: the process that establishes the mathematical correctness of the computational model with respect to a reference model

Validation: the process of determining the degree to which a model represents the empirical data from the perspective of the context of use

Although all definitions aim at the same goals ('Did we build the right item?' for validation and 'Did we build the item right?' for verification), the emphasis in systems development assurance is more on the requirements, whereas the emphasis in structures is more on accuracy.

These differences should be kept in mind when discussing V&V topics; however, the V&V objectives defined in this document (Sections C.3.1.2.3, C.3.1.3.6, C.3.1.5, C.3.1.6, C.3.1.8, C.3.1.9, C.3.1.10) are to be applied.

3.1.14.2. Considerations on the determination of criticality level

The determination of criticality level is driven by the objectives related to the safety assessment under Section C.2.2, in particular through MOC SA-01-1. However in some domains, a safety assessment process might not be guided: for those domains, the intent is to define a generic layer of safety assessment objectives, including considerations on criticality level determination.

3.1.14.3. Considerations on software assurance

Software development assurance processes are driven by the objectives related to implementation of the ML model into software (Section C.3.1.7), and in particular IMP-05. However, in some domains, software development assurance might not be guided: for those domains, the intent is to draw a link to existing 'tool qualification' industry standards (such as ED-215/DO-330) in the next Issue of this document. Based on the criticality level determination, the applicable objectives will be modulated through a tailored Tool Qualification Levels table to be developed also in a future Issue.

3.1.14.4. Considerations specific to surrogate modelling

Surrogate models are typically models that approximate the simulation output of more complex, higher-fidelity models that are, for instance, physics based (i.e. based on first principles). A surrogate model is computationally less expensive to run than the associated high-fidelity model, and is therefore more suitable to perform sensitivity analyses, design optimisations, risk analyses or uncertainty quantification.

Considering the existence of a high-fidelity reference model, the use of surrogate models raises specific concerns beyond those addressed for ML models in the generic set of learning assurance objectives from Sections C.3.1.1 to C.3.1.13. Thus, the two following additional objectives are identified as necessary:

Objective SU-01: The applicant should capture the accuracy and fidelity of the reference model in order to support the verification of the accuracy of the surrogate model.

In some cases the simulation output of the high-fidelity model would be used to derive the required analytical results and the surrogate model is, for example, only used for design optimisation. In other cases the surrogate model completely replaces the high-fidelity model in the simulation process and the output of the surrogate model defines the required results. Consequently, the level of accuracy of

a surrogate model should be at the same level as the credibility of the (high-fidelity) physics-based model, or be compensated for.

Moreover, the training (and test and validation) data set for the (data-driven) surrogate model is delivered by the (typically physics-based) high-fidelity model. Within the design space, a number of locations (sampling) are chosen where the high-fidelity model defines the input-output relationship that constitutes this training set. This sampling process is known as Design of Experiments (DOE).

Objective SU-02: the applicant should identify, document and mitigate the additional sources of uncertainties linked with the use of a surrogate model.

The management of uncertainties is already addressed under MOC SA-01-4 and SA-01-5. Nevertheless some additional sources of uncertainties may be specifically triggered by the use of surrogate modelling. Indeed, the use of a surrogate model may introduce an additional uncertainty; for example, if it replaces a physics-based model, as the ML model is then ‘another step away from reality’.

In addition, strong discontinuities or non-linearities in the design space may pose a challenge to properly define a surrogate model.

Furthermore, calibration is seen as an important element to address the overall uncertainty quantification of surrogate models. An ML model is considered to be calibrated if it produces calibrated probabilities²¹.

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.

²¹ [A guide to model calibration | Wunderman Thompson Technology \(wttech.blog\)](#)

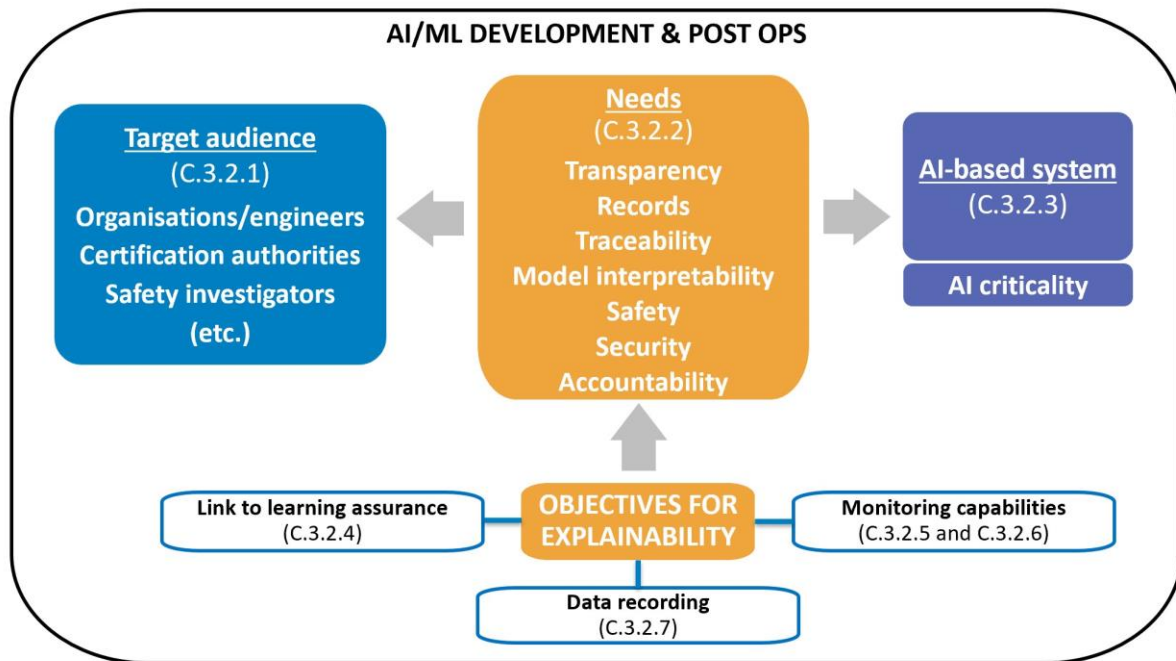


Figure 19 — Development & post-ops explainability view

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.3.1.13 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 behaviour 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 bias in the trained and/or inference model; and
- Absence of unintended behaviours.

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 (EASA and 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 related to the level of confidence in the AI/ML constituent output

Objective EXP-04: The applicant should design the AI-based system with the ability to deliver an indication of the level of confidence in the AI/ML constituent output, based on actual measurements or on quantification of the level of uncertainty.

3.2.6. Specific objectives related to the ODD monitoring and performance monitoring provisions

As mentioned in Section C.3, learning assurance aims at ensuring the intended function of the AI-based system 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 operating conditions remain within acceptable ODD boundaries (both in terms of input parameter range and distribution) and the performance is aligned with the expected level.

The feedback of this monitoring is a possible contributor to the operational AI explainability guidelines, as described in Section C.4.1.4.2.

The following objectives are anticipated:

Objective EXP-05: The applicant should design the AI-based system with the ability to monitor that its inputs are within the specified ODD boundaries (both in terms of input parameter range and distribution) in which the AI/ML constituent performance is guaranteed.

Objective EXP-06: The applicant should design the AI-based system with the ability to monitor that its outputs are within the specified operational AI/ML constituent performance boundaries.

Objective EXP-07: The applicant should design the AI-based system with the ability to monitor that the AI/ML constituent outputs (per **Objective EXP-04**) are within the specified operational level of confidence.

Anticipated MOC EXP-07: 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 level of confidence in its outputs. Such an indication should be part of the elements provided within the explanations as needed.

Objective EXP-08: The applicant should ensure that the output of the specified monitoring per the previous three objectives are in the list of data to be recorded per **MOC EXP-09-2**.

3.2.7. 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 three distinct types of use should be addressed:

- Data recording for the purpose of monitoring the safe usage of the AI-based system in operations (as part of safety management of the end-user organisation)
 - 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 performed by (or on behalf of) the organisation using the AI-based system.
 - This monitoring is meant to be part of the safety management system (SMS) of the organisation using the AI-based system.
 - An example of recording and using data for monitoring operational safety is the collection of parameters from the AI-based system through a flight data monitoring (FDM) programme in order to evaluate the use of the system by the end user. An FDM programme is required for some categories of large aeroplanes and it must be part of the operator's SMS. An FDM programme does not require a crash-protected recorder.

- Data recording for the purpose of the continuous safety assessment by the applicant
 - 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 interaction with human users or other systems.
 - This monitoring 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.
 - This monitoring is performed by (or on behalf of) the applicant of the product embedding the AI-based system.
 - An example of recording and using data for the purpose of the continuous safety assessment, is the evaluation of possible drift in the distribution of AI-based system inputs in operations, compared to the initial ODD assumptions, that would impact the generalisation capabilities of the system. Continuous safety assessment processes do not require a crash-protected recorder.
- 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.
 - An example of data recording for the purpose of accident or incident investigation is the crash-protected flight recorders (flight data recorder and cockpit voice recorder), which must be fitted to large aeroplanes and large helicopters.

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-09: 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, as well as the means to retrieve this data.

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

Anticipated MOC EXP-09-1: 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.

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

This section provides anticipated MOC for the monitoring of the safe usage the of AI-based system (for the organisations of end users), as well as for the continuous safety assessment (for applicants).

Anticipated MOC EXP-09-2: 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 during the deviation, and to the extent possible, before the deviation;
- (c) for monitoring trends regarding deviations over longer periods of time.

Anticipated MOC EXP-09-3: The means to retrieve the recorded data should be provided 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 the recorded data over longer periods of time and possibility to automatically process part of this data for trend analyses and statistical studies.

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

This section provides anticipated MOC for the purpose of accident or incident investigation. **Anticipated MOC EXP-09-4:** 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 HAT 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-09-5: 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) if the AI-based system is airborne, a crashworthy memory medium on board the aircraft;
- (b) recording technology that is reliable and capable of retaining data for long periods of time without electrical power supply;
- (c) if the AI-based system is airborne, means to facilitate the retrieval of the data from the memory medium 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.OR.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 (SESAR JU, 2018), 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
- Customisation of human-AI interface

Note: In this Section C.4, the use of the wording ‘the applicant should design’ is to be understood as ‘the applicant should ensure that the AI-based system is designed’, in case the applicant is not the designer of the AI-based system or AI/ML constituent or underlying ML models.

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 validity. 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.

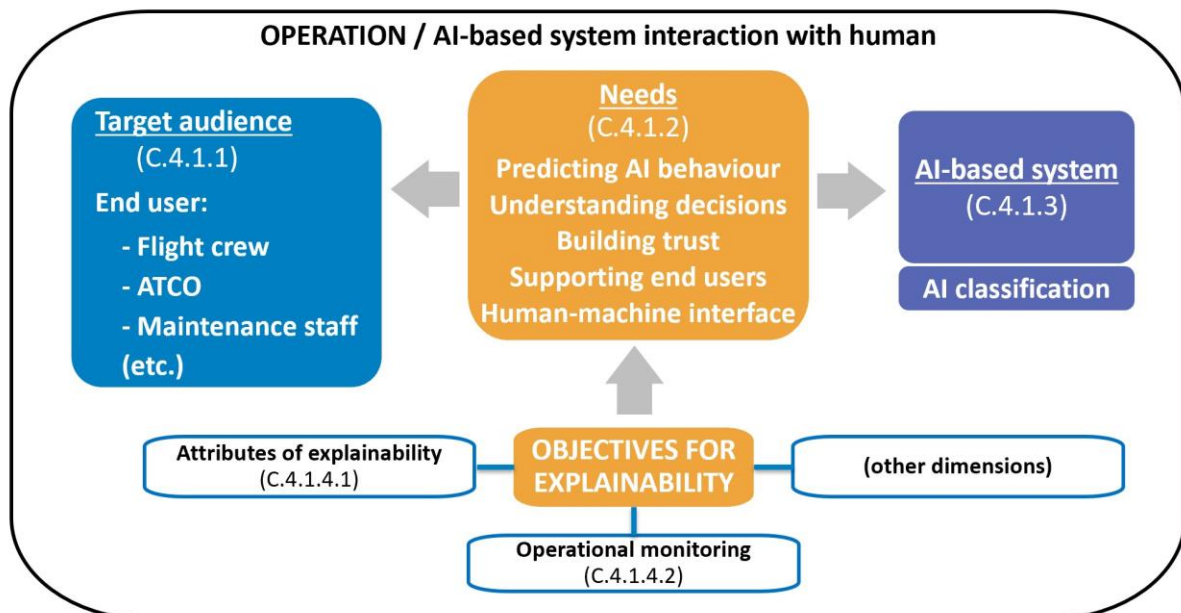


Figure 20 — Operational explainability view

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 mainly affect the task allocation scheme by progressively

giving more authority to the AI-based systems. 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 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 is the one expected by the end user
- 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, system 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 abstraction of 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.

	OVERALL IMPACT ASSESSMENT	HAI	EXPLAINABILITY	GUIDANCE
		Expected level of evolution in the human-AI interaction (HAI) compared to existing interactions	Expected level of explainability needed during operation	Need for specific human factors certification guidance linked with the introduction of AI-based systems
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 abstraction of 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 HAI 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 5 — 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-10: 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-11: The applicant should ensure that the AI-based system presents explanations to the end user in a clear and unambiguous form.

Anticipated MOC EXP-11: 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-12: 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-12: The explanation of a system output is relevant if 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 criteria 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 the level of confidence in their outputs may vary, it is fundamental that both pilots are aligned using the same sources of data.

Level of abstraction

Objective EXP-13: 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-13: 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 more time-critical situations, the end user will require concise explanations to efficiently understand and follow the actions and decisions of the AI-based system. 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 abstraction of an explanation and the cognitive cost seems to be essential to maintain an efficient HAI.

Objective EXP-14: Where a customisation capability is available, the end user should be able to customise the level of abstraction as part of the operational explainability.

Anticipated MOC EXP-14: 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 detail 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. This may be possible as pre-setting by the end user. 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 be adapted based on context-sensitive mechanisms. The AI-based system will have the capabilities to adapt to its environment in a predefined envelope set by design.

Timeliness of explainability

Objective EXP-15: 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-16: 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-15 & EXP-16: 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 situation 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 (for acknowledgement or not) 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 situation awareness of the end users.

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

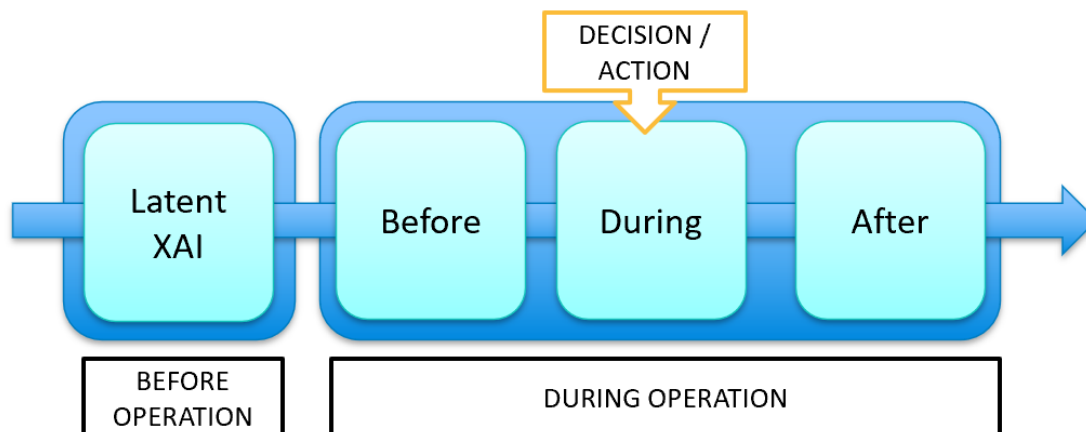


Figure 21 — Timeliness of the explainability

Validity of the explanation

Objective EXP-17: For each output relevant to the task(s), the applicant should ensure the validity of the specified explanation.

4.1.4.2. Objectives related to the monitoring of the ODD and of the output confidence in operations

As mentioned in Section C.3.2.6, an important objective is to monitor whether or not the operational input conditions remain within acceptable boundaries and the level of confidence in the output 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-18: The training and instructions available for the end user should include procedures for handling possible outputs of the ODD monitoring and output confidence monitoring.

Objective EXP-19: Information concerning unsafe AI-based system operating conditions should be provided to the 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 situation awareness and to readjust strategies and task allocation in real time. Communication is paramount to share valuable information needed to achieve the goal.

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 situation awareness;
- Identification of abnormal situation and performance of diagnostics;
- Evaluation of the relevance of the solution proposed by the end user;
- Negotiation/argumentation;

— Adaptivity.

Note: While AI-based systems supporting cooperation (AI Level 2A) are likely to be at a level of complexity commensurate with Level 1 AI-based systems, the objectives of this section, which are applicable to AI-based systems supporting collaboration (AI Level 2B), have been defined with the idea of a holistic support system (system of systems) in mind.

Note: The objectives HF-xx generally use the formulation ‘The applicant should design the AI-based system’, without consideration on the industrial organisation. The fulfilment of this objective is transferable to sub-tier suppliers as necessary, with the responsibility of meeting the objective remaining with the applicant.

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

Anticipated MOC HF-01

Situation representation refers to the machine equivalent of situation awareness. To avoid personification of AI, the term ‘situation representation’ is used to refer to ‘the collection of the environment and system state as well as of the state of the end user, processing of this information, with the aim of enabling extrapolation of a target status in the near future’.

Computer-based systems can collect, analyse, fuse, report on, monitor or store data from multiple sources at a rate that is significantly greater than that of the end user. An AI-based system therefore has the potential to form a representation of a situation that is significantly more detailed, reactive and thorough than that of an end user. The situation representation of an AI-based system should therefore form a valuable resource within an operational environment.

The applicant should ensure that the representation of the system state and the environment within which the system is operating form the basis of the interaction with both the system and the end users of the system. The number of parameters to be monitored will be driven by the proposed application area of the AI-based system and the proposed allocation scheme / pattern.

The applicant should ensure that for each potential allocation pattern, the AI-based system has access to relevant and appropriate system parameters to allow the development of system representation. Access to that data over a sufficient duration to allow and accurate situation representation to be formed should also be ensured.

The applicant has also to ensure that the AI-based system is hosted on a platform that has sufficient processing capacity to deal with the number of data points to be collected, collated and analysed in the timescales that the end user would deem acceptable.

Objective HF-02: The applicant should design the AI-based system with the ability to reinforce the end-user individual situation awareness.

Anticipated MOC HF-02: Situation awareness should be reinforced using appropriate means; for example, conversational interface or visualisations using appropriate modalities.

As stated in Anticipated MOC HF-01, an AI-based system will have the capability to monitor, simultaneously, more system parameters than the end user. The AI-based system could, predictably, have a greater awareness of a developing or rapidly changing situation than the end user.

The end user will develop situation awareness by analysing system parameters. As the AI-based system will analyse multiple systems more rapidly than the end user, it is not unreasonable to expect the end user to refer to the AI-based system to reinforce their own situation awareness.

While an AI-based system has the ability to monitor multiple parameters at rapid speed and analyse the data, it is not always necessary for the end user to analyse all possible data to develop an appropriate and adequate situation awareness. The system must therefore be able to exchange with the end user on data that is relevant to the specific situation or subject that the end user requires.

The system should therefore be designed to be sensitive to:

- information relevant to the phase of operation;
- information requested by the end user to support a current or future task;
- information that is relevant to the end user based on tasks being performed by the end user.

Given the broad potential range of parameters and interaction means available to an AI-based system, situation awareness should be reinforced using appropriate means. Applicants should ensure that the AI-based system can present information in a format that supports the end user's need to reinforce his or her situation awareness. These might include but are not limited to conversational/natural language interfaces or visualisations.

Objective HF-03: The applicant should design the AI-based system with the ability to enable and support a shared situation awareness.

MOC HF-03

Human-AI shared situation awareness refers to the collective understanding and perception of a situation, achieved through the integration of human and AI-based system capabilities. It involves the ability of both humans and AI systems to gather, process, exchange and interpret information relevant to a particular context or environment, leading to a shared comprehension of the situation at hand. This shared representation enables effective collaboration and decision-making between humans and AI based systems.

The applicant should ensure that shared situation awareness reflects that of the end user.

Moreover, the applicant should design the AI-based system with the ability to modify its individual situation awareness on end user request.

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

Anticipated MOC HF-04: In current operations both in the air and in the ground a ‘two man rule’ is often used to reduce the likelihood of error. This requires certain decisions or actions about to be made to be cross-checked by another user. This cross-check, apart from reducing the likelihood of error, also has the effect of ensuring that the second user remains aware of the action sequence being performed by the first user. In current operations, this modus operandi is embedded in procedures and operating manuals, and has safety and performance outcomes.

Within the scope of AI-based systems, a similar requirement exists. Within the scope of AI Level 2A, AI-based systems are required to provide relevant and timely feedback to the end user that allows for the opportunity to double check or veto any decision/action taken by the AI-based system.

The system applicant should develop the system and the means of interaction that:

- provide information to the end user in a timely manner;
- allow for timely intervention by the end user given the amount of information provided to the end user and the critical nature of the action proposed;
- ensure that the end user maintains situation awareness following this integration of cross-checks .

The AI-based system should allow for multiple modes of interaction with the end user and select the one that is most appropriate given the situation when the cross-check or validation is required. This might be visual when auditory channels are loaded or tactile when auditory channels are occupied.

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 improved solution.

Corollary objective HF-05: The applicant should design the AI-based system with the ability to process and act upon a proposal rejection from the end user.

Complex situations from the end user perspective can be those associated with high workload, and stress can result in cognitive tunnelling. In this situation the end user becomes overly focused on one solution or path of action and may not have sufficient capacity to consider alternative solutions or actions. End users’ fixation on one potential solution can result in workload peaks being maintained, more complex situations being created subsequently, thus reducing the safety margins.

Similarly, constraints or targets can be placed on the end user which result in operational constraints that are difficult to manage. These might include:

- providing continuous descent approaches in complex TMA situations;
- minimising fuel burn by managing climb and descent profiles;
- stabilising an approach when flying a fuel-optimised aircraft configuration;
- reducing the impact of contrails in busy en-route airspace.

Anticipated MOC HF-05

To support the end user in complex situations, the AI-based system should be capable of monitoring the situation and determining what the current situation is, as part of its situation representation.

The system should also monitor the actions taken by the end user in complex situations and determine what the outcomes of the actions taken by the end user will be.

If the AI-based system's solution differs from the end user's solution, then the AI-based system should propose one or more alternative solutions. The outcome of alternative solutions will be defined in terms of the constraints of the operational system: time, workload, route miles flown, time at preferred altitude, etc.

The underpinning reasons for the solution being considered should be available to the end user.

The presentation of the 'reasoning' to the end user can be interpreted as argumentation and serve the purpose of operational explainability (per Objective EXP-05).

The end user on being presented with one or more alternative solutions should have the opportunity to accept or reject the proposal from the AI.

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 process and act upon arguments shared by the end user.

Anticipated MOC HF-06

Complex situations have been defined above in **Objective HF-05**. Abnormal situations are (within the context of safety assessments) defined as 'those conditions relevant to a change which cover occurrences such as: loss of external input to the system, interference, extreme weather conditions, unforeseen airspace closures, system failures, industrial actions/strikes.'

AI-based systems should be operated within the context of an allocation scheme and within that a more tightly defined allocation pattern. Within the allocation pattern, it is expected that the AI-based system is capable of assisting end users during complex and abnormal operations. This will include providing information relating to system status that is within a specific allocation pattern. Thus for any allocation pattern, the AI-based system should:

- identify the problem; within its allocation pattern the AI-based system should be able to determine which of its functions is compromised. The AI-based system should be able to determine whether:
 - it is not functioning;
 - it is functioning but providing inappropriate output;
 - it is capable of functioning, but the information it relies upon is unreliable;
- share the diagnosis; the AI-based system should provide the diagnosis of the problem to the end user, and ensure that the issue is recorded and communicated for subsequent analysis to technical users. In sharing the diagnosis, the AI-based system should be sensitive to the ongoing complex situation and use the most appropriate communication means available. The diagnosis should be communicated at a level that is anticipated to generate action from the end user, meaning that it has to be shared in a manner that meets the needs of the end user in anticipation of subsequent actions to be taken;
- identify the root cause; an AI-based system given responsibility to issue clearance to an aircraft via voice communications in an ATC environment should be able to determine that a transmitter is non-functional and that in fact messages to aircraft are not being sent. In the same way, an AI-based system should be capable of determining whether the services it relies upon within the overall operational system are functioning as expected. Where the services are not functioning as expected, the AI-based system should be able to identify where the problem arises;
- describe a resolution strategy; during abnormal conditions an appropriate resolution strategy may be to withdraw the AI-based service from the operational environment or suspend the particular allocation scheme the AI-based system is operating under. Similarly, the AI-based system may suggest a re-start of some service via the implementation of an emergency checklist. The specific case of complex and abnormal situations covers a broad range of scenarios and outcomes which may not be definable in advance. In all cases however, the important element is that the resolution strategy is communicated to the end user;
- provide any anticipated consequences; where an AI-based system is removed from service (for example) the information that it provides to the end user will not be available. Any resources that the end user relied upon will not be available for subsequent implementation and, such as automation failure, the end user will be in a 'degraded mode'. In de-activating an AI-based system there will be operational consequences, some of which the end user will be aware of, and others that the end user will not be aware of. The AI-based system should

make the end user aware of the operations and system consequences of implementing the recommendations made from implementing a resolution to an abnormal condition.

Within this MOC, the point of allocation pattern is referred to in order to hint that the applicant should be designing AI-based systems capable of diagnosing and providing information only within the remit given to the AI-based system, i.e. not an all-encompassing AI-based system (i.e. Level 3B AI).

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, alert and assist the end user.

Anticipated MOC HF-07

Time-critical phases may arise from the phase of flight when considering aircrew, or from the time of day when considering ATS operations. Critical phases of flight include take-off and landing, and handling emergency situations. For the ATM environment, time-critical situations are not only driven by traffic flow (typically morning and evening) but are also situations that require an immediate manoeuvre and emergency situations. In both cases prompt decision-making may result in less than optimum solutions.

The definition of poor decision-making will depend upon both the operational context and information available to the decision-maker, i.e. any decision can only be judged in terms of the information available to the end user to make the decision, and the operational goals that the decision must be balanced with. The end user in making time-critical decisions is likely to be focused on finding a solution, and finding a solution that matches the information available – context-based decisions. In today's operations, typically any single decision made does not need to be perfect as it can be corrected with a series of other decisions where time allows. In time-critical situations decisions may be hurried, and not necessarily take into account all information available, and therefore result in a poor outcome.

AI-based systems should maintain situation representation in order to be capable of alerting in time-critical decisions. This means that the AI-based system should:

- be able to detect time-critical situations, through, for instance, phase of flight, rate of system interaction or voice stress pattern;
- put into context the decision being made by the end user; what will the outcomes of the decision be, what other alternatives might suit the end goal more;
- identify, based on its situation representation, a more optimal suggestion that will achieve the goal more effectively;
- communicate with the end user to deliver the solution in a manner that fits the time-critical situation the end user is facing;
- provide sufficient time to enable the end user to execute the appropriate action.

To assist the end user, AI-based systems should be capable of one or more of the following enhanced decision-making options:

- The system has access to broad learning to which the individual end user may not have by e.g. exploiting large previously analysed datasets and machine learning.
- Perform decision-making based on the assessment and comparison of the risks associated with one option for a decision versus another.

Detect levels of voice stress and thereby recognise where a decision may be being made too quickly or without possibly assessing a number of solutions.

Objective HF-08: The applicant should design the AI-based system with the ability to propose alternative solutions and support its positions.

Anticipated MOC HF-08: Several dimensions for collaboration have been identified:

- Human and AI work together on an agreement to achieve the shared goal.

There is a need to design an AI-based system that can exchange in case of inconsistent tasks or strategies to propose alternative solutions to achieve the shared goal.

- Human and AI work individually on allocated task(s) and, when ready, share their respective solution for agreement.

In order for the AI-based system to be able to support its positions, it should be designed to select what information to provide for argumentation and which modality to use.

Objective HF-09: The applicant should design the AI-based system with the ability to modify and/or to accept the modification of task allocation pattern (instantaneous/short-term).

Anticipated MOC HF-09

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, adaptivity and need for collaboration are foreseen.

The following should be considered:

- Context-sensitive task allocation: AI-based systems 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. 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 conversation may imply a bidirectional communication. 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.

If spoken natural language is used, the applicant should design the AI-based system with the ability to acknowledge the end-user intention.

Objective HF-10: If spoken natural language is used, the applicant should design the AI-based system with the ability to process end-user requests, responses and reactions, and provide an indication of acknowledgement of the user’s intentions.

Anticipated MOC HF-10

Where spoken natural language is used to issue instructions or initiate an interaction with the AI-based system, it should be able to detect the verbalisations, recognise within it the requirement for interaction, extract the intent of the interaction and then act upon the request or instruction.

The applicant should ensure that the natural language interface (e.g. language model) is transparent so that erroneous interpretations can easily be detected by the end user.

The requirement to include reactions provides an additional means to correlate an interaction with an outcome and provide evidence of a successful interaction.

Objective HF-11: If spoken natural language is used, the applicant should design the AI-based system with the ability to notify the end user that he or she possibly misunderstood the information.

Understanding a spoken exchange involves recovering the literal meaning of the exchange. However, that literal meaning may have many missing elements that the speaker thinks the listener can fill in in the way the speaker intended given the speaker's contribution. Both of these processes can go wrong: misunderstanding arises when the listener is unable to recover the totality of the literal meaning of the exchange; misinterpretation results when the listener fills in the literal meaning in a way unintended by the speaker. For example, an intermitted radio frequency can lead to a lack of understanding.

Misunderstandings generally arise from multiple sources that include:

- a lack of clarity and conciseness in spoken messages;
- limited attention afforded by the end user to the message;
- lack of, or inability to use, clarifying questions; and
- limited opportunities to provide feedback to the speaker on whether the message has been understood.

Within a busy operational environment with already loaded communication channels, it is clear that misunderstandings may arise.

Anticipated MOC HF-11

The applicant should design the AI-based system with the ability to identify a user's misunderstanding of spoken communication. This implies that the system should be capable of monitoring the crew's actions in response to a communication, both verbally in terms of any reply, but also physically in terms of interaction with controls. Where the system detects an inappropriate response or input, the AI-based system should engage in strategies that reinforce the users' understanding such as:

- attract the users attention before issuing spoken communication;

- provide clear and concise messaging;
- question the end user's intent following an interaction that the system deems to have been misunderstood;
- afford the user the opportunity to ask clarifying questions following the interaction.

Objective HF-12: If spoken natural language is used, the applicant should design the AI-based system with the ability to identify through the end user responses or his or her action that there was a possible misinterpretation from the end user.

Misinterpretation refers to the 'act' of understanding something incorrectly or inaccurately often due to a mistake in the interpretation of information, signals or communications. Misinterpretation by the end user can result in inputs to the AI that are inaccurate or irrelevant for the scenario of interaction between the members of the HAT, or in a set of actions that are also inappropriate. The opportunity for misinterpretation can arise when the AI-based system is addressing an issue that the end user is not focused on, or when the end user is addressing an item that is not part of the situation representation of the AI-based system.

This creation of mismatch of scenarios can result in situations where the end user responds in a manner that is out of context for the AI-based system. The reaction from the end user will therefore, at best, be confusing for the AI-based system, or at worst, an inappropriate input to an appropriate situation.

Anticipated MOC HF-12

For any given scenario where misinterpretations could arise, the applicant should demonstrate that the AI-based system can support the end user by:

- detecting that the speech interaction did not provide the input / output expected;
- identifying multiple requests for the same or similar instructions occurring in a short period of time;
- detecting subsequent corrections to actions that arise as a result of the interaction;
- offering a means to the end user to clarify the instruction or interactions with the system.

Objective HF-13: In case of confirmed misunderstanding or misinterpretation of spoken natural language, the applicant should design the AI-based system with the ability to resolve the issue.

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.4.3.4).

Anticipated MOC HF-13

The applicant should design the system so that it is capable of detecting misinterpretation. When a misinterpretation is detected, the AI-based system should have access to multiple resolution

strategies and the most appropriate should be deployed at the appropriate time. These might include:

- repeating the message verbally with additional emphasis on the element that has been misunderstood;
- requesting explanation and clarification possibly by both the AI-based system and the end user;
- giving explanations by e.g. providing additional relevant information that will clarify the original communication:
 - additional context
 - rationale for the communication / data
 - relative information based on previous interactions that were not understood
- Asking questions for clarification
- The modality of the interaction may be modified by:
 - including an additional sense to key attention;
 - increasing the volume and emphasis of the spoken word;
 - including haptic or visual feedback to draw the user's attention to the parameters or information that should be attended to.

These additional sources of reference / context are not described in a hierarchy of relevance, neither need they be provided individually. Multiple additional sources of information and data can be made available for clarity and to reduce misunderstanding.

Objective HF-14: If spoken natural language is used, the applicant should design the AI-based system with the ability to not interfere with other communications or activities at the end user's side.

Managing operational tasks dictates that at different times end users will be occupied in other tasks or activities. Some of these activities will be complex and require the 'space' to process information cognitively, and some of which will load the aural channel and hinder perception of spoken information from a second source. In addition, conversational exchanges between humans are constructed to allow natural pauses; these serve to indicate to other participants that they can speak or enter the conversation, or at least that the speaker has finished speaking.

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 end user. Verbal communication also includes non-verbal cues that humans pick up during normal conversations. These non-verbal cues may not be available to AI-based systems.

Anticipated MOC HF-14

An AI-based system communicating using natural language will have to be sensitive to and follow conversational norms. Natural language norms include:

- Turn-taking: Participants speaking take turns with one participant speaking at a time. An AI-based system should not monopolise spoken communication or interrupt when end users are engaged in other conversations, or when it is the end user's turn.
- Respect the end user's request to have the floor and to be silent when it is not one's turn.
It might, however, be that the AI-based system interjects with a higher priority message than ongoing background cockpit discussions, in the case where a priority message or emergency situation is not being addressed.
- The AI-based system should follow general prosody, e.g. tone of voice, as speakers' voices tend to reduce in tone when a speaking turn comes to an end.
- Verbal cues: speakers may use phrases such as thank you, goodbye, roger-out, etc. when they have finished speaking to other parties. Question tags may indicate when a speaker has finished speaking, but expects a response from another party (not the AI-based system).

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.

4.3.2. 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 on the 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: If spoken procedural language is used, the applicant should design the syntax of the spoken procedural language so that it can be learned and applied easily by the end user.

4.3.3. Design criteria for gesture language

Gesture language is considered in this paper as a non-verbal, unidirectional communication tool where the end user would have their body movements and/or eye 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 communication.

Objective HF-17: If gesture language is used, the applicant should design the gesture language syntax so that it is intuitively associated with the command that it is supposed to trigger.

Anticipated MOC HF-17

Gesture language is complex and has many cultural inferences and differences. The design of gesture language is therefore sensitive, and requires significant planning to establish not only gestures, but a logical syntax.

Where it is implemented, designers should ensure that:

- gesture interaction is not required in situations where one or both hands are occupied with other tasks – throttle, trackball, mouse, etc.;
- the system should be capable of recognising left- and right-handed gestures and be clear if gestures of the left and right hand are specific or can be generalised to either hand;
- where a system design envisages multiple gesture inputs, these should be designed within the context of an overarching syntax, i.e. move beyond recognition of single gestures to incorporate a series of inputs to build a ‘compound’ message;

- inputs that have the same meaning should use the same gesture, e.g. the gesture for numbers should be consistent between instructions;
- actions that have opposite meanings should have related and opposite actions; as an illustration it might be that an instruction to climb is a finger pointing up, so an instruction to descend should be a finger pointing down;
- a means to correct or annul gesture-based inputs, and also to revert to other means of input, should be provided.

Objective HF-18: If gesture language is used, the applicant should design the AI-based system with the ability to disregard non-intentional gestures.

Anticipated MOC HF-18: Non-intentional gestures include spontaneous gestures that are made to complement spoken language, or made in the context of non-related tasks.

Objective HF-19: If gesture language is used, the applicant should design the AI-based system with the ability to recognise the end-user intention.

Objective HF-20: If gesture language is used, the applicant should design the AI-based system with the ability to acknowledge the end-user intention with appropriate feedback.

Anticipated MOC HF-20

Feedback from the AI-based system to the end user should be provided through the most appropriate and available channel. In a heavily loaded auditory environment, the feedback may be primarily visual and colour based (for example, pulsing a green light within a cockpit, highlighting a green outline on a track-data-block). If the end user is engaged in a task that is loading visual channels, then auditory feedback might be provided.

To impart feedback, where an AI-based system is limited to auditory tones and/or colours, the use of both colours and auditory tones will also have to be aligned with the design philosophy of the existing system within which they are embedded, and existing regulation (e.g. CS 25.1322 – the use of red for alerts and warning). Thus, the use of red should be avoided, and the auditory system feedback should be aligned with other auditory warnings and priorities established for their production in layered audio environment.

The possibility also exists to provide verbal feedback to non-verbal inputs; these include both written and spoken feedback.

In summary, the applicant should ensure that:

- where a gesture language input is made to the AI-based system, the system should not be limited to only providing non-verbal system feedback;

- non-verbal feedback should be aligned with existing system design constraints, colour use and auditory warnings;
- feedback from the system should take account of the tasks being performed and provide appropriate and adequate feedback taking account end user channel loading.

4.3.4. 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.

Adaptive interaction modality is the AI capacity to adapt the modality of interaction to external/internal factors with the objective of optimising the HAIL. The following attributes have been identified: context-sensitive criteria (e.g. if pilot is speaking with ATC, the AI-based system will communicate using other interfaces than natural language), task-sensitive criteria, pilot-state-sensitive criteria.

The AI-based system should propose modality of interaction according to the perception of the pilot's preferences and expectations. As an example, during ATC-pilot communication, the AI-based system may avoid using natural language to interact with the pilot and instead display the collected information.

Objective HF-21: If spoken natural language is used, the applicant should design the AI-based system so that this modality can be deactivated for the benefit of other modalities.

The natural language interface relies on several complex processes: speech detection, recognition, comprehension and production. In normal operations the processes evidently take place in a noisy environment (cockpit or ops-room) and potentially when the end users' speech pattern can be affected due to stress or workload. Noisy environments and affected speech can lead to a degradation in performance of the spoken natural language interface, for which spoken procedural language is not an alternative. While an end user might be tolerant of lower levels of system performance, a point will arise when the interface no longer supports interaction in a manner that is sufficiently accurate or timely for the safety of the operation.

The decision point to shift to alternative means of interaction maybe be driven by either the end user or the system. Where end users become aware of degraded performance in natural language interactions, the design of the system should facilitate a shift to other means of interaction upon request. Degradation in performance will most likely be experienced as a reduced capacity of the

system to be 'right first time' i.e. multiple attempts will be required to complete the same interaction with the system.

Anticipated MOC HF-21

Upon making the decision to stop using natural spoken language, the system should:

- provide an appropriate means to end spoken natural language;
- provide an alternative means to interact with the end user;
- confirm to the end user that the 'reversionary' mode of interaction is operational;
- provide a means to switch back to (natural or procedural) spoken language when circumstantial factors allow.

The (natural or procedural) spoken language system should also maintain some measure of integrity and the extent to which it is successfully interpreting the verbal interactions with the end user. When the system determines that its own precision is not sufficient to continue operations, it should announce to the end user the intention to suspend this modality of interaction. On the decision to suspend (natural or procedural) spoken language, the system should make clear which other modes will be preferred.

Alternative means of interacting with the system should include traditional input/output devices i.e. cursor control devices, screens, and HMIs, printed output, other aural and visual displays or haptic devices. Thus the system should be designed not only to remove a degraded natural spoken interface, but also provide alternatives that reduce the impact of the loss of natural spoken language.

Within the context of a (natural or procedural) spoken language interface with the AI-based system, muting the AI-based system during a degraded situation is neither sufficient nor acceptable. End users should be able to select, or be offered, an alternative method for interacting with the AI-based system when spoken language is not available. The AI-based system's design should move beyond a 'mute' function to ensure that information is still communicated to the human and that attention is drawn to its presentation.

Objective HF-22:

If spoken (natural or procedural) language is used, the applicant should design the AI-based system with the ability to assess the performance of the dialogue.

Objective HF-23: If spoken (natural or procedural) language is used, the applicant should design the AI-based system with the ability to transition between spoken natural language and spoken procedural language, depending on the performance of the dialogue, the context of the situation and the characteristics of the task.

Anticipated MOC HF-23: The applicant should design the AI-based system with the ability to transition from spoken natural language to spoken procedural language in case the performance is degraded.

Objective HF-24: 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.

Human-to-human interaction in the cockpit and ATM environment is primarily performed using voice. Briefings are performed verbally between crew members before taxi, take-off, approach and landing, and are put in place to support safe operations and ensure that key elements of information are not only transmitted, but also understood. Even though natural spoken language is used in operations, if multiple simultaneous messages are passed verbally, there is an increased likelihood that one or more of them will be masked or interfered with such that the content will be misunderstood or missed altogether. It is necessary therefore to consider more than one (multi) modality of exchange when dealing with AI-human interaction.

Anticipated MOC HF-24

AI support should be adaptable and use different modes of interaction at different stages that reflects the available channels and resources for each modality of communication:

- During periods when end users are physically occupied with managing system interaction (flight controls or physical maintenance activities), voice is used.
- During periods when auditory channels are loaded, visual or haptic feedback is provided.
- When a change in operations is detected (or signalled by the end user), the AI-based system should be able to shift its interaction.

The applicant should ensure that multiple modes of interaction are available between the end user and the AI system and that both the AI-based system and end user are able to select the most appropriate mode for the scenario or context at the time.

Objective HF-25: The applicant should design the AI-based system with the ability to automatically adapt the modality of interaction to the end-user states, the situation, the context and/or the perceived end user's preferences.

Anticipated MOC HF-24

End-user states, both temporary (e.g. workload, fatigue) and permanent (e.g. user preferred settings) will affect the preferred channel for interacting with the AI-based system. In addition, as discussed in MOC HF-23, the situation and context will also vary along the duration of an operation. End users will experience periods of high workload, periods of intense verbal communication, management of system failures and handling of emergencies (diversions, etc). The end user, when time allows, will make adaptations to their preferred modes of interaction; however, there will also be occasions that the AI-based system should automatically adapt the modality of interaction.

By inference, the following design possibilities have been identified:

- The modality of interaction can be predefined by the applicant in adaptation to the characteristics of the task or the flight event.
- The modality of interaction should adapt to the pilot's state:
 - permanent state (crew personal setting);
 - instantaneous state (workload, stress, cognitive resources).

The shift in modality of interaction should be initially automatically proposed, with the option for the end user to override the proposal from the system.

4.4. Error management

4.4.1. Contribution of AI-based systems to a new typology of human errors

Taking the illustration in the IAW domain, 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 should 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.

Typically, the introduction of AI could result in errors in several ways:

- Over-reliance on the system: AI-based systems should become increasingly reliable and increasingly prevalent, thus the end user will rely on them without fully understanding their limitations or potential failure modes. This will increase the likelihood of error when the AI system does fail.
- Increasingly complex decision-making: the data and information that an AI-based system can access in order to make a decision or perform an action could be significantly beyond the comprehension of the end user. Therefore if an end user does not understand significantly more about the system than today, may he or she make errors in judgement or execution.
- Data bias and misinterpretation; where bias or restricted data sets are not addressed in learning assurance (Section C.3.1), the AI-based system can perpetuate or amplify existing bias leading to bad decisions.
- Unexpected — as yet unknown — failure modes; as AI-based systems have increasingly broad remit of allocation schemes and patterns, the likelihood that failures will occur in unexpected ways increases. End users may not be trained to recognise and respond to cascading AI failures and similarly may not be able to troubleshoot or mitigate the consequences.
- Human-machine interface issues; AI-based systems will require new modes of interaction, some of which may not yet have been deployed, and others that may have additional requirements

arising from interacting with AI. The end user may misinterpret or misunderstand the AI and this will increase the risk of human error.

- Transparency; as the sophistication of AI-based systems increases, the extent to which the AI operates as a black box will also increase. The lack of transparency can hinder the end users' ability to trust and verify the outputs of the system. This will increase the likelihood of errors in decision-making or action implementation.

When looking at classical descriptions of human error, the following comments can be made on the likely impact of AI:

- Errors of commission: these may occur when end users trust an AI-generated outcome without critically evaluating it. AI-based systems providing incorrect but relevant output can lead end users to making errors. This type of error will become increasingly likely as the level of AI moves from 1B to 2A and from 2A to 2B.
- Errors of omission: in automating tasks that end users currently perform, the AI-based system could lead to a user-based error of omission in task-sharing. The end user is under the impression that the AI-based system will perform a task whereas the allocation pattern requires the end user to perform the task.
- Slips: within the context of operating an aircraft, the notion of slips (inappropriate actions) is excluded from the considerations in CS 25.1302. Largely, it is assumed that the pilot should be able to operate the aircraft at a level of skill required of the type. However with the introduction of AI, particularly when migrating between level 1B and 2B, there will be an increased requirement to acknowledge system outputs and approve AI-based system activities. This will increase the significance of slip-based errors as the sophistication of AI-based systems increases.
- Lapses: referring to errors caused by forgetfulness or memory failure, the introduction of AI-based systems is not expected to reduce the impact of lapse errors. However, where an end user relies heavily on AI they may eventually forget to access relevant information, believing that the AI 'has it covered'.
- Mistakes: may increase if end users misinterpret or misapply AI-generated solutions or rely too heavily on AI recommendations without considering other relevant information or situational variables. Similarly, where the AI learning database is not updated, then the AI-based system may lead to the end user making mistakes.

The nature and manifestation of any error, on the side of either the AI-based system or the end user will be specific to the implementation of the AI-based system, the operational environment, and the allocation scheme and pattern in use. The following objectives address at a generic level the intentions of design to minimise the likelihood of error occurrence of all the types of errors described above.

4.4.1.1. Design-related human errors

Objective HF-26: The applicant should design the AI-based system to minimise the likelihood of design-related end-user errors.

Note: The minimisation of the likelihood of errors made by the AI-based system is addressed through the AI learning assurance Section C.3.

Anticipated MOC HF-26

To prevent and mitigate design-related errors in AI-based systems, the design and development process should be rigorous and systematic. Specifically, the design process should include collaboration between (as a minimum) data scientist, domain experts, ethicists and user experience designers.

This MOC focuses on end-user errors that are induced by the design of the system. End-user errors induced by system design arise, generally, from poor system design and lack of consideration of the end user, their tasks and the environment within which they are performed. End users will often blame the design of a system for leading them to creating an error; this may be input, reading, a poor scan pattern, or the requirement to read, remember and re-input data from one field to another. Design-related errors can be minimised by incorporating a systematic approach to consideration of the user in the development phase of the AI-based system.

To avoid end-user errors arising from system design, the system designer (or software engineer) should:

- follow a user-centred design process and understand clearly what the user needs are. In the absence of user needs (many systems may not have user needs), the designer should establish how users will interact with a proposed system and develop accepted user needs, preferences, and take account of user capabilities and limitations;
- provide simple and intuitive interfaces that minimise complexity and cognitive load through familiar design patterns, consistent layouts and clear labelling;
- provide feedback mechanisms to inform the user about their actions, and the AI-based system should continue standard design practices of incorporating visual cues, notifications and error messages to provide immediate feedback and help users correct errors;
- incorporate, during the design process, adequate user testing and human error analysis in order to predict the types of common errors that end users will make and eliminate them from the design. Additional validation checks of user input should be employed to ensure that the system cannot be 'derailed' by erroneous end-user data input;
- ensure consistent terminology, concepts and means of interaction to avoid confusion. Language and terminology should be familiar to the end user / target audience; and
- employ rigorous user testing with as broad a range of predicted end users as possible.

4.4.1.2. Operation-related errors

In the aviation environment the use of two people on 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, situation 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

To best anticipate and manage the errors arising from HAIRM, the following examples should be taken into account:

- Failure to respect the predefined task allocation pattern: if the end user fails to respect the task allocation pattern, misunderstandings will arise between the end user and the AI-based system as to what actions each of them is performing. The applicant should ensure that it is clear to the end user which allocation pattern is implemented and which, exhaustively, tasks are allocated to the AI-based system and which ones to the end user.
- Teamwork breakdown: the team is not working properly, e.g. due to a communication issue, lack of trust, poor definition of tasks and of task allocation pattern. Teamwork breakdown can be avoided by following the guidance described in the MOC to Objective HF-25 i.e. by fully understanding the team roles and impact on the end user before implementing the AI-based system.
- Incomplete or incorrect cross-checking process: the introduction of an AI-based system will require modification to operational procedures. Replacing one human user with an AI-based system will require operational procedures to be reviewed to ensure at least the same level of effectiveness as with two human team members.
- Human-AI communication issues: for the errors resulting from the communication established between the AI-based system and the end user, refer to Objectives HF-11 to HF-13.
- Mismatch between situation awareness/representation: the applicant should demonstrate through user trials and testing that the situation representation held by the AI-based system (irrespective of its AI Level [2A or 2B]) is representative of the situation awareness of the end user.
- Decision-making authority issues: within the scope of AI Level 2B, the AI-based system will have the ability to present arguments to the end user as to the reasons for a preferred outcome or result.

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 as intended despite human errors. The introduction of AI-based systems will potentially offer systems that can anticipate and gracefully handle end-user errors. The applicant should ensure that the AI-based system:

- provides a clear and intuitive interface: increasing error tolerance begins with ensuring that the interface is less likely to lead to the end user making an error, and when an error has been made to allow the end user to understand simply how to correct it;
- incorporates feedback mechanisms: the end user should be informed by the AI-based system when an inappropriate input is provided. Feedback should be provided in a manner that is timely and allows the end user to correct their input appropriately;
- incorporates validation: when an input is made to the system, it should check that the input is in a format expected, i.e. numeric rather than alphabetic;
- prompts 're-do': the AI-based system should request from the end-user a re-entry or the data required, and make clear what data is needed, and what error was made by the user;
- analyses error logs: to ensure continuous improvement, the AI-based system should log the errors made by the end user for subsequent analysis. It is vitally important that the logs are saved after any form of system shutdown and are only removed at operator request. The logs should be analysed to determine common error patterns and the AI-based system improved to account for the errors.

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.

Anticipated MOC HF-29

It is inevitable that end-user errors will occur when interacting with all systems. Therefore, the AI-based system, in addition to being largely insensitive to erroneous user input, should also facilitate the opportunity to detect and correct the error. Several methods exist for detecting human error in system interaction, and the applicant should ensure that methods are integrated into the AI-based system that include, but are not limited to:

- input validation: the AI-based system should provide the opportunity to validate the input provided to the system. This can be a simple visualisation of a data input, or an audio

readback, before the system acts upon the data that the user has input allowing the end user to correct or cancel the input;

- error messages: these provide the opportunity for the end user to detect an erroneous data entry when flagged up by the system. Error messages will inform the users of the nature of the error, incorrect data type, etc. and require that a new entry is made to the system;
- usage patterns: where an AI-based system expects a specific type of input and this is not provided, the system can flag this to the end user. For example, where an aircraft is entering a TMA, a radio frequency exchange might be expected. When the frequency change is not made, the AI-based system may flag this to the end user. Where an erroneous input is provided, before changing frequency, the user should have the opportunity to change the input to the correct one.

When an erroneous input has been detected, the end user should be offered the opportunity to correct the error. Error correction should:

- provide system suggestions, if appropriate, for a corrected end-user input;
- allow for the same entry method as previously used;
- allow for a different mode of data entry if the end user so prefers;
- have confirmation of the revised data entry values;
- indicate to the end user that the error is corrected; and
- allow a further cycle of correction if the end user makes a further 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.

4.5. Failure management

Objective HF-31: The applicant should design the system to be able to diagnose the failure and present the pertinent information to the end user.

Anticipated MOC-HF-31

Users should be informed about the nature and scope of the failure. This includes details about what went wrong, why it happened, and how it affected the system's performance or output.

End users need to understand how the failure affects the results or services provided by the AI-based system. This could include inaccurate predictions, delayed responses, or complete unavailability of the system.

The attraction of the end user to the occurrence of a failure should happen via visual alerts (colours) or auditory alerts (verbal and non-verbal). The use of both colours and auditory tones will also have

to be aligned with the design philosophy of the existing system within which they are embedded, and with existing regulation (e.g. CS 25.1322 – the use of red for alerts and warning).

Objective HF-32: The applicant should design the system to be able to propose a solution to the failure to the end user.

Anticipated MOC-HF-32

If applicable, end users should be provided with contingency plans or alternative workarounds to address the effects of the failure. This could involve using backup systems, manual processes, or alternative sources of information.

Clear instructions on how to access and utilise these contingency measures help users navigate the situation effectively.

End users need to know the expected timeline for resolving the issue and restoring the system to normal operation. This includes information about ongoing efforts to address the problem and any updates or milestones in the resolution process.

End users should be provided with channels for seeking support or assistance in case they encounter difficulties or have questions related to the system failure.

Objective HF-33: The applicant should design the system to be able to support the end user in the implementation of the solution.

Anticipated MOC-HF-33

The AI-based system should be able to accept a request from the end user to perform an action. The AI-based system proposes a solution and after confirmation from the end user the AI based system acts on the proposal and resolves the failure.

Objective HF-34: The applicant should design the system to provide the end user with the information that logs of system failures are kept for subsequent analysis.

5. AI safety risk mitigation

5.1. AI safety risk mitigation concept

AI safety risk mitigation 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 safety risk mitigation. 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).

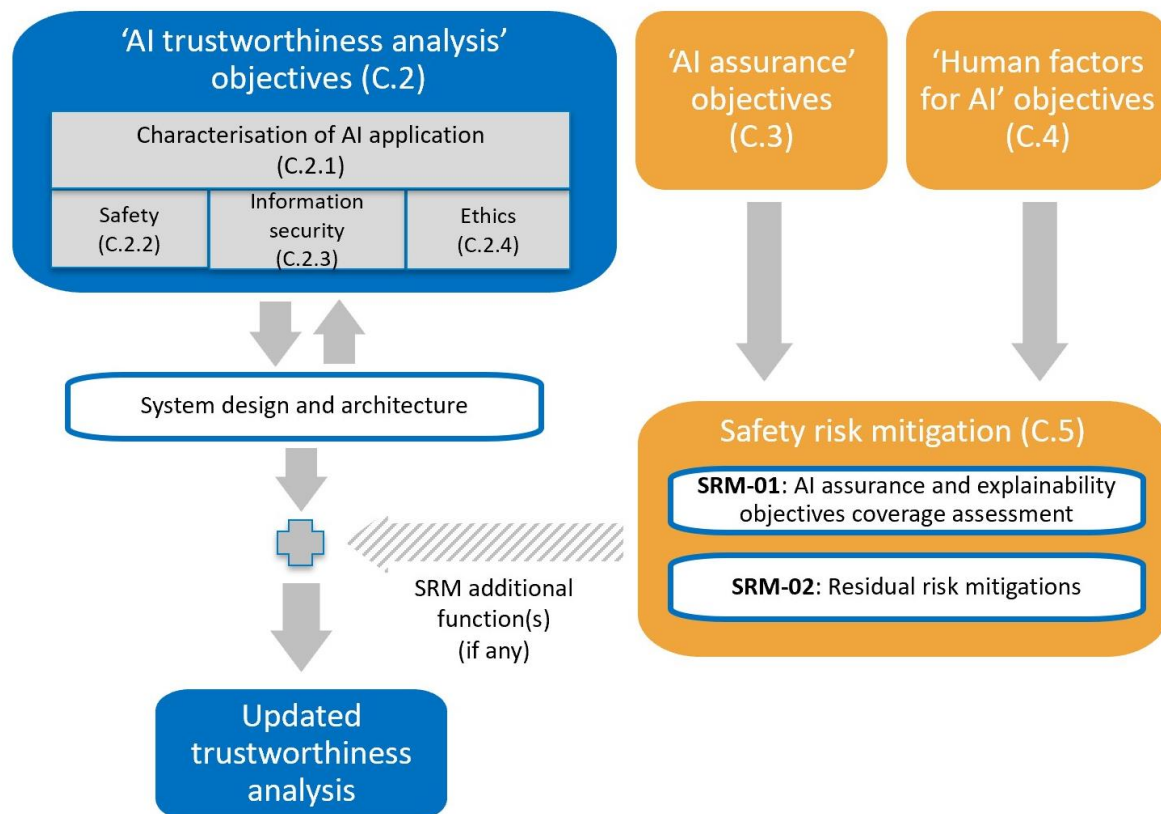


Figure 22 — Safety risk mitigation block interfaces with other building blocks

Note that safety risk mitigation is solely meant to address a partial coverage of the applicable explainability and learning assurance objectives. Safety risk mitigation 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 safety risk mitigation 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, would be necessary to mitigate the residual risks to an acceptable level.

Anticipated MOC SRM-01: In establishing whether AI safety risk mitigation 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 safety risk mitigation 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 safety risk mitigation will be needed.

In particular, the qualitative nature of some building block mitigations/analysis should be reviewed to establish the need for an safety risk mitigation.

The safety risk mitigation strategy should be commensurate with the residual risk/unknown.

Objective SRM-02: The applicant should establish safety risk mitigation means as identified in **Objective SRM-01**.

Anticipated MOC SRM-02: 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 safety risk mitigation functions should be evaluated as part of the safety assessment²², 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 safety risk mitigation architectural mitigations from the AI/ML constituent

²² 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 continuously 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.

Provision ORG-03: Implement a data-driven ‘AI continuous safety assessment’ process based on operational data and in-service events.

Anticipated AMC ORG-03:

The applicant should use the collected data (per **Objective EXP-09**) to perform a continuous safety assessment. This includes:

- 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.

Moreover, with the AI continuous safety assessment system, the organisation should:

- ensure gathering of data 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 continuous safety assessment should enable a refinement of the identification of risks based on the results of previous interactions with the AI-based systems and incorporating the human evaluation inputs.

—

Provision ORG-04: 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**.

Anticipated AMC ORG-04:

In particular, the applicant should put in place:

- an AI ethics review board;
- a process to discuss and continuously monitor and assess the AI-based system’s adherence to the ethics-based assessment guidance.

Provision ORG-05: 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 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 ensure that the safety-related AI-based systems are auditable by internal and external parties, including especially the approving authorities.

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 HAI 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 should be delivered to all users (all levels of personnel, including top management), to ensure the correct approach to the introduction of AI-based technology in the organisation.

Provision ORG-07: The organisation should adapt the training processes to accommodate the specificities of AI, including interaction with all relevant stakeholders (users and end users).

Anticipated AMC ORG-07:



In particular, the applicant should put in place for all identified users and/or end users:

- the competencies needed to deal with the AI-based systems;
- the adaptations to the training syllabus to take into account the specificities of AI.

At the individual level, the elements above should be addressed in the initial training of each domain-specific licence or certificate (refer to **Provision ORG-07**). Furthermore, device- or environment-specific elements should be considered for the final use cases.

Provision ORG-08: 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.

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:

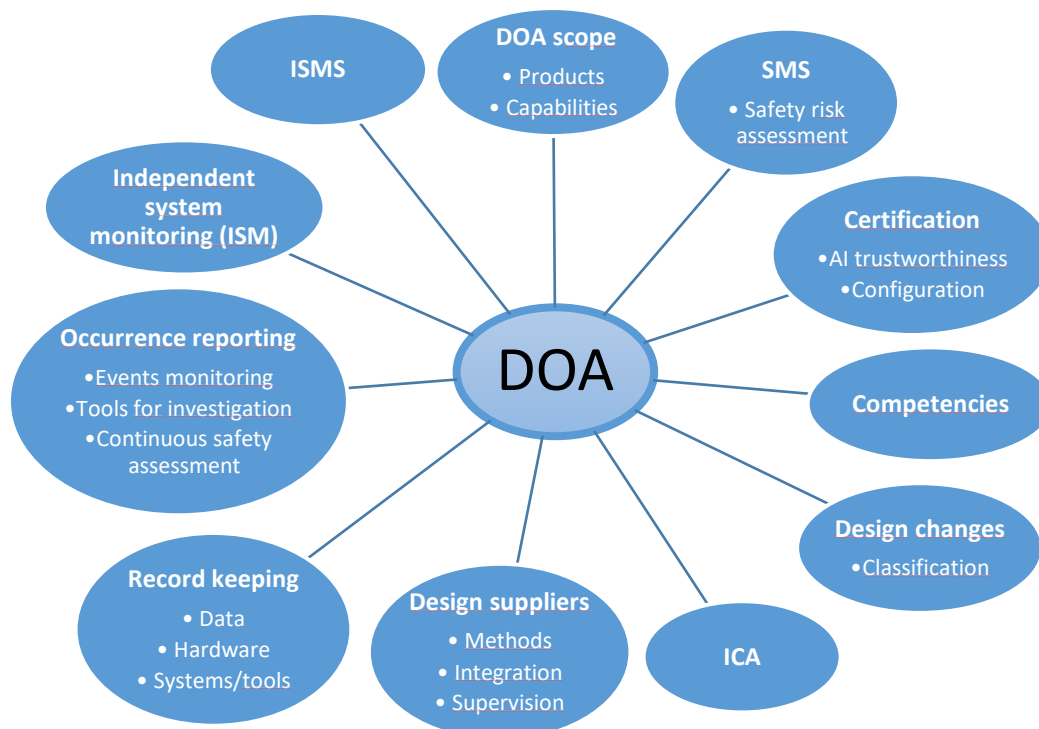


Figure 23 — DOA processes potentially affected by the introduction of AI/ML

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:

- Considering the limited experience gained from operations on the guidance proposed in this document and the unavailability of some MOC for a number of challenging objectives applicable to the highest levels of criticality, EASA will initially accept only applications where AI/ML constituents do not 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.
- The AI Level is rather independent from the criticality of the application. For instance, a Level 1A application could be assessed as safety critical and the modulation should be applied based on the assurance level allocated to its constituting items.
- Future work on Level 3 is likely to increase the number of objectives.

The numbering of objectives is using a set of specific set of acronyms that are defined here: 'CO' stands for ConOps, 'CL' for classification, 'SA' for safety assessment, 'IS' for information security, 'ET' for ethics-based assessment, 'DA' for development assurance, 'DM' for data management, 'LM' for learning management, 'IMP' for implementation, 'CM' for configuration management, 'QA' for quality assurance, 'RU' for reuse, 'SU' for surrogate modelling, 'EXP' for explainability, 'HF' for human factors and 'SRM' for safety risk mitigation. Finally, the suffix '-SL' refers to objectives applying only to supervised learning approaches, whereas '-UL' applies to unsupervised learning approaches only.

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.
	<i>The objective should be satisfied for AI level 1B, 2A and 2B.</i>
	The objective should be satisfied for AI level 2A and 2B.
	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 end-user roles, responsibilities (including indication of the level of teaming with the AI-based system, i.e. none, cooperation, collaboration) and expected expertise (including assumptions made on the level of training, qualification and skills).	○	○	○	○	○
	CO-02: For each end user, the applicant should identify which goals and associated 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, as well as a functional decomposition and allocation down to the lower level.	○	○	○	○	○
	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.	●	●	○	○	○
	SA-02: The applicant should identify which data needs to be recorded for the purpose of supporting the continuous safety assessment.	●	●	○	○	○
	SA-03: In preparation of the continuous safety assessment, the applicant should define target values, thresholds and evaluation periods to guarantee that design assumptions hold.	●	●	○	○	○
	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.	○	○	○	○	○
	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.	●	●	○	○	○

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	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 over-reliance, attachment, stimulating addictive behaviour, or manipulating the end user's behaviour.	○	○	○	○	○
	ET-03: 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-04: The applicant should ensure that the creation or reinforcement unfair bias in the AI-based system, regarding both the data sets and the trained models, is avoided, as far as such unfair bias could have a negative impact on performance and safety.	○	○	○	○	○
	ET-05: The applicant should ensure that end users are made aware of the fact that they interact with an AI-based system, and, if applicable, whether some personal data is recorded by the system.	○	○	○	○	○
	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), and define measures to reduce or mitigate these impacts.	○	○	○	○	○
	ET-07: 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.	○	○	○	○	○
	ET-08: 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.	○	○	○	○	○
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.3.1.14, as well as the interface and compatibility with development assurance processes.	○	○	○	○	○
	DA-02: Based on (sub)system requirements that have been allocated to the AI/ML constituent, the applicant should capture the following minimum requirements for the AI/ML constituent: <ul style="list-style-type: none"> — 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 AI/ML constituent ODD monitoring and performance monitoring, detection of OoD input data and data-recording requirements; — other non-functional requirements allocated to the AI/ML constituent; and interface requirements. 	○	○	○	○	○
A I	DA-03: The applicant should define the set of parameters pertaining to the AI/ML constituent ODD, and trace them	○	○	○	○	○

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
	to the corresponding parameters pertaining to the OD when applicable.					
	DA-04: The applicant should capture the DQRs for all data required for training, testing and verification of the AI/ML constituent, including but not limited to: [...]	○	○	○	○	○
	DA-05: The applicant should capture the requirements on data to be pre-processed and engineered for the inference model in development and for the operations.	○	○	○	○	○
	DA-06: The applicant should describe a preliminary AI/ML constituent architecture, to serve as reference for related safety (support) assessment and learning assurance objectives.	○	○	○	○	
	DA-07: The applicant should validate each of the requirements captured under Objectives DA-02, DA-03, DA-04, DA-05 and the architecture captured under Objective DA-06.	●	●	○	○	○
	DA-08: The applicant should document evidence that all derived requirements have been provided to the (sub)system processes, including the safety (support) assessment.	○	○	○	○	○
	DA-09: 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-10: Each of the captured (sub)system requirements allocated to the AI/ML constituent should be verified.	●	●	○	○	○
	DM-01: The applicant should identify data sources and collect data in accordance with the defined ODD, while ensuring satisfaction of the defined DQRs, to drive the selection of the training, validation and test data sets.	○	○	○	○	○
	DM-02-SL: Once data sources are collected, the applicant should ensure that the annotated or labelled data in the data set satisfies the DQRs captured under Obj. DA-04.	●	●	○	○	○
	DM-02-UL: Once data sources are collected and the test data set labelled, the applicant should ensure that the annotated or labelled data in this test data set satisfies the DQRs captured under Objective DA-04.	●	●	○	○	○
	DM-03: The applicant should define the data preparation operations to properly address the captured requirements (including DQRs).	○	○	○	○	○
	DM-04: The applicant should define and document pre-processing operations on the collected data in preparation of the model training.	○	○	○		
	DM-05: 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-06: The applicant should distribute the data into three separate data sets which meet the specified DQRs in terms of independence (as per Objective DA-04): — 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.	○	○	○	○	○

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-07: The applicant should ensure validation and verification of the data, as appropriate, throughout the data management process so that the data management requirements (including the DQRs) are addressed.	●	●	○	○	○
	DM-08: The applicant should perform a data verification step to confirm the appropriateness of the defined ODD and of the data sets used for the training, validation and verification of the ML model .	●	●	○		
	LM-01: The applicant should describe the ML model architecture.	○	○	○	○	○
	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; — explainability capabilities of the selected model; — activation functions; — 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 bounds.	○	○	○		
	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-SL: The applicant should account for the bias-variance trade-off in the model family selection and should provide evidence of the reproducibility of the model 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 requirements-based verification of the trained model behaviour.	●	●	○	○	○
	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, covering the whole AI/ML constituent ODD.	●	●	○	○	○

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-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.	●	●	○		
	LM-15: the applicant should capture the description of the resulting ML model.	○	○	○	○	○
	LM-16: The applicant should confirm that the trained model verification activities are complete.	●	●	○		
	IMP-01: The applicant should capture the requirements pertaining to the ML model implementation process.	○	○	○	○	○
	IMP-02: The applicant should validate the model description captured under Objective LM-15 as well as each of the requirements captured under Objective IMP-01.	○	○	○	○	○
	IMP-03: The applicant should document evidence that all derived requirements generated through the model implementation process have been provided to the (sub)system processes, including the safety (support) assessment.	○	○	○	○	○
	IMP-04: 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.	○	○	○		
	IMP-05: The applicant should plan and execute appropriate development assurance processes to develop the inference model into software and/or hardware items.	○	○	○		
	IMP-06: 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-07: The differences between the software and hardware of the platform used for training and those used for verification should be identified and assessed for their possible impact on the inference model behaviour and performance.	○	○	○		
	IMP-08: 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-09: The applicant should perform and document the verification of the stability of the inference model.	●	●	○	○	○
	IMP-10: The applicant should perform and document the verification of the robustness of the inference model in adverse conditions.	●	●	○	○	○
	IMP-11: The applicant should perform requirements-based verification of the inference model behaviour when integrated into the AI/ML constituent.	●	●	○	○	○
	IMP-12: The applicant should confirm that the AI/ML constituent verification activities are complete.	●	●	○		

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	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.	●	●	●	●	●
	RU-01: The applicant should perform an impact assessment of the reuse of a trained ML model before incorporating the model into an AI/ML constituent. The impact assessment should consider: [...]	○	○	○	○	○
	RU-02: The applicant should perform a functional analysis of the COTS ML model to confirm its adequacy to the requirements and architecture of the AI/ML constituent.	○	○	○	○	○
	RU-03: The applicant should perform an analysis of the unused functions of the COTS ML model, and prepare the deactivation of these unused functions.	○	○	○	○	○
	SU-01: The applicant should capture the accuracy and fidelity of the reference model, in order to support the verification of the accuracy of the surrogate model.	○	○	○	○	○
	SU-02: The applicant should identify, document and mitigate the additional sources of uncertainties linked with the use of a surrogate model.	○	○	○	○	○
	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 design the AI-based system with the ability to deliver an indication of the level of confidence in the AI/ML constituent output, based on actual measurements or on quantification of the level of uncertainty.	○	○	○		
	EXP-05: The applicant should design the AI-based system with the ability to monitor that its inputs are within the specified operational boundaries (both in terms of input parameter range and distribution) in which the AI/ML constituent performance is guaranteed.	○	○	○		
	EXP-06: The applicant should design the AI-based system with the ability to monitor that its outputs are within the specified operational performance boundaries.	○	○	○		
	EXP-07: The applicant should design the AI-based system with the ability to monitor that the AI/ML constituent outputs (per Objective EXP-04) are within the specified operational level of confidence.	○	○	○		

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	EXP-08: The applicant should ensure that the output of the specified monitoring per the previous three objectives are in the list of data to be recorded per MOC EXP-09-2.	○	○	○		
	EXP-09: 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, as well as the means to retrieve this data.	○	○	○	○	○
Human factors for AI	EXP-10: 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-11: The applicant should ensure that the AI-based system presents explanations to the end user in a clear and unambiguous form.	○	○	○	○	○
	EXP-12: 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-13: 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-14: Where a customisation capability is available, the end user should be able to customise the level of abstraction as part of the operational explainability.	○	○	○	○	○
	EXP-15: 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-16: 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-17: For each output relevant to the task(s), the applicant should ensure the validity of the specified explanation.	○	○	○	○	○
	EXP-18: The training and instructions available for the end user should include procedures for handling possible outputs of the ODD monitoring and output confidence monitoring.	○	○	○	○	○
	EXP-19: Information concerning unsafe AI-based system operating conditions should be provided to the end user to enable them to take appropriate corrective action in a timely manner.	○	○	○		
	HF-01: The applicant should design the AI-based system with the ability to build its own individual situation representation.	○	○	○	○	○
	HF-02: The applicant should design the AI-based system with the ability to reinforce the end-user individual situation awareness.	○	○	○	○	○
	HF-03: The applicant should design the AI-based system with the ability to enable and support a shared situation awareness.	○	○	○	○	○
	HF-04: If a decision is taken by the AI-based system that requires validation based on procedures, the applicant should design the AI-based system with the ability to request a cross-check validation from 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	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 an improved solution. Corollary objective: The applicant should design the AI-based system with the ability to process and act upon a proposal rejection from the end user.	○	○	○	○	○
	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 process and act upon arguments shared by the end user.	○	○	○	○	○
	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, alert and assist the end user.	○	○	○	○	○
	HF-08: The applicant should design the AI-based system with the ability to propose alternative solutions and support its positions.	○	○	○	○	○
	HF-09: The applicant should design the AI-based system with the ability to modify and/or accept the modification of task allocation pattern (instantaneous/short-term).	○	○	○	○	○
	HF-10: If spoken natural language is used, the applicant should design the AI-based system with the ability to process end-user requests, responses and reactions, and provide indication of acknowledged user's intentions.	○	○	○	○	○
	HF-11: If spoken natural language is used, the applicant should design the AI-based system with the ability to notify the end user that he or she possibly misunderstood the information.	○	○	○	○	○
	HF-12: If spoken natural language is used, the applicant should design the AI-based system with the ability to identify through the end-user responses or his or her action that there was a possible misinterpretation from the end user.	○	○	○	○	○
	HF-13: In case of confirmed misunderstanding or misinterpretation of spoken natural language, the applicant should design the AI-based system with the ability to resolve the issue.	○	○	○	○	○
	HF-14: If spoken natural language is used, the applicant should design the AI-based system with the ability to not interfere with other communications or activities at the end user's side.	○	○	○	○	○
	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.	○	○	○	○	○
	HF-16: If spoken procedural language is used, the applicant should design the syntax of the spoken procedural language so that it can be learned and applied easily by the end user.	○	○	○	○	○
	HF-17: If gesture language is used, the applicant should design the gesture language syntax so that it is intuitively associated with the command that it is supposed to trigger.	○	○	○	○	○

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-18: If gesture language is used, the applicant should design the AI-based system with the ability to disregard non-intentional gestures.	○	○	○	○	○
	HF-19: If gesture language is used, the applicant should design the AI-based system with the ability to recognise the end-user intention.	○	○	○	○	○
	HF-20: If gesture language is used, the applicant should design the AI-based system with the ability to acknowledge the end-user intention with appropriate feedback.	○	○	○	○	○
	HF-21: In case spoken natural language is used, the applicant should design the AI-based system so that this modality can be deactivated for the benefit of other modalities.	○	○	○	○	○
	HF-22: If spoken (natural or procedural) language is used, the applicant should design the AI-based system with the ability to assess the performance of the dialogue.	○	○	○	○	○
	HF-23: If spoken (natural or procedural) language is used, the applicant should design the AI-based system with the ability to transition between spoken natural language and spoken procedural language, depending on the performance of the dialogue, the context of the situation and the characteristics of the task.	○	○	○	○	○
	HF-24: 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-25: 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-26: The applicant should design the AI-based system to minimise the likelihood of design-related end-user errors.	○	○	○	○	○
	HF-27: The applicant should design the AI-based system to minimise the likelihood of HAIRM-related errors.	○	○	○	○	○
	HF-28: The applicant should design the AI-based system to be tolerant to end-user errors.	○	○	○	○	○
	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.	○	○	○	○	○

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-31: The applicant should design the system to be able to diagnose the failure and present the pertinent information to the end user.	○	○	○	○	○
	HF-32: The applicant should design the system to be able to propose a solution to the failure to the end user.	○	○	○	○	○
	HF-33: The applicant should design the system to be able to support the end user in the implementation of the solution.	○	○	○	○	○
	HF-34: The applicant should design the system to provide the end user with the information that logs of system failures are kept for subsequent analysis.	○	○	○	○	○
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, would be necessary to mitigate the residual risks to an acceptable level.	●	●	○	○	
	SRM-02: The applicant should establish safety risk mitigation 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 solutions in all the domains under the remit of the Agency, in line with the Rulemaking concept described in EASA AI Roadmap 2.0, EASA intends to define a common policy that can be applied to the whole of the EU civil aviation regulatory framework, while accounting for certain domain specificities.

This calls for a mixed rulemaking approach, involving on the one hand cross-domain rules (horizontal) and, on the other hand, domain-specific rules (vertical).

This is anticipated to be developed under Rulemaking task RMT.0742 (as defined in the European Plan for Aviation Safety (EPAS) 2024-2026 Volume II), in two steps:

- Step 1 of RMT.0742: development of a transversal Part-AI and a set of ‘AI trustworthiness’ AMC and GM to reference or complement where necessary applicable industry standards.
- Step 2 of RMT.0742: analysis, per domain, of those requirements that are domain-specific and those that need to be complemented to provide an adequate regulatory basis for deploying the generic AI trustworthiness framework.

This Annex provides an analysis of the anticipated impact on aviation regulations and on the MOC to the current regulations for the various impacted domains.

1. Product design and operations

In the product design and certification domain, the current implementing rules (Part 21) and CSs (both for manned and unmanned products) 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, CS 23/SC-VTOL/SC-Light-UAS.2500, 2505, 2510 are considered to still be valid for evaluating the safety of AI-based systems, provided additional MOC 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 anticipated. For Level 2B, the notion of human-AI collaboration and the partial release of human end-user authority to the AI-based system will require new guidance in the generic AI trustworthiness framework, as well as possible adaptations to the CSs and associated AMC and GM. For higher AI Levels (3A and 3B), this assumption will need to be revisited when working on further updates to this document.

Note: 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.

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 Special Conditions and project-specific Certification Review Items (CRIs) using the guidelines from this document.

Furthermore, 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.

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.

2. ATM/ANS

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 has recently been complemented with a set of regulations, with additional requirements after Regulation (EC) No 552/2004²³ has been totally repealed. These regulations are in support of the conformity assessment framework in the ATM/ANS domain. Delegated Regulation (EU) 2023/1768 lays down detailed rules for the certification and declaration of air traffic management/air navigation services systems and air traffic management/air navigation services constituents, while Implementing Regulation (EU) 2023/1769 establishes technical requirements and administrative procedures for the approval of organisations involved in the design or production of air traffic management/air navigation services systems and constituents.

In addition to these new regulations, a set of AMC & GM to the Articles of Delegated Regulation (EU) 2023/1768 as well as Detailed Specifications (DSs) with their related AMC and GM applicable to the design, or changes to the design, of ATM/ANS equipment were published in October 2023.

All 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.

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

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.

²³ 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.

In addition, some AMC & GM to Delegated Regulation (EU) 2023/1768 as well DSs will be impacted, and these impacts will be managed when entering step 2 of RMT.0742.

3. Aircraft production and maintenance

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.

For 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, no impact on the EU aviation regulatory framework in relation to CAW is anticipated. Anticipated Rulemaking activities will focus on deploying the generic AI trustworthiness framework in the frame of the current CAW regulations for safety-related applications.

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.

‘Officially recognised standards’ as mentioned in the AMC material ‘means those standards established or published by an official body, 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

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.

In this context, the AMC for the above-mentioned implementing rules should 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 generic AI trustworthiness framework (step 1 of RMT.0742) will need to be interfaced with the Aircrew Regulation in a proportionate manner.

5. Aerodromes

In addition to the Basic Regulation, Regulation (EU) No 139/2014²⁴ 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 services²⁵;
- (d) competent authorities involved in the oversight of the above organisations, certification of aerodromes and certification/acceptance of declarations of safety-related aerodrome equipment²⁶.

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.

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.

More specifically, 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.016 Foreign object debris control programme
- ADR.OPS.B.020 Wildlife strike hazard reduction
- ADR.OPS.B.037 Assessment of runway surface condition and assignment of runway condition code
- ADR.OPS.B.075 Safeguarding of aerodromes
- ADR.OPS.D.035 Aircraft parking

Furthermore, Regulation (EU) No 139/2014 and the current CSs provide a comprehensive set of requirements for the design of aerodrome infrastructure and for some aerodrome equipment (as far

²⁴ 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.

²⁵ For groundhandling services and providers of such services, there are at this stage no detailed implementing rules. These are expected not earlier than 2024.

²⁶ 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.

as it exists stemming from the transposition of ICAO 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

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'.

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, a table summarising the use cases for each of the major aviation domains has been created at the beginning of the related sections, in order to provide the audience with a description of how each use case has been classified as per Table 2.

2. Use cases — aircraft design and operations

		Aircraft design and operations			
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	Computer vision based auto-taxi	Pilot AI teaming — Proxima virtual use case
Level 1A Human augmentation	Automation support to information acquisition	camera	ATC radio communication	Data acquisition (FPL and airport charts)	Data acquisition (displayed information, pilot state, data link t ground)
	Automation support to information analysis	Pre-processing + Runway object classification + bounding box + tracking/filtering algorithm	Voice recognition	Pre-processing + Centerline and obstacle detection + tracking/filtering algorithm	Information preparation
Level 1B Human assistance	Automation support to decision-making	x	Radio frequency suggestion for pilot validation		
Level 2A Human-AI cooperation	Directed decision and automatic action implementation	x	x	Obstacle avoidance management (reasoning)	Automatic configuration of the aircraft
Level 2B Human-AI collaboration	Supervised automatic decision and action implementation	x	x	x	Communication identification and management of failure

Table 6 — Classification applied to aircraft design and operations use cases

Where:



represents the AI-based system or subsystem; and

The **AI/ML constituent** is in blue.

2.1. Visual landing guidance system — IPC CoDANN with Daedalean AG

This use case describes Daedalean's Visual Landing System (VLS), analysed in [CODANN] (EASA and 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 (including indication of the level of teaming with the AI-based system, i.e. none, cooperation, collaboration) and 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 goals and associated 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. A 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)

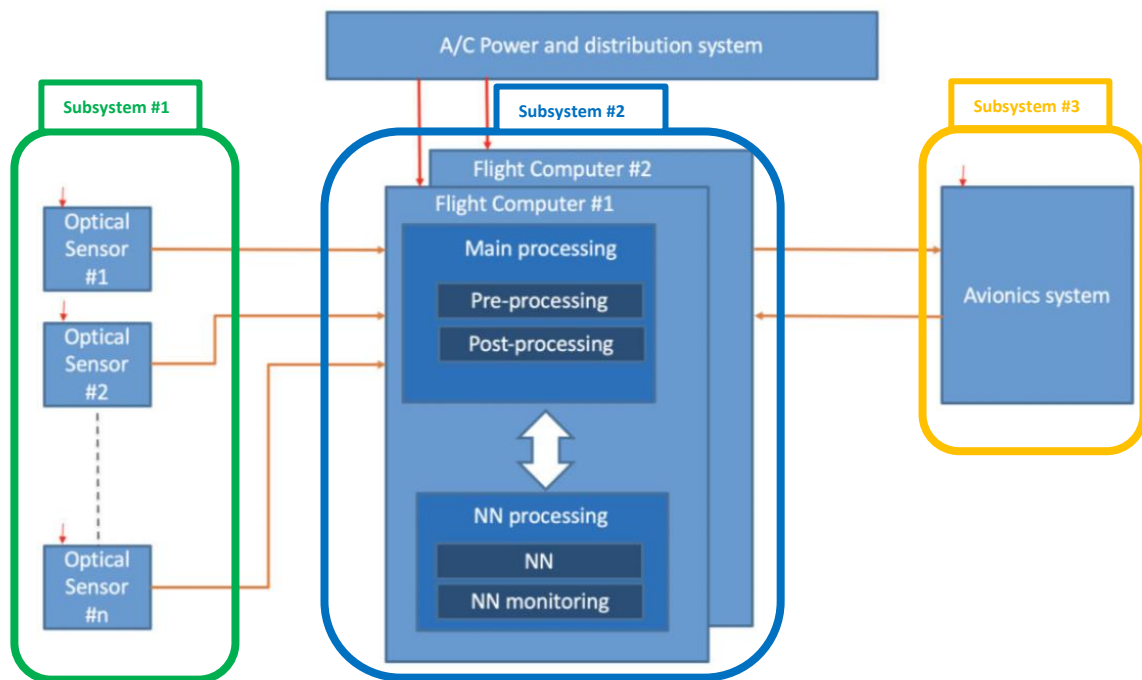


Figure 25 — System breakdown in subsystems and components (source (EASA and Daedalean, 2020))

Objective CO-06: The applicant should perform a functional analysis of the system, as well as a functional decomposition and allocation down to the lowest level.

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.

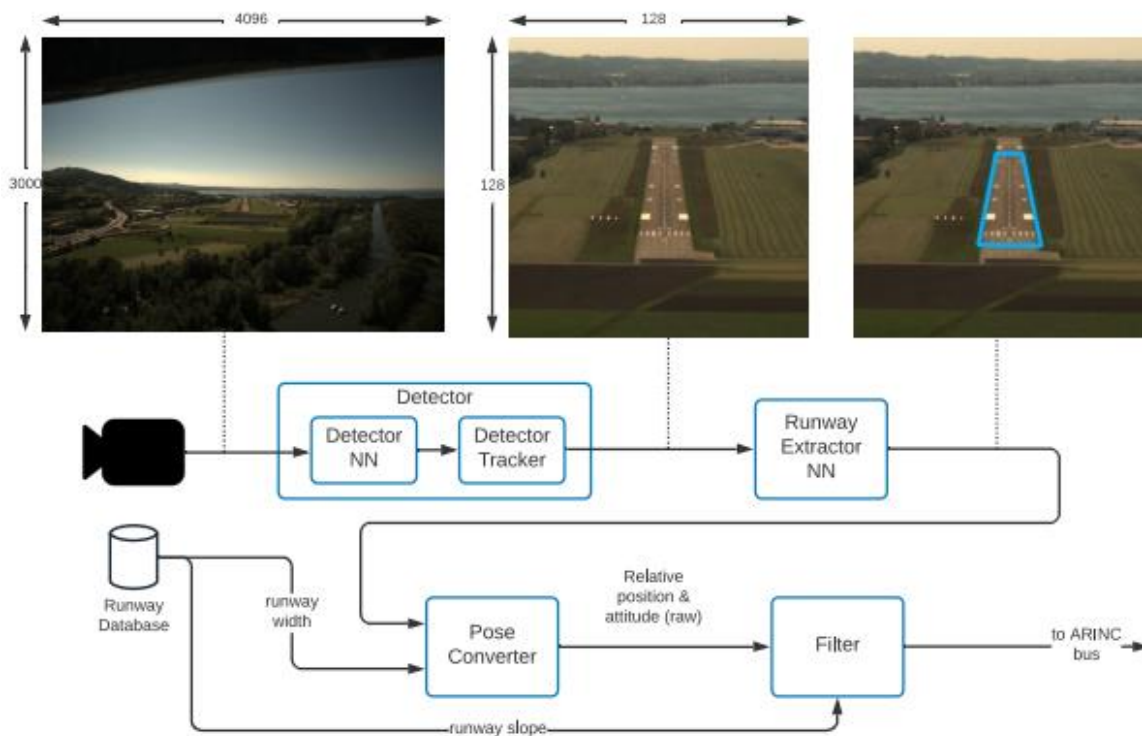


Figure 26 — System overview (source (EASA and Daedalean, 2020))

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 7 — 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, or even higher (depending on the result of the classification per **Objective CL-01**).

Based on the discussions from the CoDANN report (EASA and 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 SA-02: 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 continuously 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**, and **Objective DA-05** 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-06: The applicant should describe the preliminary AI/ML constituent architecture, to serve as reference for related safety (support) assessment and learning assurance objectives.

The systems, subsystem and AI/ML constituent 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 DA-03: The applicant should define the set of parameters pertaining to the AI/ML constituent ODD, and trace them to the corresponding parameters pertaining to the OD when applicable.

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 main relevant operating parameters pertaining to the OD include the following (on the frame level):

- Altitude
- Glide slope
- Lateral deviation
- Distance to runway
- Time of day
- Weather

In addition, at least the following operating parameters pertain to the ODD (linked to characteristics of the camera)

- Brightness

— Contrast

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-01: 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-02-SL: Once data sources are collected, the applicant should ensure that the annotated or labelled data in the data set satisfies the DQRs captured under **Objective DA-04**.

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-05: 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-06: The applicant should distribute the data into three separate data sets which meet the specified DQRs in terms of independence (as per **Objective DA-04**):

- 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-07: The applicant should ensure verification of the data, as appropriate, throughout 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 (EASA and Daedalean, 2020), Section 6.2.8, for more information.

— Data accuracy

To demonstrate that the model was provided with correct data samples during the development 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 development phase.

2.1.3.2. Learning process management

Objective LM-01: The applicant should describe the ML 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;
- explainability capabilities of the selected model;
- activation functions;
- 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 bounds.

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-SL: The applicant should account for the bias-variance trade-off in the model family selection and should provide evidence of the reproducibility of the model 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.

An example of 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-04: 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-06: 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-07: 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-08**, **Objective IMP-09**, **Objective IMP-10**, and **Objective IMP-11**, 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 of verification

Objective DM-08: The applicant should perform a data verification step to confirm the appropriateness of the defined ODD and of the data sets used for the training, validation and verification of the ML model.

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 bias, or possible misidentifications of the input space (e.g. use of non-generic runway markings or adjacent objects).

Objective EXP-09: 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, as well as the means to retrieve this data.

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-11: The applicant should ensure that the AI-based system presents 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-17: For each output relevant to the task(s), the applicant should ensure the validity of the specified explanation.

The system includes an uncertainty estimation component, estimating both the aleatoric and epistemic uncertainties.

See [FAAVLS; Chapter 5, Section 8.4].

Objective EXP-19: Information concerning unsafe AI-based system operating conditions should be provided to the 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, 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 safety risk mitigation means as identified in **Objective SRM-01**.

No safety risk mitigation 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. Auto-Taxi system — IPC with Boeing

This use case describes an experimental auto-taxi system being researched by Boeing, which is the subject of an ongoing innovation partnership contract (IPC) between Boeing and EASA. Further details will be available at a later date once the IPC is completed and a report released. This section presents a use case to help illustrate the application of the objectives discussed in this Concept Paper.

The auto-taxi system receives, via standard radio communication, taxi clearance from ground control, provides a readback of the clearance, and plans the appropriate route based on that clearance. The system then executes the plan and autonomously controls the aircraft as it travels from one location to another at an airfield, such as from the boarding gate to the departure runway.

While executing the plan, the system detects potential obstacles in the aircraft's path to which it reacts accordingly: it either stops, follows or goes around them. For the detection of obstacles, the system employs a LIDAR. Optical cameras will be added to the perception sensors for the object classification which will support an improved awareness and intent and prediction capabilities for the environment. The operation of the system is monitored by the flight crew, who retains the ability to override and disconnect the system at any time.

2.3.1. Trustworthiness analysis — description of the system and ConOps

2.3.1.1. 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).

The primary end user of the auto-taxi system is the flight crew, which is expected to be the traditional two-person flight crew of today's large commercial aircraft. It is assumed that the training requirements for, and task allocation between, the two flight crew members would be similar to standard operations. An exception though would be that the crew member normally tasked with the operation of the aircraft during taxi would instead be tasked with monitoring the auto-taxi system, with the ability to override it if necessary. Boeing acknowledges that other human interfaces will exist, such as with maintenance support personnel, ATC personnel, and training personnel; however, these are being treated as secondary interactions for the time being.

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 the traditional two-person flight deck, the tasks performed in interaction with the AI-based system are the activation, monitoring and override of the AI-based system's operations. The system will provide, via the HMI, the flight crew with feedback and controls necessary for them to monitor the operation and performance of the system, enabling the crew to react accordingly. The flight crew has the ability to amend the autonomously planned taxi route if necessary. In addition, the flight crew has the ability to deactivate the AI-based system if necessary.

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):

Similar to other use cases presented in this Concept Paper, it was decided to utilise an example scenario in order to illustrate potential operations of the auto-taxi system. While Seattle-Tacoma International airport is used in this example, the auto-taxi system requires no additional infrastructure to be installed for its use beyond what is already present at any commercial airport.

Example of a typical auto-taxi scenario

Before landing at Seattle-Tacoma International airport, the Boeing aircraft equipped with the auto-taxi system receives clearance to exit the active runway via radio communications from ground control. Once the aircraft exits the runway, it receives the clearance to taxi to the gate or to the next waypoint on its way to the gate. The auto-taxi system listens for the flight's designator or tail number, checks its validity, and listens to and processes the taxi clearance. The system can read back the clearance to the ground control and requests clarification as needed. The flight crew monitors the system's interaction with ground control and has the ability to correct if necessary.

Through its sensors, the auto-taxi system perceives the surrounding environment and detects obstacles, localises itself on the map, and plans a route based on the crew-confirmed clearance. This planned route is displayed to the flight crew, who then confirms its accuracy and activate the system. Once activated, the system operates the necessary aircraft controls (e.g. thrust, nosewheel steering, brakes, etc.) and navigates the aircraft along the planned route.

While taxiing, ground control can change the clearance or provide additional clearance details (such as stopping instructions at certain points). The system processes the clearance modifications, adjusts the route if necessary, and provides a readback. Hold-short locations have to be overridden by the clearance (i.e. the aircraft will hold short unless instructed otherwise). For all clearances, the system checks the validity of the communications as well as the entities they are addressed to. At all times the flight crew monitors the system, retaining the ability to override it and take control of the aircraft at any time. Once the aircraft reaches its gate or designated location, the system stops the aircraft and then is deactivated.

2.3.1.3. Description of the subsystems involved (inputs, outputs, functions)

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

During the development of the experimental auto-taxi system, the research team consulted with a human performance expert and a representative member of the end user group identified in Objective CO-01 to understand their assigned tasks in the aviation ecosystem, how those tasks will be affected by the introduction of the auto-taxi system, and how the auto-taxi must be designed in order to safely execute these tasks. This input is translated into requirements that are levied upon the system, and these requirements are logged and tracked as part of the validation & verification process for the system.

Objective CO-06: The applicant should perform a functional analysis of the system, as well as a functional decomposition and allocation down to the lowest level.

STPA (System-Theoretic Process Analysis), as described in the STPA Handbook is '... a relatively new hazard analysis technique based on an extended model of accident causation²⁷.' The STPA process provides a method to perform an early functional analysis of the system in question.

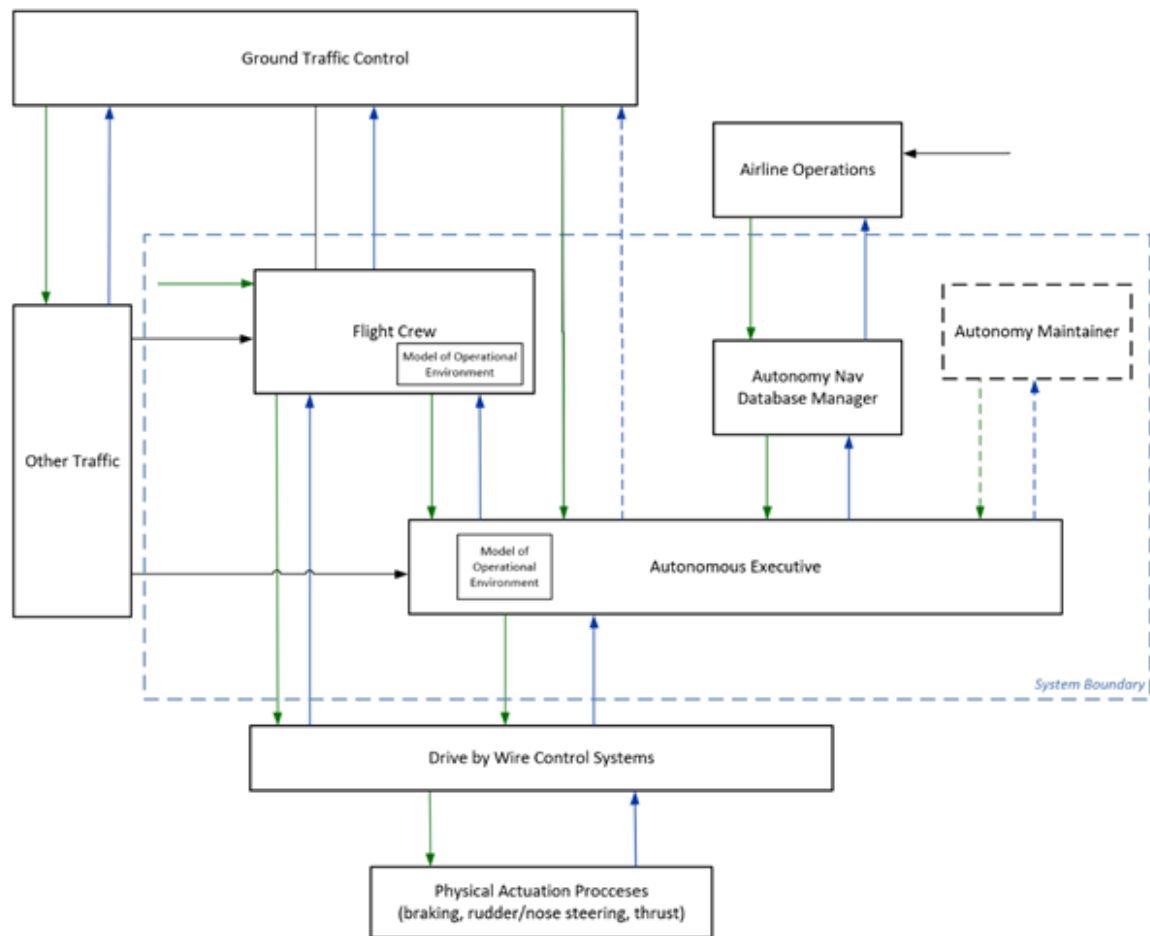
The system's architecture is represented via a control structure model, which 'captures functional relationships and interactions by modeling the system as a set of feedback control loops.' Based on stakeholder needs and the ConOps, the responsibilities of each controller in the system are documented as part of the control structure creation process, from which formal functional allocations and decomposition can be readily derived. Functions relevant to each control loop are captured in the form of control actions and feedback, as well as necessary communication between controllers with equal authority or entities outside the system boundary. It is important to note that 'The hierarchical control structure used in STPA is a functional model, not a physical model like a physical block diagram, a schematic, or a piping and instrumentation diagram (...) A control structure will emphasize functional relationships and functional interactions.'

The subsequent steps in STPA (analysis of unsafe control actions and causal scenarios) identify the functional requirements necessary to ensure properly constrained behaviour of the overall system. Below is the preliminary control diagram and responsibilities for the auto-taxi system as generated during the STPA process. As STPA is an iterative process, the control structure model and its representation of controller responsibilities and functional interactions are expected to evolve as STPA continues and the system design matures.

²⁷ PA handbook, Nancy Leveson and John Thomas.
http://psas.scripts.mit.edu/home/get_file.php?name=STPA_handbook.pdf

Auto Taxi Project

High Level Control Structure for existing system



Note: The arrows between controllers denote information flows.

Figure 27 — High-level control structure

ID	Controller	Responsibility
1	ATC	Provide oversight of multi-aircraft operations for the entire airport Manage strategic clearance/deconfliction between aircraft Understand the current position and intent of every aircraft on the airport Provide clearance to each aircraft and receive readback
2	Pilot(s)	Ensure safe operation of the aircraft, ultimate authority over operation/movement of the aircraft -Oversight of the auto-taxi system and intervening or shutting it off if required -Modification of waypoints if required -Providing intent to ATC (current state) -Understanding the intent of the auto-taxi system through requesting explanations from the system Ensure the navigation database is using current airport maps Ensure they are up to date with any applicable NOTAMs Determine whether the expected environmental conditions are appropriate for use of auto-taxi
3	Navigation database controller/airline operations (including data loading)	Provide the accurate and up-to-date airport maps to the system and pilot
4	Other aircraft traffic	Follow ATC commands Follow existing standard airport operational rules: taxi speed max 30 knots, 10 knots with turns
5	Autonomous executive	Provide readback (future state) Provide automation's situation representation and explainability (intent, goals, situation awareness, constraints) to pilot in real time and more detailed log to be used by maintenance for post-operations explainability & auditability Follow airport specific restrictions (e.g. a taxiway restricting valid aircraft type due to size) Receive clearance, plan route, control aircraft as necessary to follow route Ask pilot for approval of route before execution Obstacle detection and avoidance Determination of location in order to load correct airport map
6	Existing aircraft control systems (control laws, envelope protection, thrust, brakes, etc., drive by wire)	Translate pilot or automation commands into effector commands Provide back-drive of inceptors
7	Control surfaces (rudder, nose gear, brakes, thrust (including differential thrust))	Provide effector motion

Table 8 — Breakdown of responsibilities

2.3.1.4. Expected benefits and justification for Level 2A

The auto-taxi system may enhance the safety of the very busy taxi phase, by providing extra awareness of the environment and the contents of the clearances. It may also reduce the workload on the flight crew during the taxi phase of flight by allowing one flight crew member to monitor the system rather than having to actively control the aircraft. It may additionally reduce the workload on ground control by eliminating the need for progressive taxi instructions for aircraft equipped with the system.

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

The Level 2A classification is justified by the fact that:

- The auto-taxi system controls the aircraft in order to navigate it along the planned route; therefore, it is not an advisory system and is beyond the scope of Level 1A or 1B.
- Per the HAT discussion in Section 6.4, the auto-taxi most closely aligns with the description of human-AI cooperation where the AI-based system is assigned a predefined task to complete the flight crew's goal of safely taxiing the aircraft to a specific destination. Tasks are not dynamically allocated between the system and the flight crew as described in the Level 2B human-AI collaboration.
- One additional point of discussion is the question of authority: the Concept Paper's description of a Level 2A system shows that full authority remains with the expert pilots (human or humans). As regards the auto-taxi system, the human retains the ability to override the system at any point, except for when the system provides the clearance readback to ground control. This step is taken automatically without a separate human approval. This is authority that is allocated from the human to the machine, resulting in the human only having partial authority which would meet the description of a Level 2B system. In all other respects, the auto-taxi system matches the description of a Level 2A system not a 2B. While the classification of any particular system will be a discussion between the applicant and regulator on a case-by-case basis, in general it is thought that a system should be classified to the Level it most closely aligns with, and that having one or a small number of aspects that are demonstrative of a higher level should not entail it being classified as that higher level.

2.4. 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.

2.4.1. Trustworthiness analysis — description of the system and ConOps

2.4.1.1. 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 should 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 build its situation representation 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 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 HAI 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).

2.4.1.2. 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:

Aircraft SPO, automated flight control system failure and AP loss

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	<ul style="list-style-type: none"> - 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	Keyboard CCD Touchscreens	Conventional hardware systems	Haptic information	Pilot state detection
Galvanic Response	Skin contact with aircraft controls	'Sweat' rate – skin conductivity		
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 9 — Proxima user interface possibilities

2.4.1.3. 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:

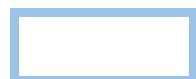
- Proxima and the PIC have a set of common goals and share tasks;
- Proxima is capable of using natural language, of taking decisions and implement actions that are overseen and overridable by the pilot; and
- Proxima is capable of having some authority in decision-making, to share situation awareness and react in real time to changes by adjusting strategies and task allocation.

3. Use cases — ATM/ANS

		ATM/ANS	
EASA AI Roadmap AI Level (subsystem)	Function allocated to the (sub)systems (adapted HARVIS LOAT terminology)	AI-based augmented 4D trajectory prediction	Time-based separation + Optimum runway delivery
Level 1A Human augmentation	Automation support to information acquisition	Data acquisition (FPL and updates, radar + weather)	Data acquisition (weather + radar)
	Automation support to information analysis	4D trajectory calculation – Curtain + Climb and descent rate	Information preparation (pairs, applicable separation)
Level 1B Human assistance	Automation support to decision-making	x	Trajectory prediction + uncertainty calculation
Level 2A Human-AI cooperation	Directed decision and automatic action implementation	x	x
Level 2B Human-AI collaboration	Supervised automatic decision and action implementation	x	x

Table 10 — Classification applied to ATM/ANS use cases

Where:



represents the AI-based system or subsystem; and

The **AI/ML constituent** is in blue.

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 11 — 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 12 — 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 Bar²⁸](#)
 - [Threat Modeling AI/ML Systems and Dependencies²⁹](#)
 - [Failure Modes in Machine Learning³⁰](#)
- [A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View \(Liu, 2018\)](#)
- MITRE [Adversarial ML Threat Matrix³¹](#).

The objective is to establish different potential attack paths and identify possible shortcomings.

As illustrated in Figure 28, 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.

²⁸ Source: <https://docs.microsoft.com/en-us/security/engineering/bug-bar-aiml>

²⁹ Source: <https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml>

³⁰ Source: <https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning>

³¹ Source: <https://github.com/mitre/advmthreatmatrix/blob/master/pages/adversarial-ml-threat-matrix.md>. Latest commit: Oct 23, 2020.

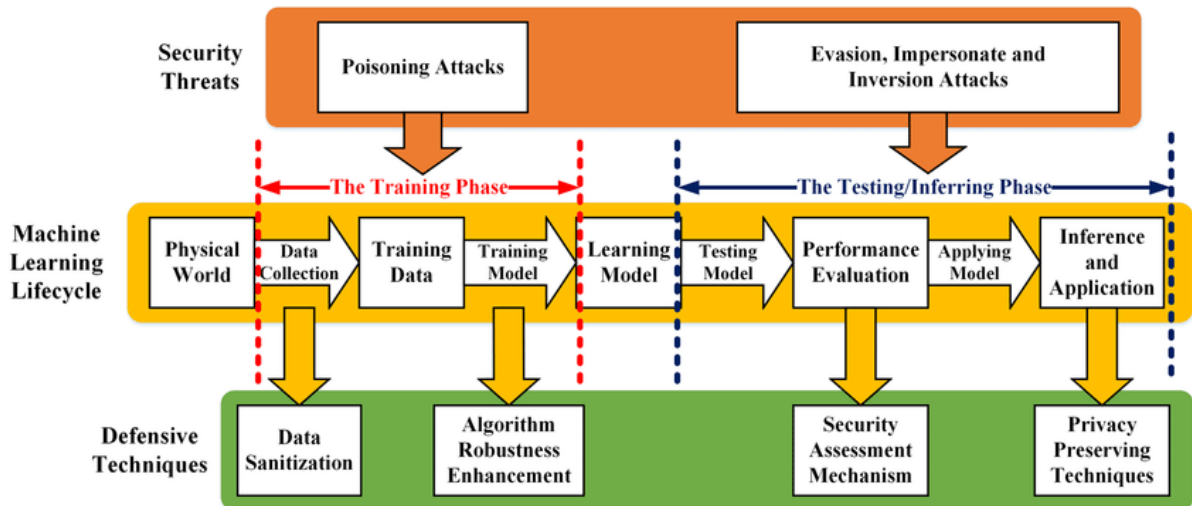


Figure 28 — Illustration of defensive techniques of ML

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 29 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.

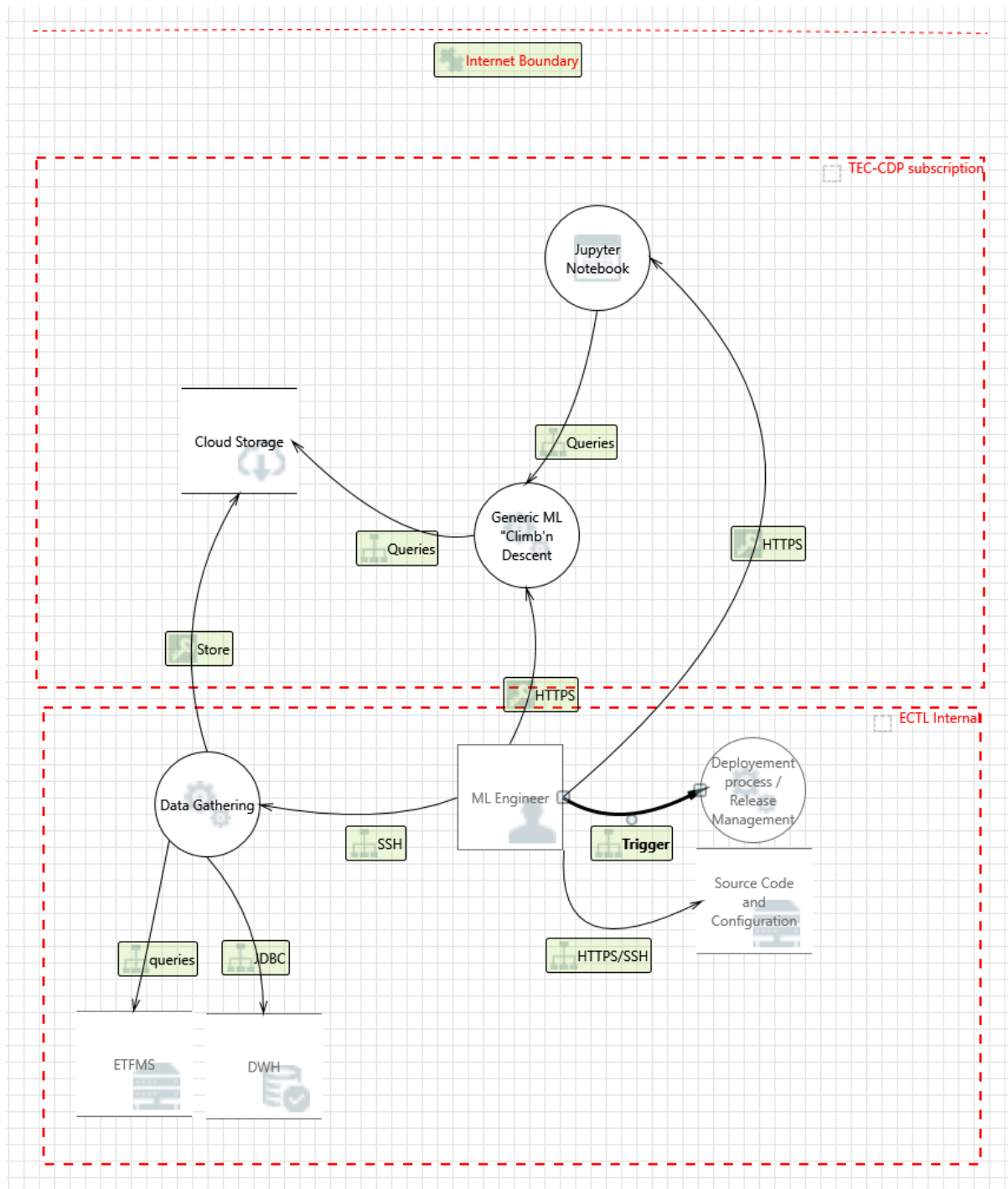


Figure 29 — Modelisation of the 'climb and descent' ML system

— 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.
- Training data 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 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 ARU³² 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 labels³³ 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:

$$T = T_{prev} + \frac{T_{next} - T_{prev}}{F_{next} - F_{prev}}(F - F_{prev})$$

³² 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).

³³ 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.

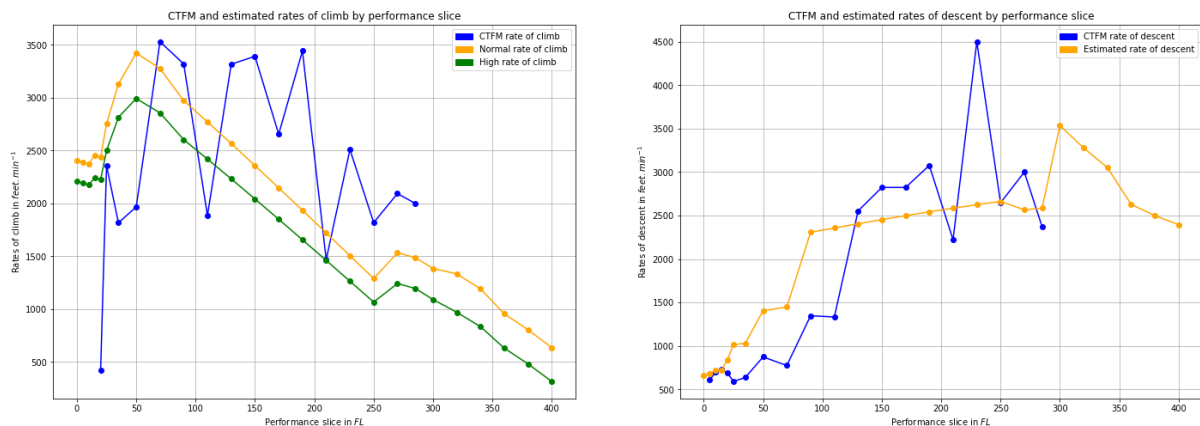


Figure 30 — Climb and descent rates by performance slice

— 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-04: The applicant should define and document pre-processing operations on the collected data in preparation of the model training.

— Data cleaning

Several data cleaning operations were performed, including the removal of yo-yo flights³⁴ (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.

³⁴ 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 13 — 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.

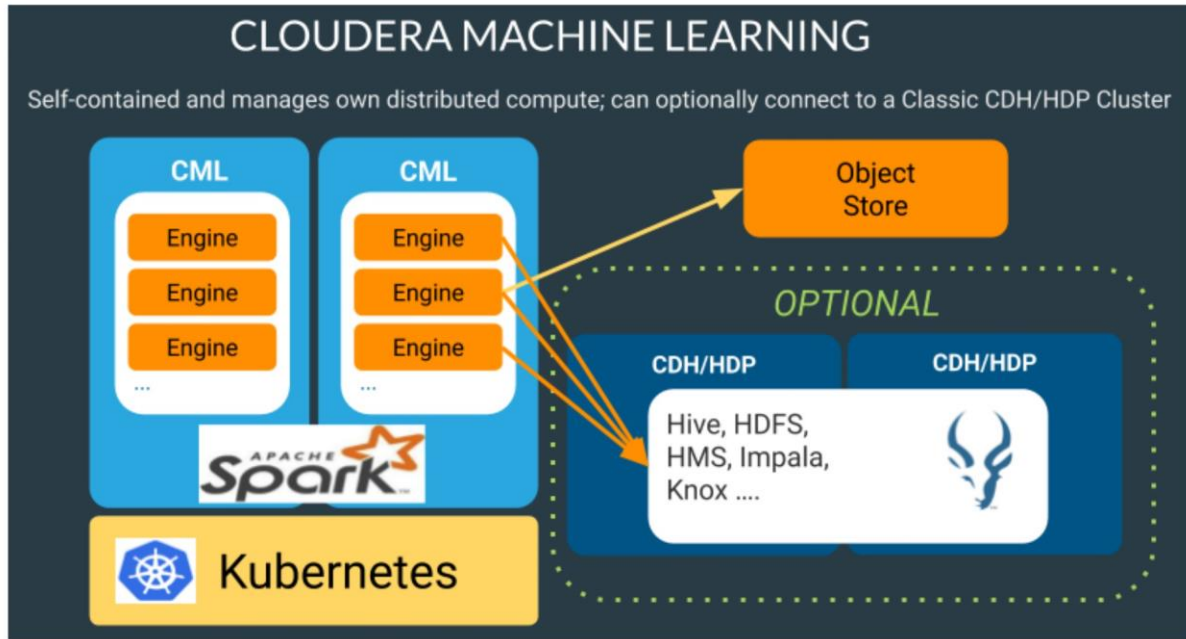


Figure 31 — Cloudera machine learning environment

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.

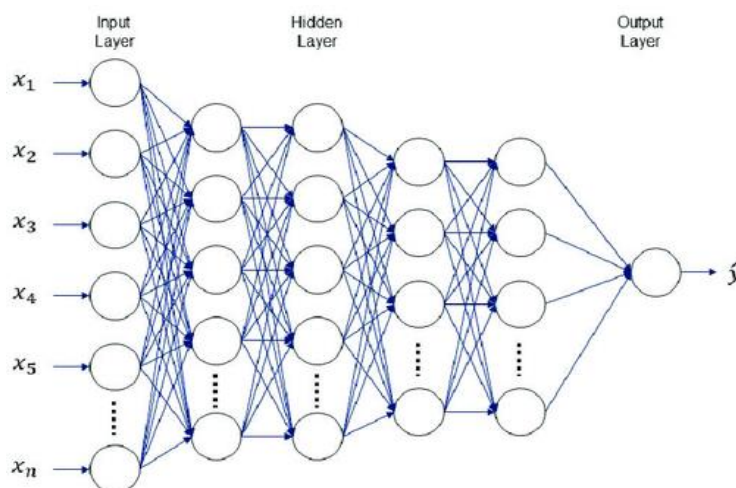


Figure 32 — DNN structure

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 14 — 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 15 — Key elements of the DNN

b. Hosting the model predictor

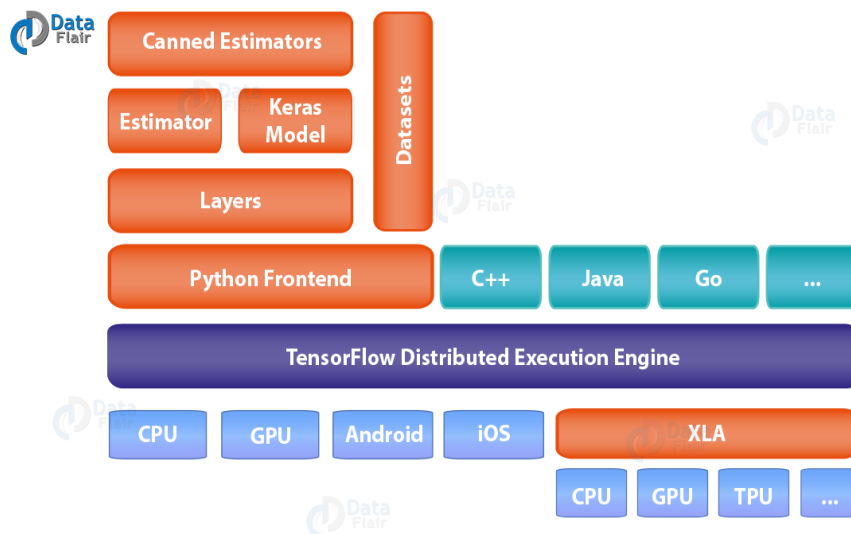


Figure 33 — Tensorflow component model and dependencies

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 16 — 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:

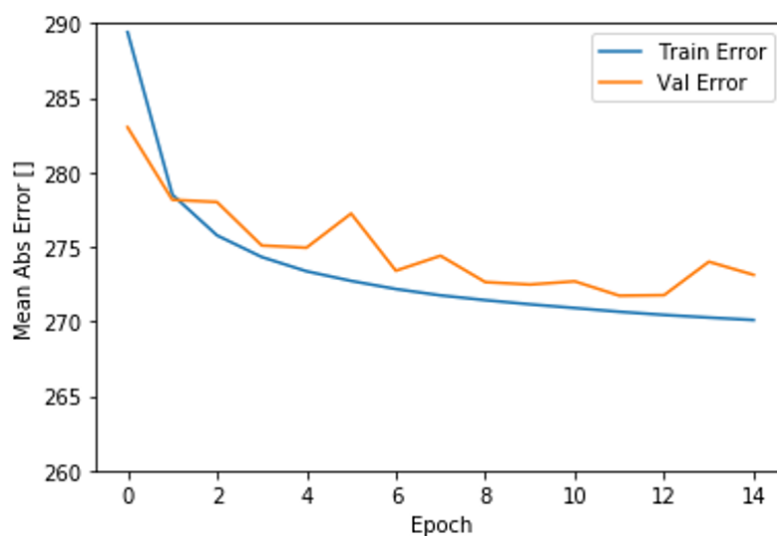


Figure 34 — Model training (mean absolute error)

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.

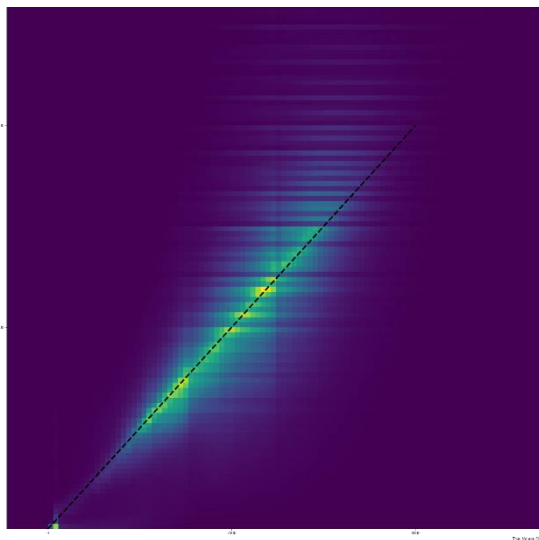


Figure 35 — Predicted climb rate (with BADA) v actual from CTFM

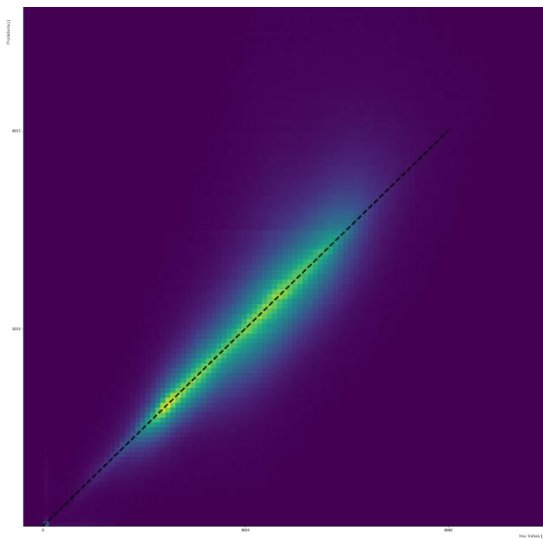


Figure 36 — Predicted climb rate (with ML) v actual from CTFM

b. Comparison of error rates between current (FTFM) and new ML calculation

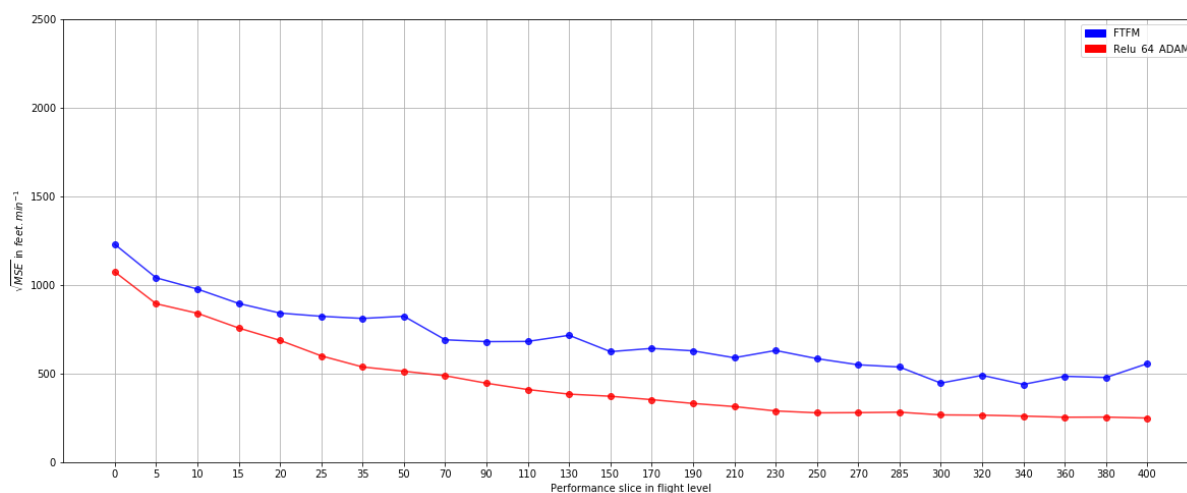


Figure 37 — Mean square error on actual climb rates (with low-pass filter)

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’:

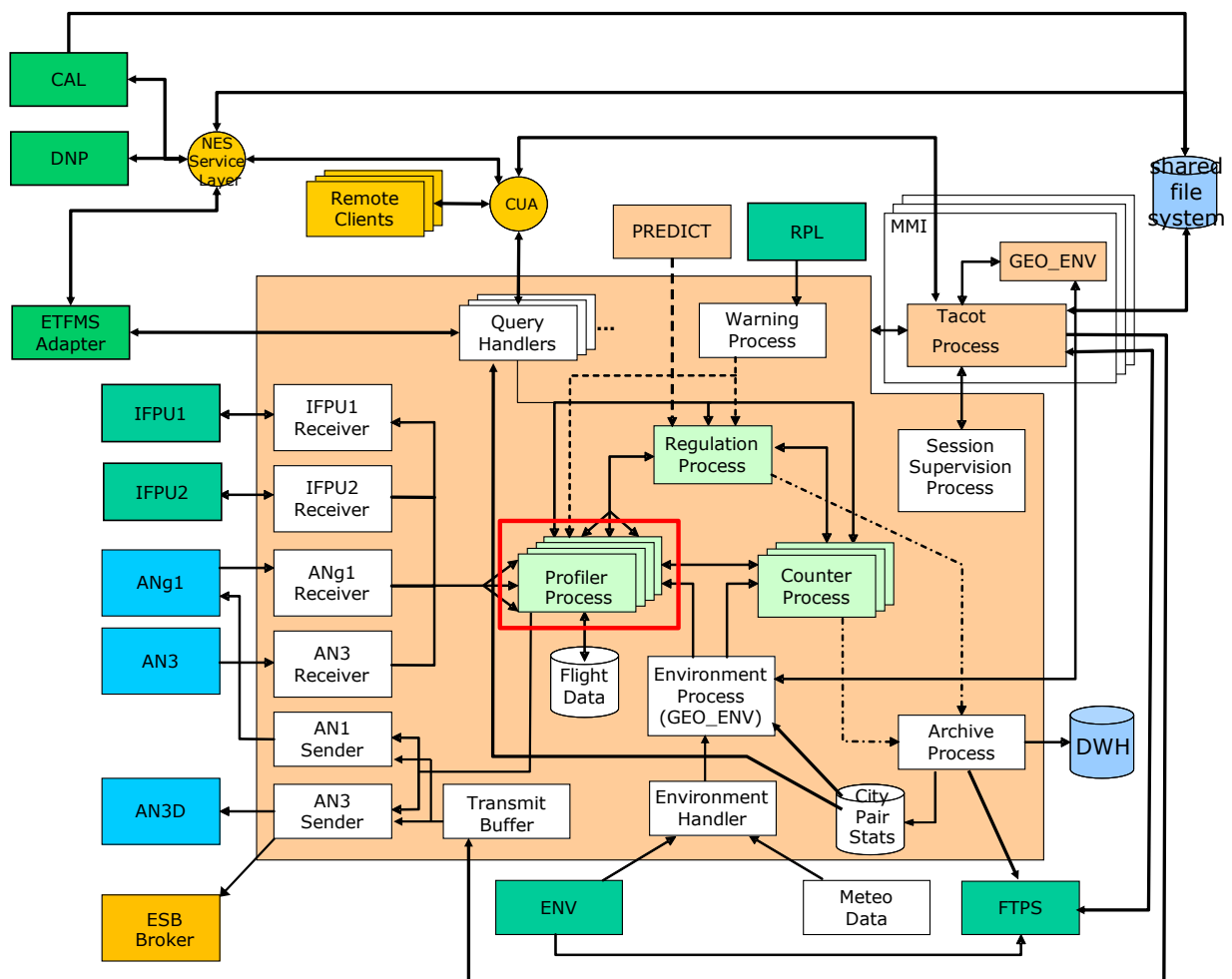


Figure 38 — ETFMS logical architecture

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:

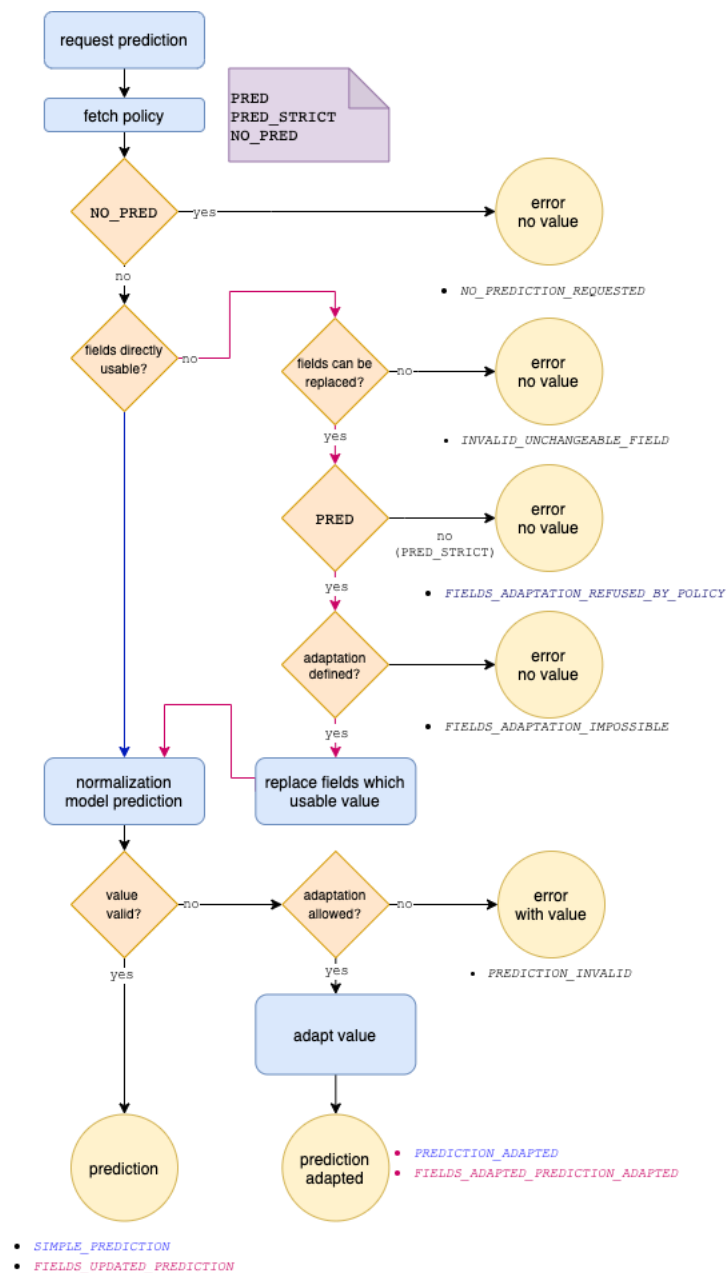


Figure 39 — Workflow for prediction delivered by the AI/ML library

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
<i>BL</i>	283 051	1 860 046	544 428	3 226 165	285 194	1 783 651
<i>ML</i>	265 889	1 655 747	514225	2 880 420	272 071	1 632 486
<i>Improv.</i>	6,06 %	10,98 %	5,54 %	10,71 %	4,60 %	8,48 %

Table 17 — Improvements on the network

Objective IMP-07: The applicant should perform 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

The objective of the use case is to extend the concept of time-based separation (TBS) on final approach which has already been developed by EUROCONTROL, integrated and deployed for certain airports. The concept is even optimised with the introduction of ML constituent(s).

Note 1: The objectives referred to in this use case are traceable (in numbering and text) to the ones developed in this Issue 02 of the EASA Concept Paper 'Usable guidance for Level 1&2 machine learning applications'.

Note 2: The following provides a partial view of a future complete description of the case, especially in the context of the transition towards the updated regulatory framework for ATM/ANS (introduction of the set of regulations for conformity assessment of ATM/ANS systems and ATM/ANS constituents, i.e. (EU) 2023/1768 and (EU) 2023/1769). Indeed, the development made by EUROCONTROL needs to be endorsed by a DPO organisation that is in charge of the integration of the developed library into its ATM/ANS equipment subject to certification, declaration or statement of compliance. It is the final responsibility of the ANSP to obtain the approval for the subsequent functional system that makes use of the ATM/ANS equipment embedding the functionality.

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.

3.2.1. Trustworthiness analysis — description of the system and ConOps

3.2.1.1. 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 end-user responsibilities and expected expertise (including assumptions made on the level of training, qualification and skills).

The main end users of the functionality are (refer to Annex A to Calibration Of Optimised Approach Spacing Tool using Machine Learning techniques):

- Tower ATC roles:
 - tower ATC supervisor
 - tower runway controller
- Approach ATC roles:
 - approach supervisor
 - final approach controller
 - intermediate approach controller

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.

The high-level tasks in relation to the AI-based system are described in the table in Annex A to Calibration Of Optimised Approach Spacing Tool using Machine Learning techniques).

For example, at final approach, the final approach controller ensures that the final approach separations are set up consistently and efficiently.

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

As illustrated in Figure 40, the TBS-ORD system is composed of two subsystems:

- A final target distance (FTD) computation subsystem that calculates, based on the sequence, traffic surveillance and meteorological (MET) data inputs, the FTD distance for each pair of the sequence and output it to the HMI system responsible for displaying the FTD chevron at a distance corresponding to the FTD behind the leader aircraft;
- An initial target distance (ITD) computation subsystem that calculates, based on the sequence, traffic surveillance, MET data inputs and the FTD value calculated by the FTD computation subsystem, the ITD distance for each pair of the sequence. It outputs it to the HMI system responsible for displaying the ITD chevron at a distance corresponding to ITD behind the leader aircraft.

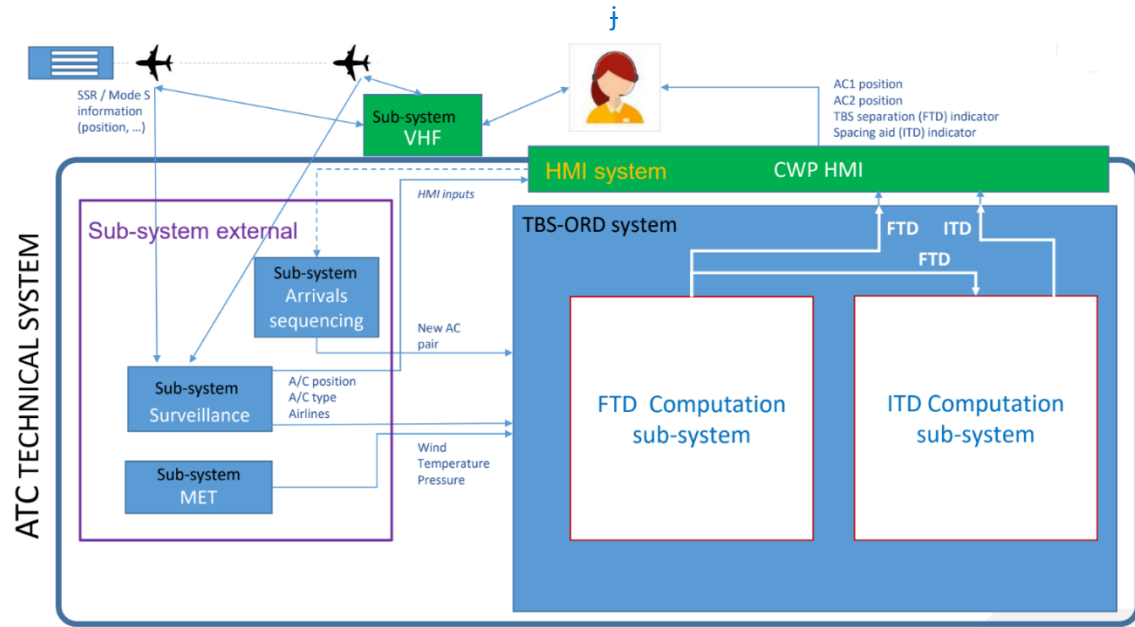


Figure 40 — TBS-ORD system functional architecture

3.2.1.2. Concept of operations for the AI application

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.

Headwind conditions on final approach cause a reduction of the aircraft ground speed which, for distance-based separation, results in an 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.

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 (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 (ITD): a spacing indicator to optimise the compression buffers ensuring optimum runway delivery (ORD). Both indicators are shown in Figure 41.

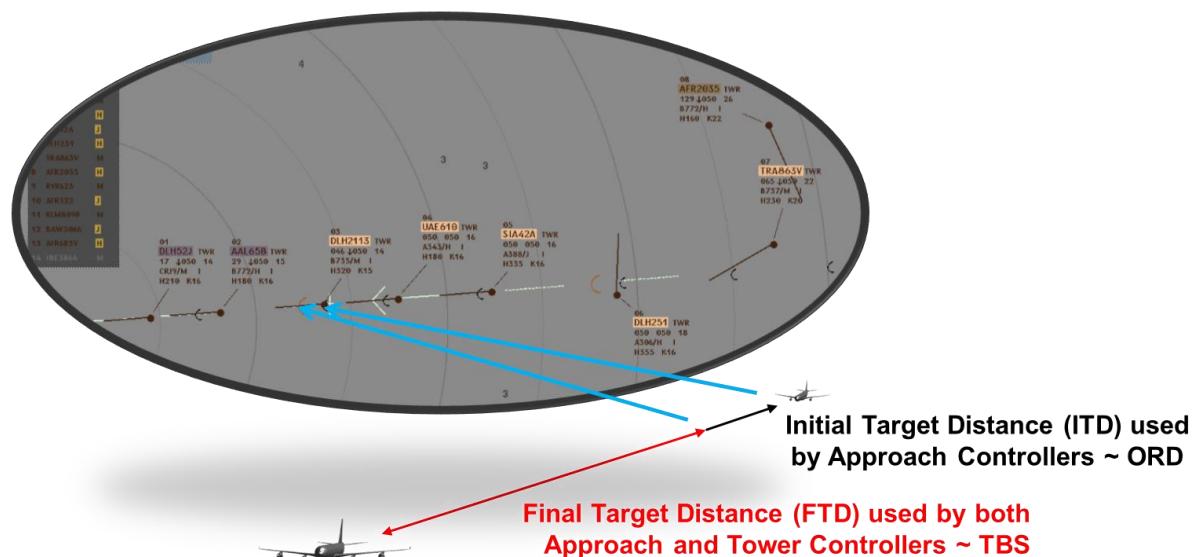


Figure 41 — Representation of FTD and ITD in the ATCO's separation delivery tool

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.

3.2.1.3. Classification

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

The trigger for classifying an application in Level 2 is the presence of human-machine teaming. This is not the case for the TBS-ORD application. There is no task reallocation, no collaboration process, no 2-way communication between human and AI-based system, no feedback loop, no situation representation for the AI-based system. The end user (i.e. the ATCO) keeps the full authority on the decision (i.e. safely managing the arrival traffic). Given these properties, it is concluded that the TBS-ORD system is Level 1.

Moreover, the introduction of ML in a TBS system does not change the operational flow for the FTD definition. The ATCO receives the target separation from the system instead of a distance matrix. The need for an indicator is a consequence of the introduction of TBS and not specifically due to the use of ML. Thus, it is considered that the impact on the operations is low. Therefore, the FTD subsystem is classified as Level 1A.

On the other hand, ORD introduces a change in operations. The ITD is a completely new indicator that did not exist in former operation flows. The ATC is typically neither used to defining this indicator by itself nor trained to do so, and must rely on the ITD subsystem. We then consider that the ITD subsystem as being Level 1B. Yet, the ITD is a spacing aid, but the controller is still free to space the aircraft at a larger or lower distance than the ITD as far as the FTD separation is guaranteed.

Since the system classification level should be at least equal to the highest level of all its subsystems, we consider the full TBS-ORD system as being Level 1B.

3.2.2. Trustworthiness analysis

3.2.2.1. Safety assessment

As mentioned under Note 2 in the introduction to the use case, the full safety assessment, also called local safety assessment, is to be performed by the ANSP aiming at deploying TBS-ORD. However, EUROCONTROL has developed a generic safety assessment in anticipation of the local one. The following presents the main elements of this generic safety assessment, organising the description based on the objectives and associated anticipated MOC set in the guidance.

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.

d. Identification of hazards

Three hazards have been identified (for details, refer to Annex A to Calibration Of Optimised Approach Spacing Tool using Machine Learning techniques):

- Wake turbulence encounter: by introducing possible dynamic variation of separation minima as function of head wind, with the application of fixed TBS minima between successive arrivals, some aircraft pairs will be exposed to different wake age compared to their current situation under the DBS scheme, and the TBS may impact the wake strength that can be potentially encountered.
- Runway collision: In TBS compared to DBS operations, due to the reduced distance spacing, more traffic pairs are expected to be spaced closer on a distance basis in the ground referential system which might be inconsistent with current aircraft runway occupancy time (ROT). As a consequence, there may be more instances of minimum time spacing at the runway threshold, resulting in more instances where there is insufficient spacing for clearance to land with potential for leading to runway incursions. This potential impact on runway collision risk needs to be mitigated in the design and safety assessment of the new solutions.
- Mid-air collision: The ATC operational application (i.e. separation delivery) of TBS might potentially lead to traffic being spaced closer than the local provisions set for airborne surveillance distance separation minima, hence affecting mid-air collision risk on approach. This potential impact on mid-air collision risk needs to be mitigated in the design and safety assessment of TBS-ORD.

e. Determination of safety criteria

Based on the identified hazards, a set of safety criteria has been determined. For details, refer to Annex A to Calibration Of Optimised Approach Spacing Tool using Machine Learning techniques. It is to be noted that these safety criteria have been quantitatively expressed in terms of proxies (see AMC1 and AMC2 to point ATS.OR.210(a) of Regulation (EU) 2017/373).

The first five safety criteria relate to wake turbulence encounter, three safety criteria are established on the risk associated with mid-air collision, and a final one relates to runway collision. The table below presents the safety criteria:

ID	Description of the safety criterion
TBS-SAC#1	The probability per approach of wake turbulence encounter of a given severity for a given traffic pair spaced at TBS minima on the final approach segment for any applicable head wind conditions shall not increase compared to the same traffic pair spaced at reference distance WTC-based minima (i.e. DBS minima) in reasonable worst-case conditions recognised for WT separation design.
SAC#F1	The probability per approach of wake alive ahead with large under separation during interception & final approach shall be no greater in operations based on TBS than in current operations applying reference DBS minima.
SAC#F2	The probability per approach of unmanaged under-separation (WT) in adequate separation mode during interception and final approach shall be no greater in operations based on TBS than in current operations applying reference DBS minima.
SAC#F3	The probability per approach of unmanaged under-separation (WT) during interception and final approach shall not increase due to inadequate selection of or transition between any adequate modes of operation.
SAC#F4	The probability per approach of imminent infringement (WT) during interception and final approach shall be no greater in operations based on TBS than in current operations applying reference DBS minima.
SAC#F5	The probability per approach of crew-/aircraft-induced spacing conflicts during interception and final approach shall be no greater in operations based on TBS-ORD than in current operations applying reference DBS minima.
SAC#F6	The probability per approach of imminent collision during interception and final approach shall be no greater in operations based on TBS-ORD than in current operations applying reference DBS minima
SAC#F7	The probability per approach of imminent infringement (radar separation) during interception and final approach shall be no greater in operations based on TBS-ORD than in current operations applying reference DBS minima.
SAC#R1	The probability per approach of runway conflict resulting from conflicting ATC clearances shall be no greater in operations based on TBS-ORD than in current operations applying reference DBS minima.

As a summary, it is to be considered that the safety strategy for the application of TBS is ‘as safe as before’ in line with the requirement in point ATS.OR.210(c) of Regulation (EU) 2017/373.

f. Safety requirements

The following design criteria are established.

- From TBS-SAC#1: FTD design criteria established in the requirements COAST-FTD-010 and COAST-FTD-020

- When using the separation delivery tool, the FTD (via the buffer used for its computation), the ITD (via the additional spacing to anticipate compression effect) and the alerts (automatic display of FTD, catch-up alert) will contribute to prevent the occurrence of under-separation.
 - This is translated in Requirements on FTD calculation COAST-FTD-010 to COAST-FTD-150, see Sections 3.1.2, 3.1.3, and 3.1.4 (FTD design criteria, FTD computation process verification, FTD generic verification) and 3.3.1 (FTD calculation) of (EUROCONTROL, 2021).
 - This is translated in Requirements on ITD calculation COAST-ITD-010 to COAST-ITD-90, see Section 3.1.5, 3.1.6, and 3.1.7 (ITD design criteria, ITD computation process verification, ITD generic verification) and 3.3.2 (ITD calculation) of (EUROCONTROL, 2021).
- The FTD minima have been capped to the surveillance distance separation minima as translated in Requirements on FTD calculation COAST-FTD-110 and COAST-FTD-150, see Section 3.3.1 (FTD calculation)
- The FTD minima also account for ROT spacing constraint as translated in Requirements on FTD calculation COAST-FTD-110 and COAST-FTD-150, see Section 3.3.1 (FTD calculation).

The table below identifies the set of safety requirements coming from design criteria. The table is limited to requirements on the FTD:

ID	Description of the safety requirement
COAST-FTD-010	In TB mode, the FTD computed by the tool to indicate the wake separation applicable at the delivery point shall take into consideration: <ul style="list-style-type: none"> • the time separation from the wake turbulence separation table; • the aircraft pair (from the arrival sequence list); • the follower time-to-fly profile obtained from modelled time-to-fly profile in the considered headwind conditions; • the time separation buffer considering uncertainties of final approach speed profiles of the aircraft pair and of the glide slope wind prediction.
COAST-FTD-020	For the TBS wake separation, the FTD shall be calculated such that the time separation distribution at FTD minima in any wind condition matches the time separation distribution obtained at DBS minima in low wind. This constraint is expressed through the matching of three quantiles: TBSp1, TBSp10, TBSp50 (respectively quantiles 1, 10 and 50 of the time separation distribution observed in low wind conditions when applying DBS minima). The application of TBS must then guarantee that no more than 1 % of the pairs are delivered below the TBSp1, no more than 10 % below the TBSp10 and no more than 50 % below the TBSp50.
COAST-FTD-030	The FTD predictions obtained using the COAST time-to-fly and FTD buffer models and FTD coverage functions shall be assessed based on an independent local test set of surveillance and meteorological data observations showing that for this database the FTD design criteria are met.
COAST-FTD-040	The FTD design criteria shall be verified on the complete test data set and also when considering subparts of it based on the most important feature selection covering at least leader and follower wake turbulence category, surface head and cross wind and most represented follower aircraft types.

COAST-FTD-050	The ML-based FTD results shall be shown to be in line (the real criteria should be defined here) from a statistical point of view with the results of an analytical physics-based model.
COAST-FTD-060	The 'extreme' ML-based FTD result cases (i.e. cases showing the largest differences) shall be investigated, characterised and understood.
COAST-FTD-070	The TBS-ORD shall compute an FTD for all pairs whatever the prevailing applicable separation/spacing constraint.
COAST-FTD-080	If using DBS mode for an individual pair, the TBS-ORD shall use the applicable DBS minima.
COAST-FTD-090	If using TBS mode for an individual pair, the TBS-ORD shall use the applicable TBS minima.
COAST-FTD-100	If using TBS mode, in case no TBS minimum is defined for the follower aircraft type considered, the TBS-ORD shall use the applicable DBS minimum.
COAST-FTD-110	For each aircraft arrival in the approach arrival sequence, all applicable in-trail and not in-trail separation and spacing rule(s) shall be selected by the TBS-ORD and the corresponding FTD shall be computed. This shall include: <ul style="list-style-type: none"> • MRS, the minimum radar distance separation • Wake turbulence separation: <ul style="list-style-type: none"> o minimum TBS o minimum DBS • ROT (of the leader flight) spacing
COAST-FTD-120	For all time-based separations and spacings, the TBS-ORD shall compute the corresponding distances using the expected time-to-fly profile.
COAST-FTD-130	For all time-based separation and spacings, additional buffer shall be added to the corresponding distances calculated by the TBS-ORD in order to account for time-to-fly uncertainty.
COAST-FTD-140	In TBS mode, if the distance corresponding to the TBS plus buffers calculated by the TBS-ORD is larger than the applicable DBS minimum, the separation shall be set to the DBS minimum.
COAST-FTD-150	For each arrival pair, the most constraining of all applicable separation or spacing distance values computed by the TBS-ORD shall be sent for FTD indication.

Note that a similar table exists for safety requirements established from design criteria on ITD.

g. Identification of the assurance level

The TBS-ORD system, the FTD computation subsystem, and the ITD computation subsystem are envisaged to be allocated with a SWAL-3 assurance level. This SWAL level should however be confirmed with the ANSP and the system integrator, and agreed upon with the local authority on the basis of the results of the local safety assessment as per ED-153.

h. Quantitative considerations

As recalled in the ConOps description, the objective of TBS-ORD is to display on the controller working position (CWP) two indicators, FTD (respectively ITD), providing the separation minimum to be applied

at runway threshold (respectively spacing to be applied before leader deceleration). As an example, we here only focus on FTD for wake separation.

The FTD design objective can thus be expressed as ‘If a follower aircraft is positioned exactly on the FTD when the leader reaches the runway threshold (separation delivery point), every separation or spacing constraint shall be respected. When operating TBS, in terms of wake separation minima, the constraint is expressed in time. The FTD shall then be calculated such that the time separation distribution at FTD minima in any wind condition matches the reference time separation distribution corresponding to what is obtained at DBS minima for that aircraft type in low wind. The reference distribution is characterised by three statistics (median TBSp50, 10th and 1st percentiles -TBSp10, TBSp1) that shall be determined per aircraft type based on local data observations in low wind conditions.

The distribution of time separation at FTD minima shall thus be such that:

- maximum 50 % of the pairs are below TBSp50,
- maximum 10 % of the pairs are below TBSp10, and
- maximum 1 % of the pairs are below TBSp1.

The assurance framework developed to meet those design objectives is summarised in Figure 42. When using ML models (denoted predictive strategy), the FTD is calculated from:

- a predictive time-to-fly model providing an estimate of the average behaviour of a considered flight;
- buffer models quantifying the variability of the flight behaviour to cope with the three statistical design criteria (e.g. p50, p10 and p1 for TBS).

Because of the lack of data, the ML models cannot be proven to safely cover all combinations of input parameters. Therefore, coverage functions are introduced characterising for which sets of input parameters the ML models can be used. If not proven safe, non-ML conservative models are then used for FTD calculation.

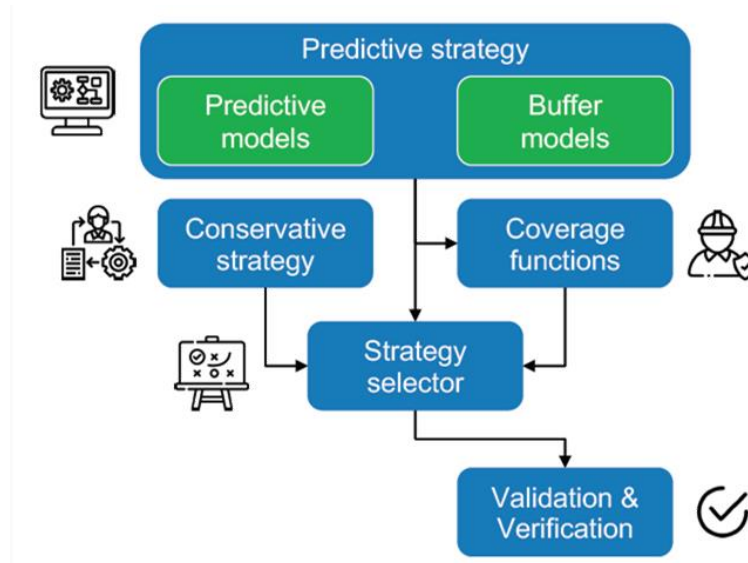


Figure 42 — Schematic view of the assurance framework

At inference stage, using a strategy selector based on the coverage functions, the ML models are then only used when proven to meet design safety criteria.

The coverage functions are established from the empirical error rates (i.e. comparison of time separation statistics with TBS references) per constraint on the data set independent from model train data set with two criteria:

- Empirical error rate below the target error rate (with a small tolerance introduced to avoid too sharp cut-off);
- Confidence interval upper bound of the error rate below the target plus epsilon (tolerance introduced to avoid too sharp cut-off).

The coverage functions are computed for each feature of interest (or combination of features). The error rates are compared to target constraints for several subsets (defined by the value of one or several features). In case the target error rates are respected with enough confidence, the subset is considered covered.

The parameters of the coverage functions have been defined based on expert knowledge determined as impacting the flight behaviour. For the FTD, the coverage functions encompass: surface wind conditions, runway, tuple DBS / ROT / follower RECAT / runway head wind, follower aircraft type, and follower airline.

Figure 43 (respectively Figure 44) shows examples of aircraft types considered as covered (respectively non-covered) based on the comparison of the time separation at FTD minima to the reference TBS minima statistics.

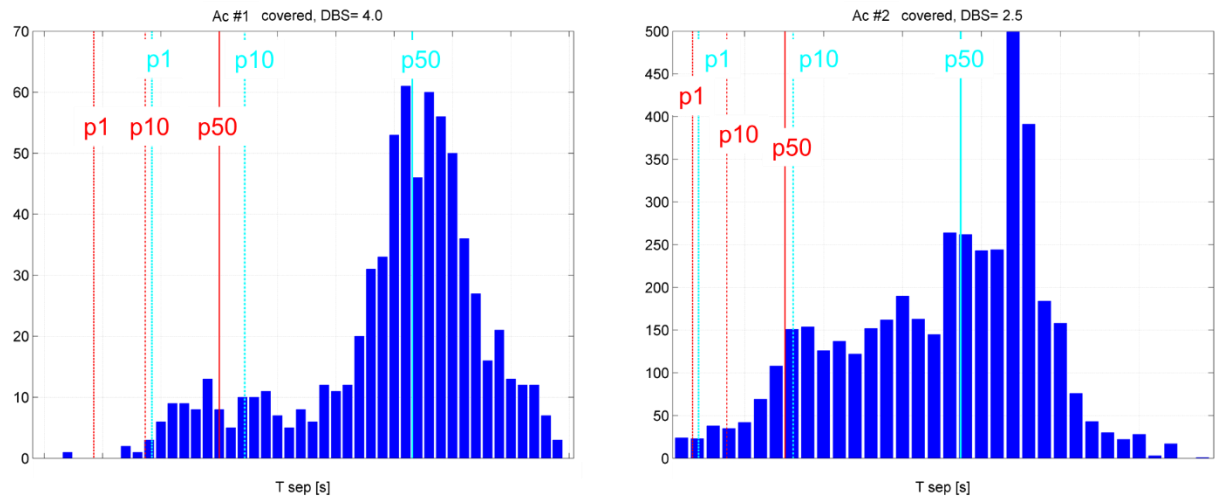


Figure 43 — Example of covered aircraft types: blue bar empirical distribution of time separation at FTD minima; red lines: TBS reference statistics; cyan line: FTD time separation statistics

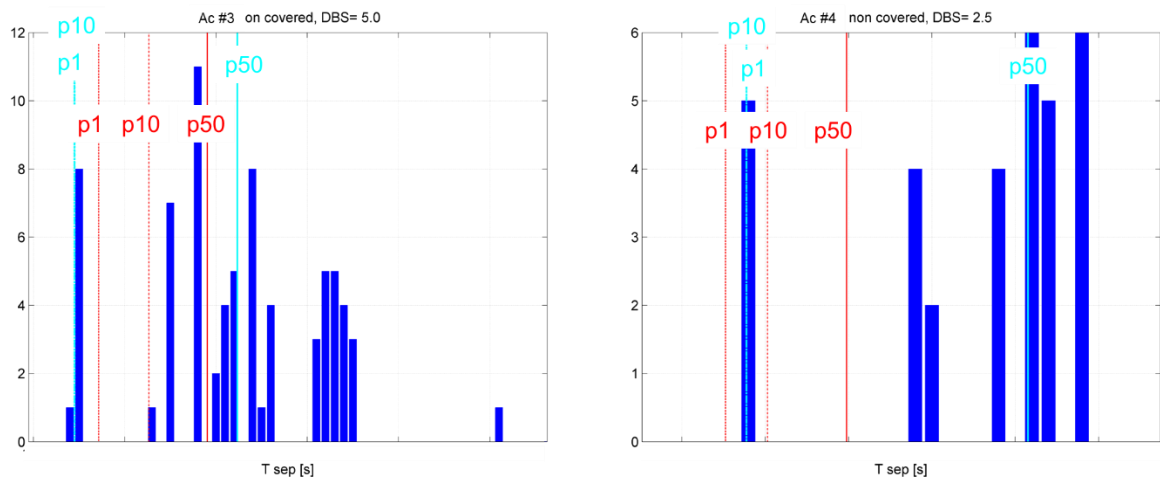


Figure 44: Example of non-covered aircraft types: blue bar empirical distribution of time separation at FTD minima; red lines: TBS reference statistics; cyan line: FTD time separation statistics

Figure 44 provides examples of covered and non-covered airlines based on the distribution of the error on TBS p50, p10 and p1.

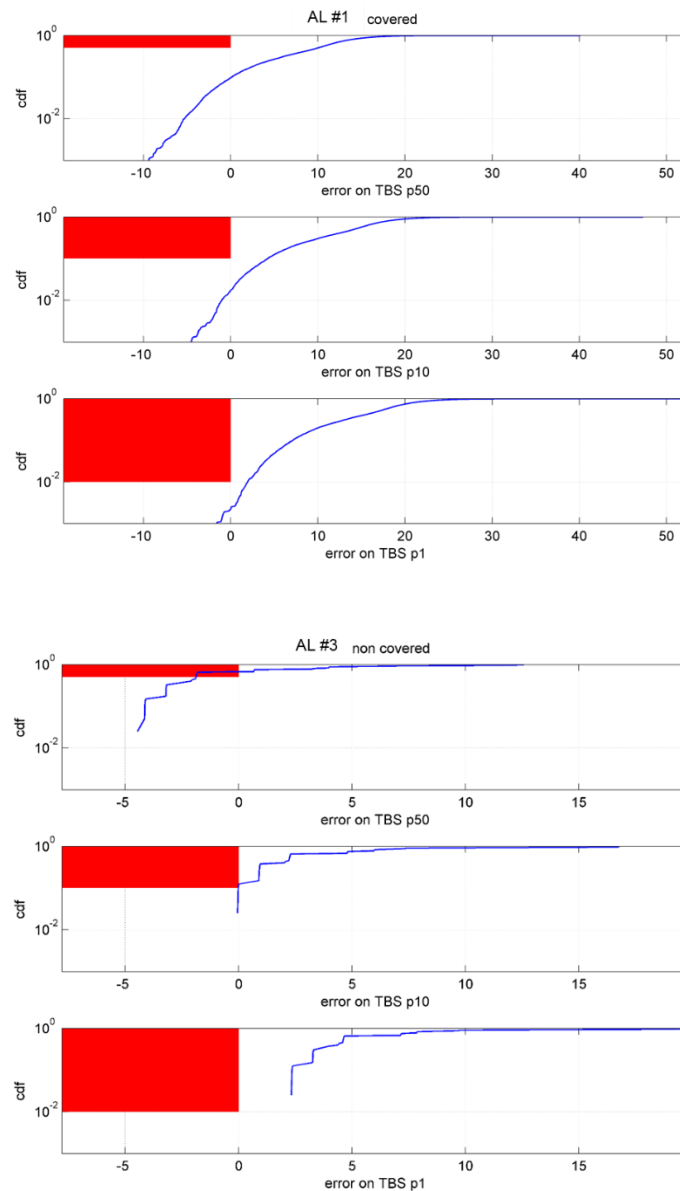


Figure 45 — Example of covered (top) and non-covered (bottom) airline. The graphs show the cumulative density function (cdf) of error on TBS p50 (top subplot), p10 (middle subplot) and p1 (bottom subplot). Red squares indicate the region above targets (50 % for p50, 10 % for p10 and 1 % for p1)

3.2.3. Learning assurance

3.2.3.1. Capture of the AI/ML constituent requirements

Based on the AI/ML constituent architecture presented under Section F.3.2.3.2, it was found valuable to discuss the definition of the ODD, as per the following Objective DA-03.

Objective DA-03: The applicant should define the set of parameters pertaining to the AI/ML constituent ODD, and trace them to the corresponding parameters pertaining to the OD when applicable.

Figure 46 depicts the pipeline for predictive (i.e. ML-based only) FTD training. As the FTD calculation is based on two ML models, the calibration is performed in two steps: time-to-fly model training followed by buffer models' training.

For the time-to-fly (TTF) model, the data set is built from three input data sets:

- follower surveillance radar tracks;
- follower flight data; and
- meteorological data;

An ODD is defined for each input data set. This ODD shall be defined also based on local expert knowledge and historical data analysis. While outliers data is discarded, the remaining data is used for building data sets for TTF training. Once the TTF data set is built, the predictive TTF model is trained.

For the buffer models (four models for TBSp50, TBSp10, TBSp1 and ROT constraints), the buffer data set is built from:

- follower surveillance radar tracks (as used for the TTF model);
- follower flight data (as used for the TTF model);
- meteorological data (as used for the TTF model);
- leader surveillance radar tracks (as used for the TTF model);
- leader flight data;
- separation/spacing constraints; and
- outputs of the TTF predictive model applied to follower data.

This last data set is key as the buffer model aims to determine the uncertainty of the TTF model. The four buffer models are then trained based on the four created buffer data sets.

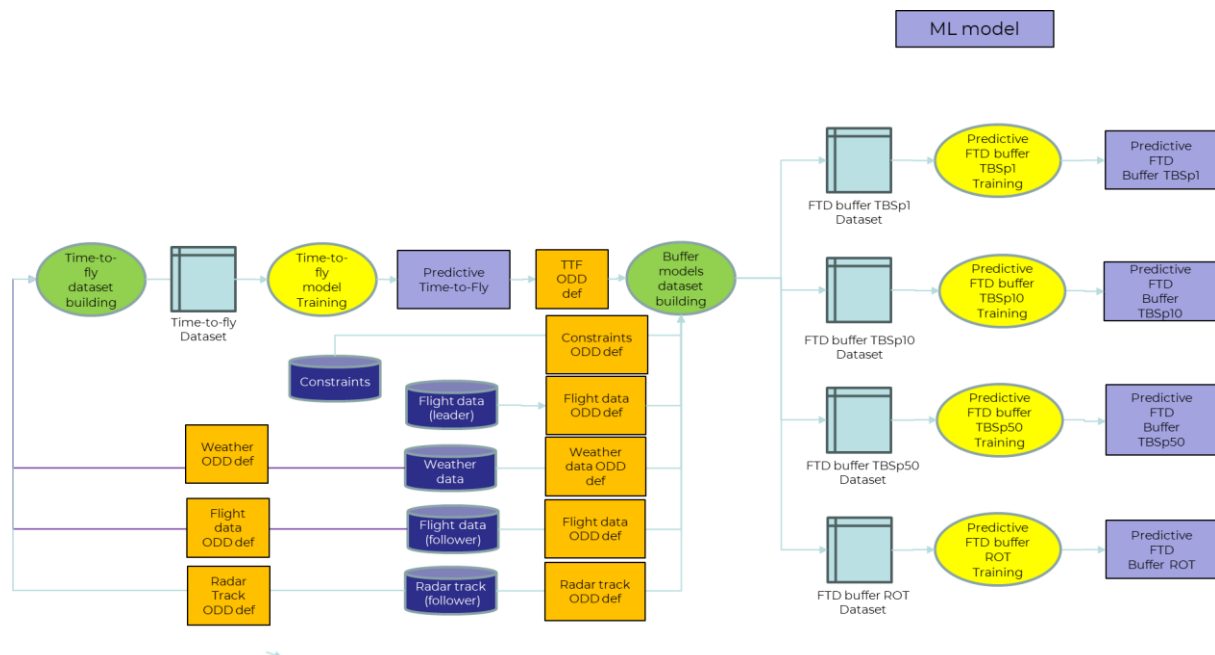


Figure 46 — Pipeline for predictive FTD training

3.2.3.2. Preliminary AI/ML constituent architecture development

At inference stage, as illustrated in Figure 47, the FTD calculation follows several steps:

1. From the MET, follower and constraints inputs, the strategy selector uses the coverage function to determine which type of model to use (predictive or conservative) — ML (predictive) models in the case of Figure 47.
2. A TTF profile is predicted by the ML TTF model for the follower aircraft.
3. This predicted TTF profile is interpolated to calculate:
 - a. a first estimate of the expected FTD for the different applicable constraints (here TBSp50, TBSp10, TBSp1 and ROT);
 - b. an expected average speed on DBS distance used as input for buffer model calculation.
4. Four buffer values are calculated using the input features including the average speed calculated in 3b.
5. The buffer values of 4 are added to the estimators calculated in 3a.
6. For the wake separation, the selected wake FTD value corresponds to the minimum between the DBS and the maximum of the FTDs related to TBSp50, TBSp10 and TBSp1.

7. The final FTD value (to be displayed on CWP) then corresponds to the maximum between the minimum radar separation (MRS), the wake FTD computed in 6 and the FTD related to ROT constraint.

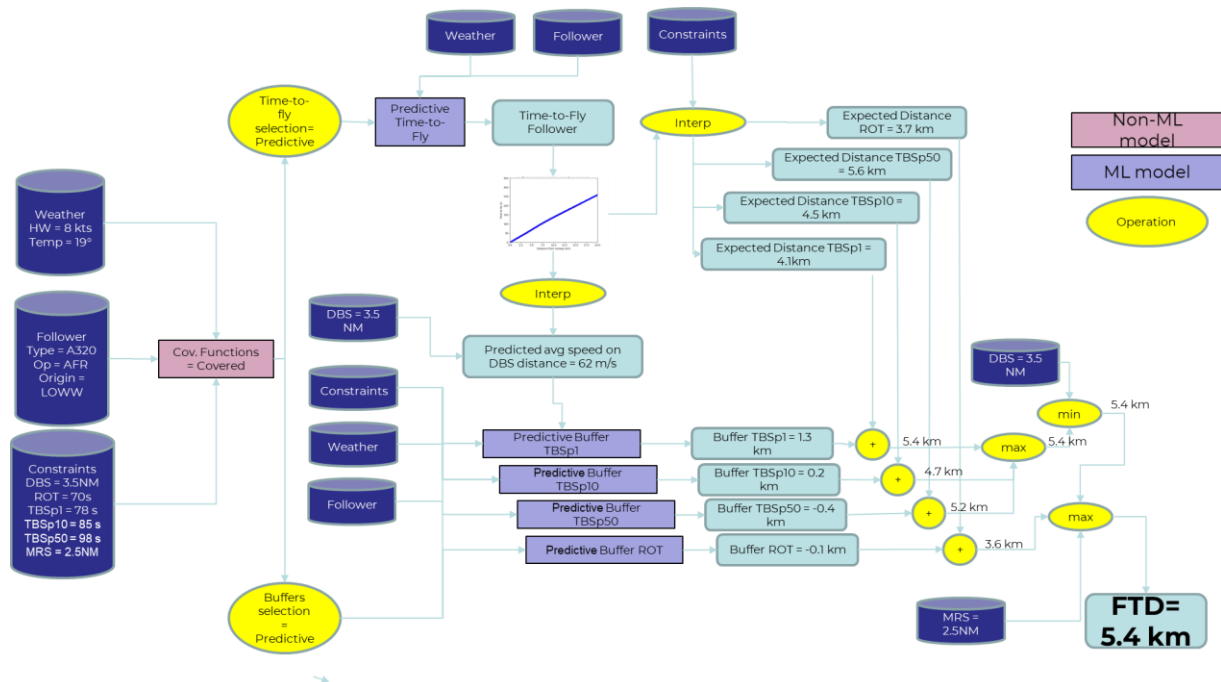


Figure 47 — Example of predictive FTD inference

Objective DA-06: The applicant should describe a preliminary AI/ML constituent architecture, to serve as reference for related safety (support) assessment and learning assurance objectives.

Based on the typical workflow presented above, a candidate FTD computation subsystem architecture is proposed in Figure 48. The figure also presents a candidate architecture for the two AI/ML constituents that are embedded into the AI-based subsystem (i.e. a TTF constituent, and a buffer constituent).

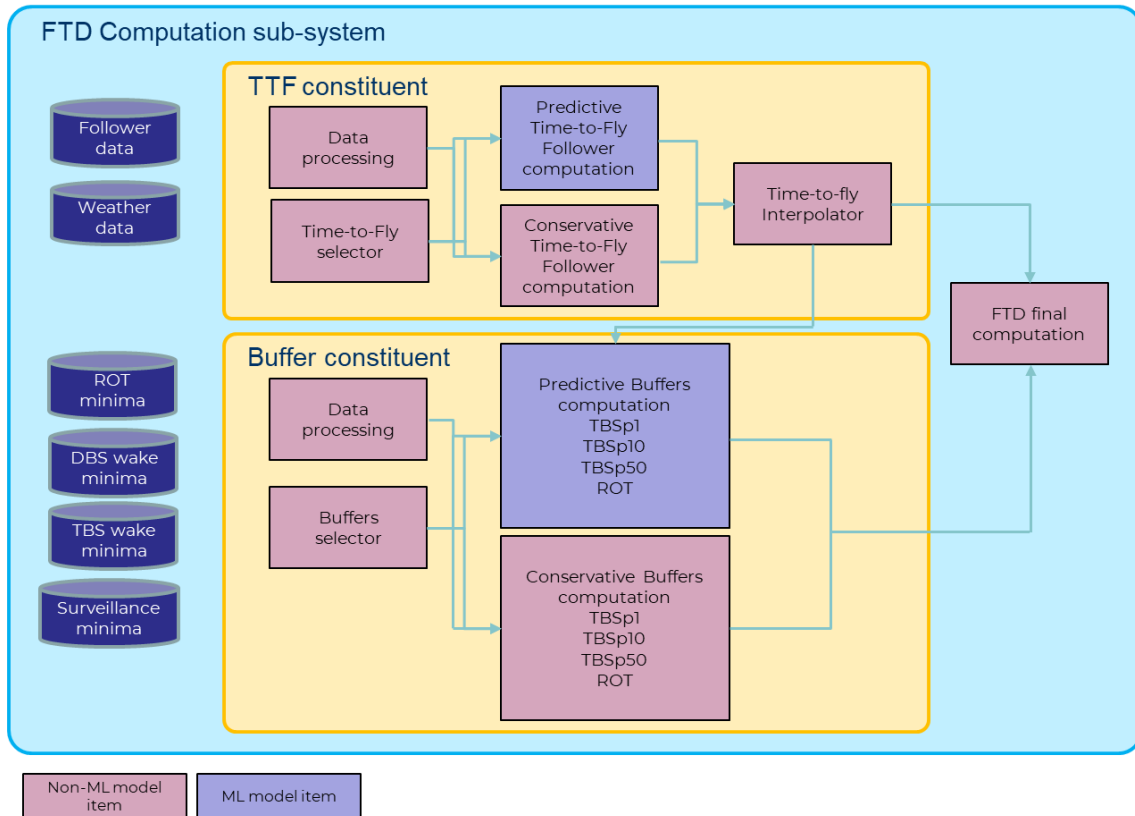


Figure 48 - FTD computation subsystem and its constituents

A similar subsystem and constituent architecture is candidate for the ITD computation subsystem.


It has to be noted that this FTD computation subsystem does not embed any item that will be in charge of the data recording in operations for the different purposes expressed in the document. It is indeed the expectation that these functions will be initially endorsed by the DPO organisation that will be in charge of the integration of the developed library into its ATM/ANS equipment, subject to certification, declaration or statement of compliance (see Note 2 in the introduction to the use case).

4. Use cases — aircraft production and maintenance

Aircraft production and maintenance			
EASA AI Roadmap AI Level (subsystem)	Function allocated to the (sub)systems (adapted HARVIS LOAT terminology)	Controlling corrosion by usage-driven inspections	Damage detection in images
Level 1A Human augmentation	Automation support to information acquisition	Maintenance, environment, operator / manufacturer databases	infrared camera
	Automation support to information analysis	Predicted corrosion level + Time to inspect for corrosion	Damage classification
Level 1B Human assistance	Automation support to decision-making	x	Support decision to repair for inspector validation
Level 2A Human-AI cooperation	Directed decision and automatic action implementation	x	x
Level 2B Human-AI collaboration	Supervised automatic decision and action implementation	x	x

Table 18 — Classification applied to production and maintenance use cases

Where:

 represents the AI-based system or subsystem; and
The AI/ML constituent is in blue.

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).

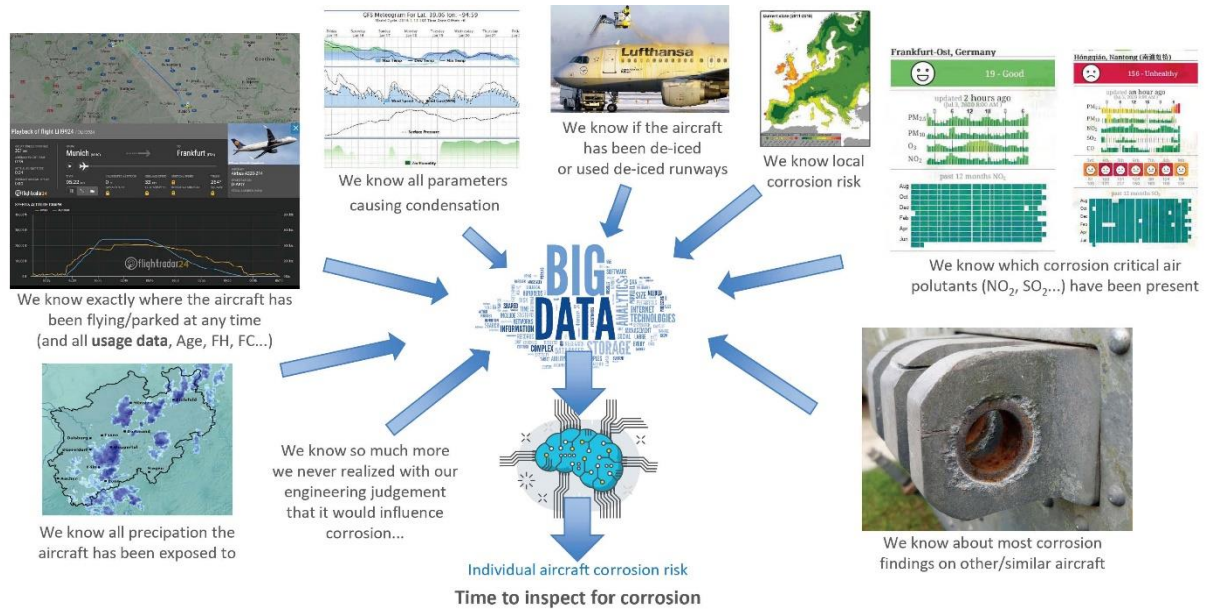


Figure 49 — General philosophy of CPCP by utilisation of data and AI

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.

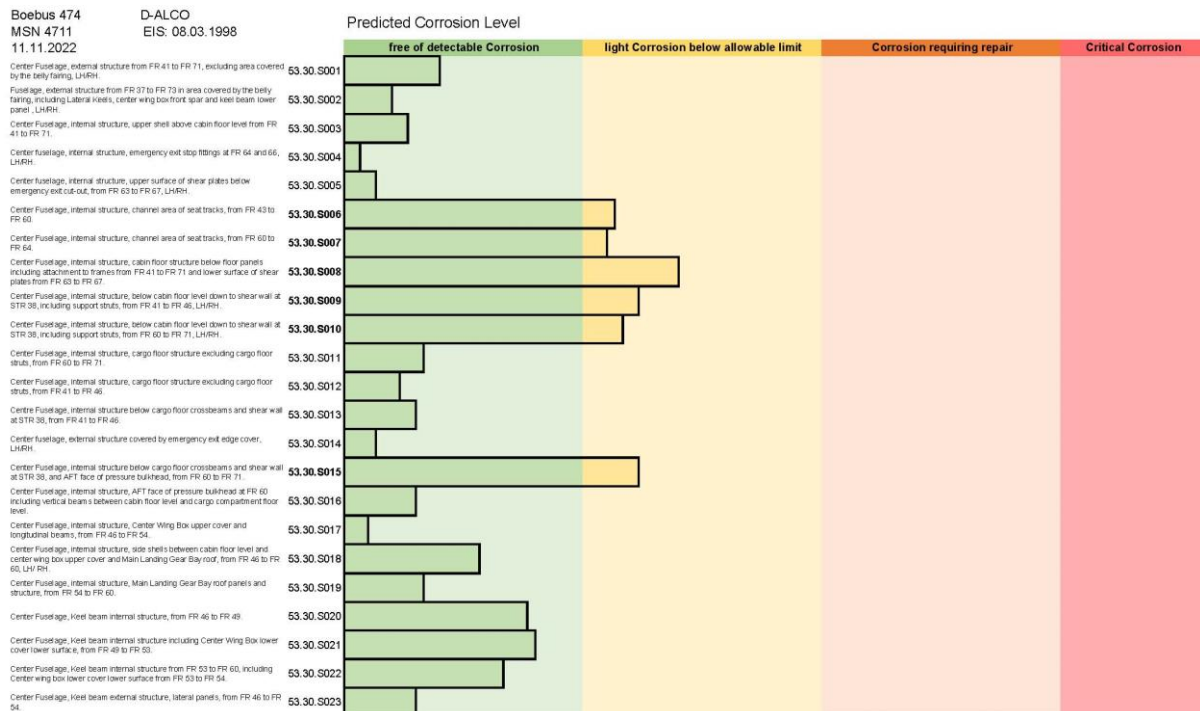


Figure 50 — Example of a possible system output: predicted corrosion in specific areas

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.

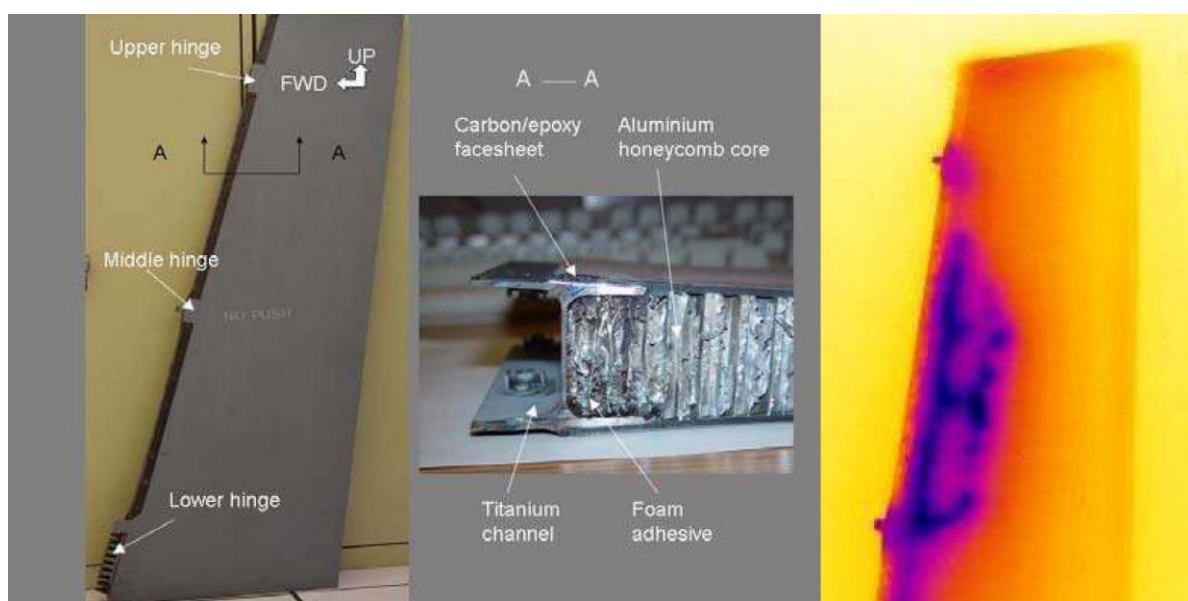


Figure 51 — Thermographic images of a fighter aircraft rudder showing water ingress in honeycomb cells

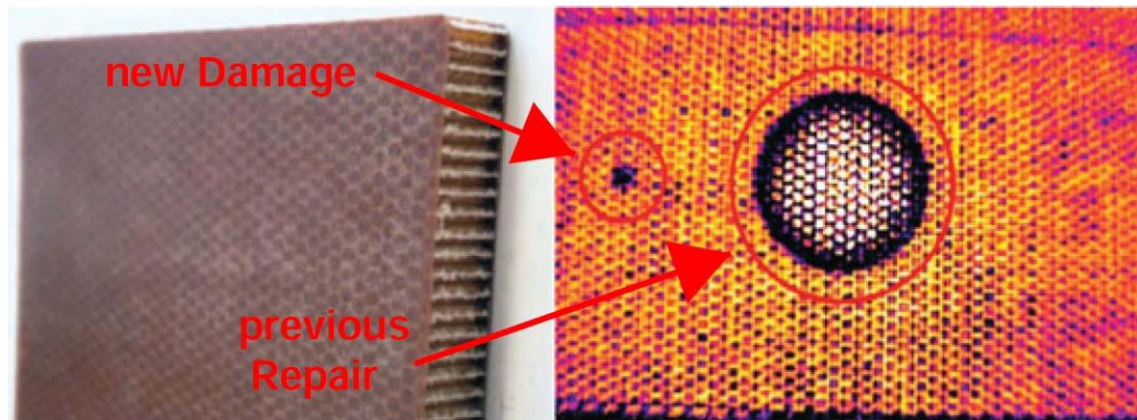


Figure 52 — Optical and thermographic image of a GFRP sandwich panel

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/manufacture databases could be envisaged in the future.



Figure 53 — Portable thermographic test equipment, potentially including an image evaluation system

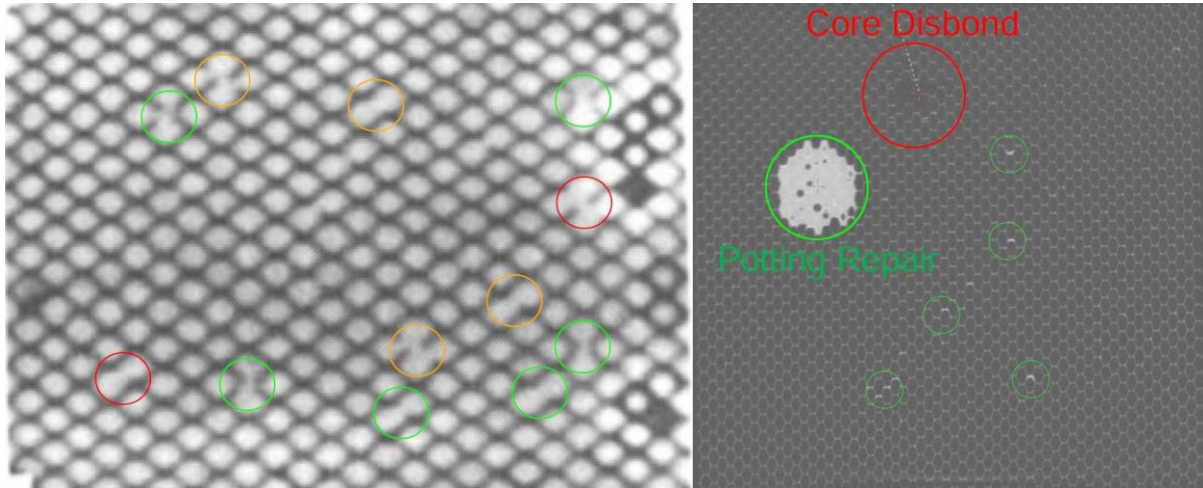


Figure 54 — Example of how the system could mark some areas in images to support inspection of honeycomb sandwich

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 — that 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 related to 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 situation awareness and allowing for a 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. 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 is sufficient to support the system to perform 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)³⁵.

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 fail³⁶.

Accuracy (of the data) — The degree of conformance between the estimated or measured value and its true value.

Adaptivity (of the system) — A system capability to change its active feedback process in order to maintain the desired performance in response to failures, threats or a changing environment³⁷. Note: This is not to be confused with adaptivity of the learning process which is the ability to learn 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 with³⁸.

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 with³⁹.

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 bias). 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:

³⁵ Source: adapted from (EU High-Level Expert Group on AI, 2020).

³⁶ Source: adapted from (EU High-Level Expert Group on AI, 2020).

³⁷ Adapted from (Siddhartha Bhattacharyya and Darren Coffey, 2015)

³⁸ Source: adapted from (EU Commission, 2021).

³⁹ Source: (EU Commission, 2021)

- **Convolutional neural networks (CNNs)** — A specific type of deep neural networks that are particularly suited to process image data, based on convolution operators. (EASA and Daedalean, 2020)
- **Recurrent neural networks (RNNs)** — A type of neural network that involves directed cycles in memory.

Attachment — Is the state of strong emotional bond between the end user and the AI-based system⁴⁰.

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 auditability⁴¹.

Authority — The ability to make decisions 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 oversight⁴².

Advanced automation — The use of a system that, under specified conditions, functions without human intervention⁴³.

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).

Commercial-off-the-shelf machine learning model (COTS ML model) — A hardware and/or software machine learning model product that is ready-made and available for purchase by the general public (reused from NIST COTS software definition).

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.

⁴⁰ Source: adapted from WordReference.com LLC

⁴¹ Source: adapted from (EU High-Level Expert Group on AI, 2020).

⁴² Source: adapted from ISO/IEC 22989:2022(en), 3.1.7.

⁴³ Source: adapted from ISO/IEC 22989:2022(en), 3.1.7.

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 algorithm⁴⁴.

Concept of operations (ConOps) — A ConOps is a human-centric document that describes operational scenarios for a proposed system from the users' operational viewpoint.

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).

Correctness — Different definitions of correctness have to be considered depending on the context:

- **Correctness** (in the data) — is a synonym of accuracy.
- **Correctness** (of a requirement) — is the degree to which an individual requirement is unambiguous, verifiable, unique, consistent with other requirements, and necessary for the set of requirements⁴⁵.

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 been 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 relates to 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 organisation⁴⁶.

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 or ethnic origin, political opinions, religious or philosophical beliefs (ii),

⁴⁴ Source: adapted from (ENISA, December 2021).

⁴⁵ Source: extracted from ED-79B/ARP4754B.

⁴⁶ Source: adapted from (EU High-Level Expert Group on AI, 2020).

processing of personal data relating to criminal convictions and offences (iii) and systematic monitoring of a publicly accessible area on a large scale (iv)⁴⁷.

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 or ethnic origin, political opinions, religious or philosophical beliefs, or refers to criminal convictions and offences. A DPO must be independent of the appointing organisation⁴⁸.

Data set⁴⁹ (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 facilitates 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 consideration⁵⁰. A choice that is made about something after thinking about several possibilities⁵¹.

⁴⁷ Source: adapted from (EU High-Level Expert Group on AI, 2020).

⁴⁸ Source: adapted from (EU High-Level Expert Group on AI, 2020).

⁴⁹ Source: adapted from (ER-022 - EUROCAE, 2021).

⁵⁰ Source: OxfordLanguages.

⁵¹ Source: adapted from the Cambridge Dictionary.

Decision-making — The cognitive process resulting in the selection of a course of action among several possible alternative options⁵². Automated or automatic decision-making is the process of making a decision by automated means without any human involvement⁵³.

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, it produces identical outputs.

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 product⁵⁴. Example: a pilot in an aircraft or an ATCO in an ATC centre are typical end users.

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), and related attacks are often called adversarial attacks. 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 runway⁵⁵.

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.

⁵² Source: adapted from Wikipedia.

⁵³ Source: adapted from ico.org.uk.

⁵⁴ Source: adapted from (EU High-Level Expert Group on AI, 2020).

⁵⁵ Source: adapted from (ENISA, December 2021)

Fairness — Refers to ensuring equal opportunities and non-discriminatory practices applied to individuals or groups of users (or end users). (Definition based on the EU guidelines on non-discrimination⁵⁶)

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.

Feature engineering — Feature engineering is the process of transforming raw data into features that best represent the underlying problem (ISO/IEC TR 24028:2020).

General Data Protection Regulation (GDPR) — EU's data protection law, refer to <https://gdpr.eu> for more details.

Generalisation capability — Ability of a learning algorithm to produce models that perform well on unseen data encountered during the operational phase⁵⁷. In this context, 'unseen data' is from the same distribution as the training data set.

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 bias, are the result of the training process⁵⁸.

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'⁵⁹ in statistics).

⁵⁶ [Article 21 'Non-discrimination' | European Union Agency for Fundamental Rights \(europa.eu\)](#)

⁵⁷ Source: adapted from (EASA and Daedalean, 2020).

⁵⁸ Source: adapted from (Goodfellow-et-al, 2016).

⁵⁹ 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.

Infeasible corner case (see also corner case) — Corner case that is not part of the functional intent, thus outside the ODD.

Inference — The process of feeding the machine learning model an input and computing its output. See also the 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.

Input space — Given a set of training examples of the form $\{(x_1, y_1) \dots (x_N, y_N)\}$ 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.

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.

Level of abstraction — In the context of this document, the level of abstraction corresponds to the degree of detail provided within an explanation.

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 during the operations, where the ML model parameters are updated continuously based on data acquired during operation (also known as continual or adaptive learning).

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).

Nominal (data) — Data points that are inside the ODD and are not inliers, singular points, or data points corresponding to edge cases or corner cases.

Novelty — Data which is within the ODD according to the existing ODD parameters, but which should have been considered outside the ODD if it had been correctly described with the introduction of at least one new ODD parameter. A novelty is in general due to a lack of characterisation of the ODD. It could be integrated to the ODD after analysis following the upgrade policy of the ODD. A novelty that is already outside the ODD is therefore an outlier.

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. For instance, in the airworthiness domain, the Certification Specification for large transport aircraft, CS 25.1309 requires the identification of ‘the aeroplane operating and environmental conditions’. A definition of ‘foreseeable conditions’ can be found under AMC 25-11 and generalised: ‘Foreseeable Conditions - The full environment that the [...] system is assumed to operate within, given its intended function. This includes operating in normal, non-normal, and emergency conditions.’

Operational design domain (ODD) — Operating conditions under which a given AI/ML constituent is specifically designed to function as intended, including but not limited to environmental, geographical, and/or time-of-day restrictions⁶⁰. 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

⁶⁰ Source: adapted from SAE J3016, Level of driving automation, 2021.

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 data⁶¹.

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.

Over-reliance — is the state when the end user is excessively relying on, depending on or trusting in the AI-based system⁶².

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 way⁶³.

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 system⁶⁴.

Reliability — The probability that an item will perform a required function under specified conditions, without failure, for a specified period of time⁶⁵.

⁶¹ Source: adapted from (ENISA, December 2021)

⁶² Source: adapted from Merriam-Webster Inc.

⁶³ Source: adapted from (ENISA, December 2021).

⁶⁴ Source: adapted from (EU High-Level Expert Group on AI, 2020).

⁶⁵ Source: ARP 4761 Guidelines and Methods for Conducting the Safety Assessment Process on Civil Airborne Systems and Equipment, 1996.

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 oversight⁶⁶.

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 taken⁶⁷. 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 safety risk mitigation 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 — The ability of a system to maintain its level of performance under all foreseeable conditions. At AI/ML constituent level, the robustness objectives are further split into two groups: the ones pertaining to ‘generalisation’ and the ones pertaining to ‘robustness in adverse conditions’. In this context, adverse conditions refer to the singular points, edge and corner cases, out-of-distribution cases and adversarial cases.

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).

Shared situation awareness — Human-AI shared situation awareness refers to the collective understanding and perception of a situation, achieved through the integration of human and AI-based system capabilities. It involves the ability of both humans and AI systems to gather, process, exchange

⁶⁶ Source: [Cambridge Dictionary](#).

⁶⁷ Source: IEV ref 903-01-11 — https://std.iec.ch/iev/iev.nsf/ID_xref/en:903-01-11.

and interpret information relevant to a particular context or environment, leading to a shared comprehension of the situation at hand. This shared representation enables effective collaboration and decision-making between humans and AI-based systems.

Singular point — A point at which a given mathematical object is not defined, or a point where the mathematical object ceases to be well-behaved; for instance, lacking differentiability or analyticity⁶⁸.

Situation awareness — Situation awareness is the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and a projection of their status in the near future⁶⁹.

Situation representation — Situation representation is the collection of the environment and system state as well as of the state of the end user, processing of this information, with the aim of enabling extrapolation of a target status in the near future.

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.:

$$\|x' - x\| < \delta \Rightarrow \|\hat{f}(x') - \hat{f}(x)\| < \varepsilon, \text{ where } x, x' \in X \text{ and } \delta, \varepsilon \in R_{>0}.$$

Subject — A subject is a person, or a group of persons affected by the AI-based system⁷⁰.

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 defined combination of subsystems, equipment or items that perform one or more specific functions [ED-79B/ARP4754B]

Traceability (of data) — The ability to track the journey of a data input through all stages of sampling, labelling, processing and decision-making⁷¹.

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

⁶⁸ [https://en.wikipedia.org/wiki/Singularity_\(mathematics\)](https://en.wikipedia.org/wiki/Singularity_(mathematics))

⁶⁹ Source: Endsley, M.R.: Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors Journal 1995, 37(1), 32-64.

⁷⁰ Source: adapted from (EU High-Level Expert Group on AI, 2020).

⁷¹ Source: adapted from (EU High-Level Expert Group on AI, 2020).

maps the input to the output, by optimising its internal parameters. See also the related definition of **Inference**.

Transfer learning — The process where an ML model trained for a task is reused and adapted for another task.

Unintended behaviour — Unexpected operations of a system in ways contrary to intended functionality.

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 technicians⁷².

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).

⁷² 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	design or production organisation
DQR	data quality requirement
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 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
IP	intellectual property
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	situation awareness
SAL	security assurance level
SLT	statistical learning theory
SMS	safety management system
SPO	single-pilot operation
SAE	Society of Automotive Engineering
STRIP	STRong Intentional Perturbation
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

AVSI. 2020. *AFE-87 – Machine Learning*. 2020.

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.

EASA and 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.

EASA. 2023. *Machine Learning Application Approval - Unified deliverable Phase 2*. 2023.

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 - Office of Next Gen. 2022. *Certification Research Plan for AI Applications.* 2022.

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 Diulio, 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 University 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.

SESAR JU. 2018. *SESAR Human Performance Assessment Process V1 to V3- including VLDs.* 2018.

Siddhartha Bhattacharyya and Darren Coffey. 2015. *NASA/CR-215-218702 - Certification Considerations for Adaptive Systems.* s.l. : NASA, 2015.

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-08. 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-10 to EXP-16, Provision ORG-08. 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-10 to EXP-16, Provision ORG-08. 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-08. 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-06, Provision ORG-08. Rationale: Over-reliance 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-08) 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-10 to EXP-16. 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 (<i>a process of reciprocal stimulation or response between two people</i>) 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 should 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-08. 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, SA-02, SA-03, IS-01, EXP-05, EXP-06, EXP-18, EXP-19, DA-01, DA-02, DA-06, DA-07. Rationale: The question is answered through the safety (SA-01), continuous safety (SA-02 and SA-03) and security assessment (IS-01) and

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
	monitoring for the adherence of operational boundaries (EXP-05, EXP-06, EXP-18, EXP-19), which result is finally fed back into the learning assurance process requirements (DA-01, DA-02, DA-06, DA-07).
G1.j. Did you ensure a <i>'stop button' or procedure to safely abort override an operation by a human end-user when needed?</i>	<p>Objective: SA-01 to SA-03, IS-01, EXP-12, DA-01, DA-02, DA-06, DA-07.</p> <p>Rationale: The override-procedure should be assessed for compliance with safety objectives (SA-01, SA-02, SA-03) and security objective (IS-01), safeguarded by the relevant explainability (EXP-12) and specified through the learning assurance process requirements (DA-01, DA-02, DA-06, DA-07).</p> <p>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.</p>
G1.k. Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI-based system?	<p>Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale).</p> <p>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.</p>

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 <i>for information security requirements</i> (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: <ul style="list-style-type: none"> ▪ 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: IS-03, applicable IS guidance Rationale: Refutation test (including red team/pentest) is addressed in the applicable information security guidance (e.g. AMC 20-42). This is part of the validation and verification strategy under Objective IS-03.
G2.f. Did you inform end users of the duration of security coverage and updates?	Objective: AMC 20-42 or Part-IS Rationale: The organisation (design or operation) should monitor the evolution of security risks as defined in e.g. AMC 20-42 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: SA-02 to SA-03. Rationale: A process for continuous risk monitoring, using defined metrics and levels (SA-02 to SA-03) 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-08. 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.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
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.
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-06 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, SA-03, EXP-06 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-06) and assessed (SA-03)
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	Objective: DM-02-SL/UL to DM-03 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.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
representative of the environment the system will be deployed in?	up to date, complete, representative and verified to be compliant with the DQRs.
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-08, DA-10, EXP-06, SA-03. Rationale: During the development phase (LM-09, IMP-08) 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-10). During the operational phase (EXP-06, SA-03), 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: SA-03 Rationale: The continuous safety assessment (SA-03) 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-06 Rationale: Relevant information concerning deviations of the AI-based system's output from the specified performance need to be indicated (EXP-06) 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-10, LM-07-SL 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 behaviour (DA-10) and the reproducibility of the learning process (LM-07-SL).
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-10 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-SL Rationale: The bias-variance trade-off should be accounted for in the model family selection in

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
	order to provide evidence of the reproducibility of the training process.
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-07, DA-10 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-07) and verified (DA-10). 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-06, EXP-18 Rationale: The AI-based system output performance should be monitored (EXP-06) and procedures put in place to act on the possible output of the AI-based system's monitoring (EXP-18).
G2.u. Is your AI-based system using (online) continual learning?	Objective: LM-02 and associated anticipated MOC. 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, data protection 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-03 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-03 Rationale: See item G3.a.

Data protection 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-03 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 <i>AI-based</i> 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 <i>AI-based</i> system?	
G3.f. Did you consider the privacy and data protection implications of data collected, generated or processed over the course of the <i>AI-based</i> system's life cycle?	Objective: IS-01 to IS-02 Rationale: Non-personal data, which is processed by an <i>AI-based</i> system should be protected for data security by assessing the security risks and installing security controls.
G3.g. Did you consider the privacy and data protection implications of the <i>AI-based</i> system's non-personal training data or other processed non-personal data?	

Data governance in aviation applications

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
G3.h. Did you align the <i>AI-based</i> system with relevant standards (e.g. ISO, IEEE) or widely adopted protocols for (daily) data management and governance?	Objective: to be developed in the future Issue 03 of this document. Rationale: to be developed in the future Issue 03 of this document.

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: SA-02, SA-03, EXP 09, EXP-05. Rationale: A process for data recording (EXP-04, SA-02) and continuous safety assessment (SA-03) 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-05) 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: SA-02, SA-03, EXP 09. 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: SA-02, SA-03, EXP 04. Rationale: A process for data recording (EXP-04, SA-02) and continuous safety assessment (SA-03) 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: SA-02, SA-03, EXP 06 and EXP-09. Rationale: A process for data recording (EXP-04, SA-02) and continuous safety assessment (SA-03) 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-06) 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: SA-02, EXP 09. Rationale: A process for data recording (EXP-09, SA-02) 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-10 to EXP-16 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 assess the end users if they understand the decision(s) of the AI-based system?	Objective: SA-03 Rationale: The understanding of decisions of the AI-based system might contribute to the end user's ability to monitor in-service events for the detection of potential issues or suboptimal performance trends (SA-03). The verb 'survey' in the ALTAI item was changed to 'assess', because in the aviation domain, there is no way to survey but to assess the understanding of the end user.

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-07 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.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
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 level of accuracy and/ or error rates?	Objective: See item G4.j. Rationale: See item G4.j.
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 bias 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) bias 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, EASA encourages applicants to check whether the AI-based system could have any impact on diversity, non-discrimination and fairness, which at the same time is of safety relevance. 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 (safety relevance), 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 <i>algorithm</i> design?	<p>Objective: ET-04, SA-03, MOC DM-05-2; MOC DM-05-3; EXP-02; LM-07-SL and LM-08.</p> <p>Rationale: The avoidance of potential unfair bias (with safety relevance) is addressed through the systematic mitigation of any potential bias in all phases of the AI-based system development and operations. All objectives mentioned above contribute to this goal:</p> <ul style="list-style-type: none"> • Learning assurance aims at detecting potential bias in the data, through data representativeness (MOC DM 05-2) and data accuracy and correctness (MOC DM 05-3). • Objectives LM-07-SL and LM-08 contribute to ensuring that bias 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. <p>The Continuous Safety Assessment (SA-03) aims at identifying bias or poor performance in the systems operation.</p>
G5.b. Did you consider diversity and representativeness of end users and/or subjects in the data?	<p>Objective: Anticipated MOC DA-07</p> <p>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).</p>
i. Did you test for specific target groups or problematic use cases?	<p>Objective: Objective LM-14 and IMP-09</p> <p>Rationale: robustness on adverse cases covering specific target groups and problematic use cases</p>
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?	<p>Objective: Not applicable.</p> <p>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</p>

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
	and are managed through the learning assurance process.
<p>iii. Did you assess and put in place processes to test and monitor for potential bias during the entire life cycle of the <i>AI-based</i> system (e.g. bias due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness)?</p>	<p>Objective: SA-03, MOC DM-05-2; MOC DM-05-3; EXP-02; LM-07-SL and LM-08, ICSA-02 Rationale: The avoidance of potential unfair bias with safety relevance is addressed through the systematic mitigation of any potential bias in all phases of the AI-based system development and operations is addressed through the systematic mitigation of any potential bias in all phases of the AI-based system development and operations. All objectives mentioned above contribute to this goal:</p> <ul style="list-style-type: none"> • Learning assurance aims at detecting potential bias in the data, through data representativeness (MOC DM 05-2) and data accuracy and correctness (MOC DM 05-3). • Objectives LM-07-SL and LM-08 contribute to ensuring that bias 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. <p>The Continuous Safety Assessment (SA-03) aims at identifying bias or poor performance in the systems operation. Learning assurance in particular related to ensure that data representativeness (MOC DM 05-2) and data accuracy and correctness (MOC DM 05-3) are not bias impacted. Also MOC EXP-02 ensuring the goals and in particular LM-07-SL 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.</p>
<p>iv. Where relevant, did you consider diversity and representativeness of</p>	<p>Objective: See item G5.b. Rationale: See item G5.b.</p>

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
end users and/or subjects in the data?	
G5.c. Did you put in place educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI-based system?	<p>Objective: Provision ORG-07.</p> <p>Rationale: The organisation should put in place training initiatives that would support the development of bias awareness and other AI-specific competencies for the users.</p>
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?	<p>Objective: MOC DM-05-2; MOC DM-05-3; EXP-02 and EXP-06; LM-07-SL and LM-08, SA-03.</p> <p>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.</p> <p>The mitigation of potential for discrimination is addressed through the monitoring of system performance (EXP-06) 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:</p> <ul style="list-style-type: none"> • Learning assurance aims at detecting potential biases in the data, through data representativeness (MOC DM 05-2) and data accuracy and correctness (MOC DM 05-3). • Objectives LM-07-SL 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. <p>The Continuous Safety Assessment (SA-03) aims at identifying bias or poor performance in the systems operation.</p>
i. Did you establish clear steps and ways of communicating on how and to whom such issues can be raised?	<p>Objective: ORG-03.</p> <p>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.</p>

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
<p>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?</p>	<p>Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale).</p> <p>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 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.</p>
<p>G5.e. Is your definition of fairness commonly used and implemented in any phase of the process of setting up the AI-based system?</p>	<p>Objective: MOC DM-05-2; MOC DM-05-3; EXP-02; LM-07-SL and LM-08, SA-03.</p> <p>Rationale: The applicable definition of fairness is defined in the glossary of the present document. Regarding the mitigation of potential unfairness, as far as safety relevant, 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:</p> <ul style="list-style-type: none"> • Learning assurance aims at detecting potential biases in the data, through data representativeness (MOC DM 05-2) and data accuracy and correctness (MOC DM 05-3). • Objectives LM-07-SL 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. <p>The Continuous Safety Assessment (SA-03) aims at identifying bias or poor performance in the systems operation.</p>

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
i. Did you consider other definitions of fairness before choosing this one?	<p>Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale).</p> <p>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.</p>
ii. Did you consult with the impacted communities about the correct definition of fairness, i.e. representatives of elderly persons or persons with disabilities?	<p>Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale).</p> <p>Rationale: Having the definition based on the EU non-discrimination guidelines, all the impacted communities are understood as considered.</p>
iii. Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness?	<p>Objective: Not addressed through the objectives of this Concept Paper (please consider the rationale).</p> <p>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.</p>
iv. Did you establish mechanisms to ensure fairness in your AI-based system?	<p>Objective: MOC DM-05-2; MOC DM-05-3; EXP-02; LM-07-SL and LM-08, SA-03</p> <p>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:</p> <ul style="list-style-type: none"> • Learning assurance aims at detecting potential biases in the data, through data representativeness (MOC DM 05-2) and data accuracy and correctness (MOC DM 05-3). • Objectives LM-07-SL 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.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
	The Continuous Safety Assessment (SA-03) 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-06. Rationale: Through Objective CO-06, 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.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
i. Did you assess whether the team involved in building the <i>AI-based</i> system engaged with the possible target end users <i>and/or subjects</i> ?	Objective: CO-06. 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-06.
ii. Did you assess whether there could be groups who might be disproportionately affected by the outcomes of the <i>AI-based</i> system?	Objective: SA-01, SA-03. 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 (SA-03) 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' <i>and/or subjects'</i> communities?	Objective: MOC DM-05-2; MOC DM-05-3; EXP-02; LM-07-SL and LM-08, SA-03. Rationale: Regarding the mitigation of potential unfairness, as far as safety relevant, it is addressed through the systematic mitigation of any potential biases in all phases of the <i>AI-based</i> system development and operations. The objectives mentioned above all contribute to this goal: <ul style="list-style-type: none"> • Learning assurance aims at detecting potential biases in the data, through data representativeness (MOC DM 05-2) and data accuracy and correctness (MOC DM 05-3). • Objectives LM-07-SL 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 (SA-03) 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?	<p>Objective: ORG-01</p> <p>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.</p>

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)?.	<p>Objective: ET-06.</p> <p>Rationale: This ALTAI question has been reworked:</p> <ul style="list-style-type: none"> — 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 MOC ET-06.
G6.c. Did you define measures to reduce or mitigate these impacts?	Objective: ET-06. 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-07 and Objective ET-08**, 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.*

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 with impacted workers and their representatives (trade unions, (European) work councils) in advance?	Objective: to be developed in the future Issue 03 of this document. Rationale: to be developed in the future Issue 03 of this document.

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
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-08. 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-08. Rationale: This risk should be mitigated through refresher training.
G6.h. Does the system promote or require new (digital) skills?	Objective: ET-07 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-07.
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.
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

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
	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.’

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, DA-04, CM-01, 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 (DA-04). 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-06. Rationale: The AI-based system should be auditable by internal and external entities.

Risk management

Quote from the ALTAI: 'Both the ability to report on actions or decisions that contribute to the AI system's outcome, and to respond to the consequences of such an outcome, must be ensured.'

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-07, ORG-05 Rationale: The organisation's training procedures and 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.(ET-07, ORG-05)
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-04 Rationale: The organisation should establish means (e.g. processes) to continuously assess ethics-based aspects as far as safety is concerned and in case of a safety-relevant conflict between different ethical principles,

ALTAI items	Link to EASA concept paper objectives Objective(s)/Rationale for the link
	warrant an explanation on the decision-making.
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-04 Rationale: See item G7.e
i. Does this process include identification and documentation of conflicts between the six aforementioned <i>gears</i> 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-07, ORG-08 Rationale: The organisation's training procedures and requirements (ORG-08, ET-07) 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), safety-relevant 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, SA-03. Rationale: Continuously assess the safety in operation (SA-03) and mitigate the risk by architectural design, e.g. safety net (SRM-02).

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