European Union Aviation Safety Agency

Notice of Proposed Amendment 2025-07 (B)

in accordance with Article 6 of Management Board Decision 01-2022

NPA 2025-07 (B) — Proposed detailed specifications and associated acceptable means of compliance and guidance material for AI trustworthiness (DS.AI)

Contents

Detailed specifications and associated acceptable means of compliance and guidance material for artificial intelligence (AI) trustworthiness (DS.AI)

DS.AI.010 Scope

(a) The purpose of DS.AI is to establish a comprehensive framework of technical requirements for ensuring the trustworthiness of AI-based systems.

(b) DS.AI applies to different aviation domains through its technical requirements and related acceptable means of compliance (AMC) and guidance material (GM) or directly through its AMC and GM.

(c) DS.AI applies to AI-based systems that:

(1) are determined by the risk assessment under DS.AI.130 as being high-risk AI systems in the sense of Regulation (EU) 2024/1689, and

(2) are classified as level 1 (human augmentation or assistance) or level 2 (human–AI cooperation or collaboration) AI-based systems under DS.AI.110.

(d) The following cases are excluded from the scope of DS.AI:

(1) AI-based systems presenting a risk directly contributing to the potential for fatalities or multiple life-threatening injuries, usually with the loss of an aircraft or massive uncontained environmental effects as applicable,

(2) AI-based systems with online learning capabilities, for which the failure contribution is more stringent than ‘no safety effect’,

(3) AI-based systems that incorporate logic- and knowledge-based AI or hybrid AI, for which failure contribution is more stringent than ‘no safety effect’, or

(4) AI-based verification tools involved in the verification of AI-generated artefacts, unless all outputs of the AI-based verification tool are independently verified by a human user.

DS.AI.020 Definitions

Addictive behaviour. Actions, often obsessive and destructive, related to one’s abuse of or dependence on a stimulus (provided by the AI-based system), that dominate one’s life. Addictive behaviours include risk taking and breaking laws in the course of sustaining one’s addiction.

AI (artificial intelligence). Technology that, for explicit or implicit objectives, infers from the inputs received how to generate outputs — such as predictions, content, recommendations or decisions — that can influence physical or virtual environments.

AI-based system. A machine-based system that is designed to operate with varying levels of automation (up to autonomy) and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers from the inputs it receives how to generate outputs — such as predictions, content, recommendations or decisions — that can influence physical or virtual environments.

AI constituent. A defined and bounded collection of hardware and/or software items, which are grouped for integration purposes to support one or more AI-based system function.

Note: The notation 'AI/xxx constituent' is used to reflect the specificities of given AI technologies, such as machine learning (xxx = ML) or logic- and knowledge-based (xxx = LKB) technologies.

AI model. An abstract representation of a given set of aspects of a system, physical phenomenon or process, created using AI technology.

Note: The notation 'AI/xxx model' is used to reflect the specificities of given AI technologies, such as machine learning (xxx = ML) or logic- and knowledge-based (xxx = LKB) technologies.

AI/ML model. A model characterised by parameters determined by a data-driven machine-learning process to satisfy one or more requirements.

Attachment. Behaviour associated with the formation of and investment in a significant bond; seeking an emotionally supportive social relationship with the AI-based system.

Authority. The ability to make decisions without the need for approval from another member involved in the operations. The different levels of authority discussed in this document are as follows.

— **Full authority.** The end user has complete control and oversight over the AI-based system. If the AI-based system is capable of taking certain decisions, the end user oversees and can intervene in all decisions being made.

— **Partial authority.** The end user has some degree of control over the AI-based system; the AI-based system can take certain decisions in relative independence, but the end user still oversees the consequences of those decisions.

— **Limited authority.** The end user can recover, upon alerting, control of the AI-based system's decision-making to ensure the safety of operations.

Automation. The use of machine-based systems, reducing the need for human intervention when performing tasks.

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

Autonomy. Characteristic of a system that is capable of modifying its intended domain of use or goal without external intervention, control or oversight.

Concept of operations (ConOps). A human-centric description of operational scenarios for a proposed system from the end users' operational viewpoint.

Corner case (see also **edge case**). A situation that, considering at least two parameters of the AI constituent operational design domain, occurs rarely when these parameters are combined (i.e. low representation of the associated values in the distribution for the combination of those parameters).

Data dimensionality. The number of dimensions of the input space, i.e. the number of independent features that define a data point.

— **Low-dimensional data.** Data in which the number of features is tractable.

— **High-dimensional data.** Data in which the number of features is intractable.

Decision-making. The process resulting in the selection of a course of action from several possible alternative options**.**

Deep learning. A type of machine learning that uses large neural networks to learn abstract representations of the input data.

Development assurance. All planned and systematic actions used to substantiate, to an adequate level of confidence, that development errors have been identified and corrected so that the system satisfies the applicable requirements.

Edge case (see also **corner case**). A situation that, considering a given parameter of the AI constituent operational design domain (ODD), occurs rarely (i.e. low representation of the associated value in the distribution for that parameter).

End user. The person who ultimately uses or is intended to ultimately use the AI-based system. This could either be a consumer or a professional within a public or private organisation. The end user stands in contrast to users who support or maintain the product. Typical end users include a pilot in an aircraft or an air traffic controller in an air traffic control centre.

Explainability (of an AI application). The capability to provide a human with understandable, reliable and relevant information with the appropriate level of detail and appropriate timing about how an AI application produces its results.

Generalisation capability. The ability of a model to perform well on unseen data encountered during the operational phase.

Hybrid (hybrid AI). The branch of AI mixing several techniques, like machine learning and logic- and knowledge-based approaches.

Learning assurance. All planned and systematic actions used to substantiate, at an adequate level of confidence, that development errors from data-driven learning processes have been identified and corrected, such that the AI/ML constituent satisfies the applicable requirements at a specified level of performance and possesses sufficient generalisation and robustness capabilities.

Logic- and knowledge-based (LKB). The branch of AI concerned with approaches where AI-based systems infer from encoded knowledge or symbolic representation of the task to be solved.

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 inferring from this data. ML includes three techniques.

— **Supervised learning.** The process of learning in which the learning algorithm processes the input dataset, and a cost function measures the difference between the ML model’s 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 an unlabelled dataset, and a cost function indicates whether the ML model has converged on 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 techniques can be further characterised as:

— **offline learning**, where the ML model is frozen at the end of the development phase; or

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

Manipulating behaviour. Behaviour intended to exploit, control or otherwise influence others to the AI-based system’s advantage.

Mental health. A state of mind characterised by emotional well-being, good behavioural adjustment, relative freedom from anxiety and disabling symptoms and a capacity to establish constructive relationships and cope with the ordinary demands and stresses of life.

ML model. A parameterised function that maps inputs to outputs, whose parameters are determined during the training process.

— **Trained model.** The ML model obtained at the end of the learning/training phase

— **inference model.** The ML model obtained after the transformation of the trained model, so that it is adapted to the inference platform.

Model family. A group of algorithms or models that share common characteristics, mathematical foundations or underlying assumptions.

OD (operational domain). The set of operating conditions under which a given AI-based system is specifically designed to function as intended, in line with the defined ConOps.

ODD (operational design domain). The set of operating conditions under which a given AI constituent is specifically designed to function as intended, including, but not limited to, environmental, geographical and/or time-of-day restrictions.

Outlier. Data point outside the range of at least one AI/ML constituent ODD parameter.

Responsibility – The state of being answerable for one’s decisions, actions or their consequences.

Robustness (of an AI-based system). The ability of an AI-based system to maintain its level of performance under all foreseeable conditions.

Robustness (of an AI constituent). The ability of an AI constituent to meet its allocated requirements under all non-nominal conditions within the ODD, including boundary, edge/corner and singularity cases.

Safety component. A component of a product or of an AI-based system that fulfils a safety function for that product or AI-based system, the failure or malfunction of which endangers the health and safety of persons or property.

Semantic approach. The characterisation and interpretation of parameters of the ODD based on their meaning and context within a given domain, rather than solely on raw data or low-level features. Such a characterisation involves the use of semantic information – such as objects, scenes, actions and other contextual elements – to define, describe and understand the ODD.

Shared human–AI situation awareness. The collective representation of a situation, achieved through the integration of human and AI-based system capabilities. It involves the ability of both humans and AI-based systems to gather, process, exchange and interpret information relevant to a particular context or environment, leading to a shared representation of the situation at hand, which 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 analyticity.

Situation awareness (as applicable to humans). The perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future.

Situation representation (as applicable to AI-based systems). The collection of the environment and system state and the state of the end user within a volume of time and space; the processing of this information, with the aim of enabling the extrapolation of a target status in the near future.

Socially unacceptable characteristics. Topics likely to evoke controversy in a community or strong emotional responses from its members. Such topics include those with ethical implications affecting subgroups or cultures within the society or that involve potential costs and consequent problems for citizens, like professional discrimination, personal intrusion or misallocation of responsibility.

Stability of the model. Refers to maintaining input–output relations when the model is subjected to small perturbations, that is when Xcorresponds to the input space that is within the ODD:

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

Tool (development or verification tool). A programme or a functional part thereof used to help develop, transform, test, analyse, produce or modify another programme, data, physical artefact or their documentation in their design phase.

Well-being. A state of happiness and contentment, with low levels of distress, overall good physical and mental health and outlook, and good quality of life.

DS.AI.030

AI	artificial intelligence
AL	assurance level
AL/TQL	assurance level / tool qualification level
AMC	acceptable means of compliance
ATCO	air traffic controller
ConOps	concept of operations
DQR	data quality requirement

EASA	European Union Aviation Safety Agency
EU	European Union
EUROCAE	European Organisation for Civil Aviation Equipment
GDPR	General Data Protection Regulation
GM	guidance material
ML	machine learning
OD	operational domain
ODD	operational design domain
SAE	Society of Automotive Engineering
UAS	unmanned aircraft system

DS.AI.040 Compliance process definition

The necessary processes to comply with the requirements DS.AI.100 to DS.AI.170 should be defined and documented, including the identification of the corresponding technical documentation.

DS.AI.100 Concept of operations

(a) The concept of operations (ConOps) for the AI-based system should be defined and documented and include the following:

(1) a list of the end users,

(2) the intended goals and high-level tasks to be achieved through the interaction between the end user and the AI-based system, involving other systems where necessary,

(3) a description of the operational scenarios, covering all high-level tasks,

(4) the task allocation scheme between the end user(s) and the AI-based system,

(5) how the end users will interact with the AI-based system, driven by the task allocation scheme, and

(6) the limitations and conditions on the use of the AI-based system.

(b) The involvement of each identified category of end user when defining the ConOps should be ensured.

AMC1 AI.100(a)(1) Concept of operations

END USERS

The list of end users who are intended to interact with the AI-based system should be identified and documented, together with:

(a) their roles; and

(b) their responsibilities.

AMC1 AI.100(a)(2) Concept of operations

GOALS AND HIGH-LEVEL TASKS

(a) For each end user, the intended goals to be achieved in interaction with the AI-based system should be identified and documented.

(b) For each goal achieved in interaction with the AI-based system, the high-level tasks that are intended to be allocated to the end user and to the AI-based system should be identified and documented.

GM1 AI.100(a)(2) Concept of operations

GOALS AND HIGH-LEVEL TASKS

(a) Goals and tasks describe the organisation and breakdown of what is expected of the human–AI team.

(b) A ‘goal’ is a predefined higher-level purpose towards which the teaming effort is directed.

(c) A ‘high-level task’ is a cluster of tasks contributing to the achievement of a goal. The high-level tasks should be identified at the highest level of interaction between the human and the AI-based system, not at the level of each single task performed by the AI constituent.

(d) A ‘task’ is any discrete activity contributing to the achievement of a high-level task.

(e) General principles in defining tasks include the following:

(1) a single goal can be achieved through one or more tasks;

(2) the same task can be allocated to either the end user or the AI-based system but not to both at the same time;

(3) a task that was previously allocated to an end user can be allocated to an AI-based system at a different time;

(4) each task should be discrete.

(f) An example of a 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 authority for the speed, and the pilot maintains authority for the aircraft attitude and trim. The tasks related to speed (e.g. airbrakes and throttle) are managed by the AI-based system. The pilot interferes with the management of the throttles only as necessary.

AMC1 AI.100(a)(3) Concept of operations

OPERATIONAL SCENARIOS

(a) The description of the operational scenarios should be end-user-centric.

(b) The description of the operational scenarios should cover all high-level tasks.

(c) The operational scenarios should not be limited to nominal modes but also consider degraded modes in which the AI-based system might not perform as expected.

GM1 AI.100(a)(3) Concept of operations

OPERATIONAL SCENARIOS — DEFINITION

'Operational scenario' is, in a given context and environment, a sequence of actions in response to a triggering event that aims to fulfil a high-level task.

AMC1 AI.100(a)(4) Concept of operations

TASK ALLOCATION SCHEME

(a) The task allocation scheme should divide the high-level tasks into as many tasks as necessary to reach a level of granularity that ensures all tasks can be allocated to either the end user or the AI-based system.

(b) The task allocation pattern(s) for the AI-based system should be described, including the characterisation of static or dynamic task allocation as applicable.

GM1 AI.100(a)(4) Concept of operations

TASK ALLOCATION SCHEME AND TASK ALLOCATION PATTERN

(a) 'Task allocation scheme' refers to the overall envelope of tasks that can be allocated to either the end user or the AI-based system. For instance, an AI-based system, which operates as a digital assistant to an air traffic controller (ATCO), could be involved in the high-level task of transiting all flights conflict free throughout the sector. For a set of selected aircraft entering the sector, the AI-based system could perform tasks such as detecting conflicts, resolving conflicts and/or communicating with the aircraft via radio or controller–pilot data link communications. The same AI-based system may not have permission to perform any other ATCO tasks.

(b) 'Task allocation pattern' refers to the set of tasks that are allocated to the AI-based system at a specific time. During certain periods of the shift, the ATCO could delegate communication with the aircraft to the AI-based system. During other periods, the ATCO could delegate conflict detection, conflict resolution and communication with the aircraft to the AI-based system for a set of aircraft entering the sector. These two situations represent different allocation patterns for one single allocation scheme.

(c) A static task allocation pattern refers to predefined task allocation; the tasks are clearly attributed to either the end user or the AI-based system. In a dynamic task allocation pattern, tasks are assigned to resources as they become available and in response to changing conditions; for example, in the event that the end user is occupied with another task, the AI-based system can perform the task.

AMC1 AI.100(b) Concept of operations

END-USER INVOLVEMENT

How end users' inputs are collected and accounted for in the development of the AI-based system should be documented, specifying the involvement of end-user representatives in the planning, design, validation, verification and certification/approval of an AI-based system.

DS.AI.110 AI-based system classification

A single AI level should be assessed and allocated to the AI-based system.

AMC1 AI.110 AI-based system classification

(a) The AI-based system should be classified based on the levels presented in Table 1, with adequate argumentation.

Table 12 – AI levels

AI level	Nature of the task allocated to the system to contribute to the high-level task	Authority of the end user	Responsibility of the end user
Level 1A Human augmentation	Automation support to information acquisition	Full	Remains with the end user
	Automation support to information analysis	Full	Remains with the end user
Level 1B Human assistance	Automation support to decision-making	Full	Remains with the end user
Level 2A Human–AI cooperation	Directed automatic decision and action implementation	Full	Remains with the end user
Level 2B Human–AI collaboration	Supervised automatic decision and action implementation	Partial	Remains with the end user
Level 3A Safeguarded advanced automation	Safeguarded automatic decision and action implementation	Limited, upon alerting	Reserved (operations dependent)
Level 3B	Reserved	n/a	n/a

Non-supervised advanced automation			

(b) Only systems incorporating one or more AI models should be classified.

(c) When classifying, the high-level task(s) allocated to the end user(s) in interaction with the AI-based system should be considered.

(d) When several AI levels apply to the AI-based system (e.g. because it is involved in several high-level tasks), the assigned AI level should be the highest level met by the AI-based system considering its full capability.

(e) The level of responsibility should match the end users' capacity to perform their duties, considering the nature of the interaction and control in the operations of the AI-based system. Misaligning AI-based system classification and actual authority can result in over-attributing responsibility to the end user(s).

GM1 AI.110 AI-based system classification

DELINEATION BETWEEN HIGHER LEVELS OF AI

When classifying AI-based systems, the following definitions should be considered.

(a) 'Directed' refers to the capability of the end user to actively monitor the operationally-relevant tasks allocated to the AI-based system, to cross-check every decision-making and to intervene in every action implemented by the AI-based system. This corresponds to full end-user authority.

(b) 'Supervised' refers to the capability of the end user to actively monitor the operationally-relevant tasks allocated to the AI-based system and to intervene in every action implemented by the AI-based system; however, the AI-based system takes some decisions and implements some actions in relative independence, while maintaining a shared situation awareness between both members. This corresponds to partial end-user authority.

(c) 'Safeguarded' refers to the capability of the end user to oversee the AI-based system's operations and override its authority (for selected decisions and actions) when necessary to ensure the safety and security of operations (upon alerting). This corresponds to limited end-user authority (upon being alerted). The end user may revert to full or partial authority depending on the ConOps and the nature of events occurring in the operations.

GM2 AI.110 AI-based system classification

BOUNDARIES BETWEEN AI LEVELS

When classifying an AI-based system, the following considerations support the delineation of boundaries between AI levels.

(a) The boundary between level 1A and level 1B is based on the notion of support for decision-making.

(1) Level 1A covers the use of AI for any augmentation of the information presented to the end user, ranging from prediction tasks to presentation of the information for the purpose of augmenting human end-user perception and cognition.

(2) Level 1B specifically supports decision-making, that is, the process through which the end user selects a course of action(s) from several possible alternative options, based on the AI-based system's output. The number of proposed alternatives could be two or more.

(3) The notion of support implies that the decisions are solely taken by the end user and not by the AI-based system. Therefore both levels 1A and 1B imply that the AI-based system has no decision-making capability in high-level task(s). However, depending on the ConOps, a level 1 AI-based system may automatically implement actions, based on decisions taken by the end user.

(4) An example of a level 1A AI-based system is a system that, in the fulfilment of a high-level task identified as 'detect and avoid non-cooperative traffic', is limited to the perception aspects (e.g. intruder detection). In the same scenario, an AI-based system that suggests, in the case of intruder detection, several alternative options for the pilot (as the end user) to then decide the necessary avoidance manoeuvre is classified as level 1B.

(b) The boundary between level 1B and level 2A is based on the distinction between support to decision-making on the one hand and automatic decision-making and action implementation on the other (e.g. for the high-level task 'detect and avoid non-cooperative traffic', a level 2 AI-based system could implement an avoidance manoeuvre through the aircraft's autopilot). At level 2A, it is important to remember that such automatic decisions or action implementations are fully monitored by the end user, who can override the AI-based system (e.g. the pilot could decide to override the autopilot system and perform a different manoeuvre).

(c) 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 these two levels lies in the capability of level 2B AI-based systems to take charge of some decisions without the end user cross-checking them; instead, the end user actively monitors the effect of those decisions and their implementation. For example, an executive ATCO may use a level 2B AI-based system to delegate conflict detection and resolution for certain aircraft entering the sector depending on the nature of their trajectories; certain resolutions could be managed independently by the AI-based system without cross-checking by the executive ATCO. However, the monitoring capabilities allow the executive ATCO to stop automatic decision-making.

(d) The boundary between level 2B and level 3A lies in the high level of authority of the AI-based system and the end user's limited oversight over the AI-based system's operations. A prerequisite for level 2 (A and B) is the end user's ability to intervene in every decision made and action implemented by the AI-based system, whereas in level 3A applications, the end user's ability to override the AI-based system's authority is limited to cases where to do so is necessary to ensure the safety of operations (e.g. an operator supervising a fleet of unmanned aircraft systems (UAS), automated through level 3A AI-based systems, terminating the operation of one UAS upon alerting that it is operating beyond its operating conditions).

GM3 AI.110 AI-based system classification

CONCEPT OF PARTIAL AUTHORITY DELEGATION FOR LEVEL 2B AI-BASED SYSTEMS

(a) Partial delegation of authority refers to the allocation of specific decision-making tasks from a human end user to an AI-based system, while maintaining the human's overall responsibility and oversight. This delegation is based on the type of decision and can be driven by the SRK decision classification scheme[1] considering three types of decisions.

(1) Skill-based decisions. These involve routine, repetitive tasks that require minimal cognitive effort.

(2) Rule-based decisions. These involve applying predefined rules or protocols to specific situations.

(3) Knowledge-based decisions. These involve complex, dynamic situations that require human expertise and conscious problem-solving. They include decisions taken in time-critical or highly uncertain contexts.

(b) In this context, the following apply.

(1) Skill-based and rule-based types of decisions are transferable to a level 2B AI-based system.

(2) Knowledge-based type of decisions are not transferable to a level 2B AI-based system.

DS.AI.120 AI-based system operational domain

The operational domain (OD) for the AI-based system should be defined and documented, based on the ConOps defined under DS.AI.100.

AMC1 AI.120 AI-based system operational domain

OPERATIONAL DOMAIN CHARACTERISATION

The OD in which the AI-based system is intended to operate should be characterised, including:

(a) the list of operating parameters, classified by type, together with their allowable values or range of values;

[1] Based on the skill–rules–knowledge (SRK) paradigm by Rasmussen, J., 'Skills, rules, and knowledge; Signals, signs, and symbols, and other distinctions in human performance models', *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13, No 3, 1983, pp. 257–266, https://doi.org/10.1109/TSMC.1983.6313160.

(b) dependencies between operating parameters in order to define correlated ranges between some parameters when appropriate.

GM1 AI.120 AI-based system operational domain

OPERATIONAL DOMAIN — DEFINITION

(a) 'Operational domain' (OD) is the set of operating conditions under which a given AI-based system is specifically designed to function as intended, in line with the defined ConOps.

(b) Capturing operating conditions, which corresponds to the conditions under which a given product or AI-based system is specifically designed to function as intended, is already an established practice in aviation. However, the process is not as formal as required to deal with AI-based systems; hence the formalisation of this notion under the term OD.

DS.AI.130 AI-based system risk assessment

(a) The risks posed by the AI-based system should be assessed, considering:

(1) the hazards identified and their severity,

(2) the hazards likelihood.

(b) Based on the consolidation of the risk assessment, mitigations commensurate with the risk should be put in place including:

(1) the definition of safety objectives,

(2) the definition of a preliminary system architecture,

(3) the allocation of assurance levels,

(4) the derivation of requirements, including performance and independence requirements to meet the safety objectives or service performance requirements and support the AI-based system architecture,

(5) the identification of sources of uncertainty,

(6) the analysis of exposure to out-of-range inputs,

(7) the definition and validation of assumptions.

(c) Mitigations commensurate with the risk should be verified, including:

(1) verification that all safety objectives, service performance requirements and assumptions, as applicable, are met;

(2) consolidation of the risk assessment to verify that the implementation satisfies the safety objectives or service performance requirements.

AMC1 AI.130 AI-based system risk assessment

In the absence of domain-specific guidance, a risk assessment should be performed and mitigations for the AI-based system should be put in place, including the following steps.

(a) Functional decomposition of the AI-based system operations, based on the ConOps and associated scenarios.

(b) Risk assessment, as follows.

(1) Step 1 – functional hazard assessment of the AI-based system operations. For each operational scenario, assess the severity of the hazard on the operation of the AI-based system or associated product, capturing the necessary assumptions and operational requirements, and classify the identified hazards into the following categories:

(i) H5: no possible impact on end users or general public,

(ii) H4: slight performance degradation with possible reduction in safety margins or functional capabilities or slight increase in end-user workload,

(iii) H3: significant performance degradation with reduction in safety margins or functional capabilities, potentially impacting safety-critical operations or a significant increase in end-user workload,

(iv) H2: potential for serious injury of end users or general public, a large performance degradation with a large reduction in safety margins or functional capabilities or a large increase in end-user workload,

(v) H1: potential for fatalities of end users or general public (defined for completeness but considered an unacceptable risk for an AI-based system operation at this point in time).

The functional hazard assessment step should account for the AI-based system, in conjunction with the operational procedures and the end-user contribution. When credit can be granted to human or procedural elements, the hazard can be reduced by one level.

(2) Step 2 – hazard likelihood determination. For each operational scenario, assess the likelihood of the hazard occurring, based on the outputs of step 1 of the risk assessment, and capturing additional assumptions and operational requirements, based on the definitions below.

(i) 'Frequent' refers to a hazard anticipated to occur regularly during the AI-based system's operational life. Frequent hazards are those having an average probability per operational hour greater than the order of 1E−3.

(ii) 'Probable' refers to a hazard anticipated to occur one or more times during the AI-based system's operational life. Probable hazards are those having an average probability per operational hour of the order of 1E−3 or less, but greater than 1E−5.

(iii) 'Remote' refers to a hazard that is unlikely to occur to each instance of the AI-based system during its total operational life, but may occur several times when considering the total operational life of a number of systems of this kind. Remote hazards are those having an average probability per operational hour of the order of 1E−5 or less, but greater than 1E−7.

(iv) 'Extremely remote' refers to a hazard that is not anticipated to occur to each AI-based system during its total operational life, but may occur a few times when considering the total operational life of all systems of this kind. Extremely remote hazards are those having an average probability per operational hour of the order of 1E−7 or less, but greater than 1E−9.

(v) ‘Extremely improbable’ refers to a hazard so unlikely that it is not anticipated to occur during the entire operational life of all AI-based systems of this kind. Extremely improbable hazards are those having an average probability per operational hour of the order of 1E−9 or less.

The hazard likelihood determination step can also account for credit based on human or procedural elements; however, elements used in the functional hazard assessment step should not be reused in the hazard likelihood step.

(3) Step 3 – scenarios consolidation. Table 2 should be used to determine the risk classification for the operational use of the AI-based system, considering the most stringent severity hazard category after the aggregation of all scenarios.

Table 2. Risk levels for AI-based systems

	Likelihood of the hazard				
Hazard classification	**Frequent**	**Probable**	**Remote**	**Extremely remote**	**Extremely improbable**
H1	Unacceptable risk	Unacceptable risk	Unacceptable risk	Unacceptable risk	Unacceptable risk
H2	Unacceptable risk	Unacceptable risk	Unacceptable risk	Acceptable risk	Moderate risk
H3	Unacceptable risk	Unacceptable risk	Acceptable risk	Moderate risk	Moderate risk
H4	Unacceptable risk	Acceptable risk	Moderate risk	Moderate risk	Moderate risk
H5	No risk	No risk	No risk	No risk	No risk

(i) ‘Acceptable risk’ or ‘moderate risk’ corresponds to a ‘safety component’ (high-risk AI systems) in the sense of Regulation (EU) 2024/1689.

(ii) An aggregated ‘acceptable risk’ for the most stringent hazard category supports a direct allocation of an assurance level (AL) or tool qualification level (TQL), as reflected in the second column of Table 3.

(iii) An aggregated ‘moderate risk’ for the most stringent hazard category supports a reduction by one AL or TQL as reflected in the third column of Table 3.

(iv) An aggregated ‘no risk’ for all operational scenarios involving the AI-based system supports a determination of AL6, and consequently no further requirements apply to the associated AI constituents, as they are not ‘safety components’ (of a high-risk AI system) in the sense of Regulation (EU) 2024/1689.

(v) If one operational scenario is determined an ‘unacceptable risk’, the operation is not acceptable, and the AI-based system or associated product cannot be approved.

(c) Definition of the safety objectives or service performance requirements for the AI-based system, proportionate to the hazard classification.

(d) Definition of a preliminary system architecture to meet the safety objectives or service performance requirements.

(e) Allocation of assurance or tool qualification level (AL/TQL), proportionate with the hazard classification, considering the following guidelines.

(1) An AL should be allocated to an AI-based system, based on Table 3.

Table 3. Assurance levels for AI constituents

Hazard classification	**AI constituent AL (acceptable risk level)**	**AI constituent AL (moderate risk or substantiated safety benefit)**
H1	n/a	n/a
H2	AL2/TQL2	AL3/TQL3
H3	AL3/TQL3	AL4/TQL4
H4	AL5/TQL5	AL5/TQL5
H5	AL6	AL6

(2) When a system is developed using AI technology, an AL/TQL should be allocated, based on Table 4.

Table 4. Assurance levels and tool qualification levels

Nature of the AI-based system	**Hazard contribution of the AI-based tool**	**AL / TQL**
Development tool	H3	AL3/TQL3
Development tool	H4	AL4/TQL4
Development tool	H5	AL6
Verification tool	H3	AL5/TQL5
Verification tool	H4	AL5/TQL5
Verification tool	H5	AL6

(3) If the AI-based system can be demonstrated to provide a safety benefit, AL/TQL allocation can be done as reflected in the third column of Table 3. The safety benefit should be duly substantiated, including a quantification of the expected benefit to existing risk, demonstrating a significant expected reduction in risk likelihood or impact. Requests will be considered on a case-by-case basis.

(4) Depending on the AI technology, the allocated AL/TQL should not be more stringent than the level indicated in Table 5.

Table 5. Assurance levels and tool qualification levels

AI technology	**AL / TQL**
Supervised learning	AL3/TQL3

Unsupervised learning	AL5/TQL5
Reinforcement learning	AL6
LKB	AL6

(5) No further requirements apply to an AI constituent allocated to AL6, as it is not a 'safety component' (of a high-risk AI system) in the sense of Regulation (EU) 2024/1689.

(f) Derivation of requirements, including performance and independence requirements to meet the safety objectives or service performance requirements and support the AI-based system architecture.

(g) Identification and classification of sources of uncertainty.

(h) Analysis and mitigation of exposure to data outside the OD or ODD.

(i) Analysis and mitigation of the effect of the AI-based system (respectively AI constituent) exposure to input data outside of the AI-based system OD (respectively AI constituent ODD).

(j) Definition and validation of assumptions.

(k) Verification that all safety objectives, service performance requirements and assumptions, as applicable, are met.

(l) Consolidation of the risk assessment to verify that the implementation satisfies the safety objectives or service performance requirements.

GM1 AI.130(a)(2) AI-based system risk assessment

HAZARD LIKELIHOOD DETERMINATION (RISK ASSESSMENT STEP 2 IN AMC1 AI.130(a))

(a) The quantitative hazard likelihood is expressed per operational hour. To support the hazard likelihood determination, a usage profile based on the operational scenarios (including the duration of each phase) and an average usage duration can be defined.

(b) For various reasons, data may not be precise enough to enable accurate estimates of a hazard's probability. This results in some degree of uncertainty. Significant uncertainties also apply when assessing the impact of operational procedures or end-user contribution on hazard likelihood; therefore, the following apply.

(1) When estimating the probability of each hazard, this uncertainty should be accounted for in a way that does not compromise safety or service performance, as applicable.

(2) The risk assessment of the whole AI-based system – composed of technical, human and procedural elements – is by nature a qualitative assessment. The analytical tools used in determining numerical values for the technical elements are intended to supplement, but not replace, qualitative methods based on engineering and operational judgement.

AMC1 AI.130(b)(4) AI-based system risk assessment

AI SPECIFICITIES FOR PERFORMANCE REQUIREMENTS DEFINITION

The risk assessment for the AI-based system should be complemented by the use of AI techniques, including the following steps.

(a) When a quantitative safety assessment is required, define the performance metrics of the AI constituent(s), and associated acceptance targets, thresholds and tolerances.

(b) The quantitative safety assessment should account for:

(1) uncertainties, and

(2) elevated values of error metrics.

AI-BASED SYSTEM PERFORMANCE METRICS, ACCEPTANCE TARGETS, THRESHOLDS AND TOLERANCES

(c) In the absence of dedicated domain standards, the section of ED-324/ARP6983 related to the definition of AI-based system 'performance requirements including the associated metrics, acceptance targets, thresholds and tolerances' is recognised as an acceptable means of compliance (AMC) with regard to the definition of AI-based system performance metrics, acceptance targets, thresholds and tolerances.

(d) When using ED-324/ARP6983, the following apply:

(1) the corresponding objectives of ED-324/ARP6983 associated with the AL assigned to the AI constituent should be satisfied, and corresponding life-cycle data should be developed in order to demonstrate compliance with the applicable objectives of ED-324/ARP6983;

(2) activities that satisfy each objective related to ED-324/ARP6983 should be planned and executed;

(3) the life-cycle data related to ED-324/ARP6983 should be made available upon the competent authority's request.

ACCOUNTING FOR ELEVATED VALUES OF ERROR METRICS

(e) Study the elevated values of the model's error metrics on the training/validation (eventually testing) datasets, and develop adequate mitigations, for example by:

(1) characterising regions of the operational design domain (ODD) where elevated values of the error metrics are gathered;

(2) establishing margins on performance targets and thresholds based on the evaluated generalisation gap;

(3) proposing architectural mitigations or limitations.

AMC1 AI.130(b)(5) AI-based system risk assessment

ACCOUNTING FOR IDENTIFIED UNCERTAINTIES

(a) Aleatory uncertainties should be minimised to the extent practical. Effects of aleatory uncertainties should be assessed at the system level.

(b) In the quantitative assessment, the aleatory uncertainties should be accounted for in a way that does not compromise safety (e.g. through the establishment of margins on performance targets and thresholds).

Note: Epistemic uncertainty is addressed through the learning assurance objectives.

GM1 AI.130(b)(5) AI-based system risk assessment

IDENTIFICATION AND CLASSIFICATION OF SOURCES OF UNCERTAINTIES

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

(a) 'Epistemic uncertainty' refers to deficiencies due to a lack of knowledge or information. In the context of ML, epistemic uncertainty occurs when the model has not been exposed to data adequately covering the whole ODD or where the ODD definition needs to be refined or completed.

(b) 'Aleatory uncertainty' refers to 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 containing uncertainties.

AMC1 AI.130(b)(6) AI-based system risk assessment

AI-BASED SYSTEMS EXPOSURE TO DATA OUTSIDE THE OPERATIONAL DOMAIN / OPERATIONAL DESIGN DOMAIN

The risk assessment for the AI-based system should be complemented by analysing and mitigating the effect of the AI-based system exposure to input data outside the AI-based system's OD and AI constituents' ODD.

GM1 AI.130(b)(6) AI-based system risk assessment

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

(a) establish monitoring capabilities to detect when input data is outside the AI constituents' ODD or the AI-based system's OD;

(b) put functions in place for the AI-based system to continue delivering the intended behaviour when input data is outside the ODD.

DS.AI.140 AI-based system ethics-based assessment

An ethics-based assessment should be performed to identify potential risks introduced by the use of the AI-based system.

AMC1 AI.140 AI-based system ethics-based assessment

A preliminary ethics-based assessment of the AI-based system should be performed that includes the following steps:

(a) ensure that the creation or reinforcement of unfair bias during operations involving the AI-based system, regarding both the datasets and the trained models, is avoided;

(b) for level 2 AI-based systems, ensure there is no risk of the AI-based system creating attachment, stimulating addictive behaviour or manipulating the end user's behaviour;

(c) for level 2 AI-based systems, ensure that the AI-based system and its associated ConOps do not present socially unacceptable characteristics.

GM1 AI.140 AI-based system ethics-based assessment

CHECKLIST FOR PRELIMINARY ETHICS-BASED ASSESSMENT

The following checklist can be used to perform the assessment needed to identify potential ethics-based concerns introduced by the use of the AI-based system. The answers to each question should be briefly justified.

(a) Risk of creation or reinforcement of unfair bias when using an AI-based system.

(1) Are the datasets in any form creating and/or reinforcing unfair bias or creating and/or reinforcing discrimination or any deviation from equal opportunities?

(2) Are the training models in any form creating and/or reinforcing unfair bias or creating and/or reinforcing discrimination or any deviation from equal opportunities?

(b) In the case of a level 2 AI-based system, verify and discuss with the end users the AI-based system's ConOps to ensure that unacceptable characteristics are not present.

(1) Is the end user anticipated to be totally free to exercise their human autonomy during the operations with the AI-based system?

(2) Is the end user anticipated to be free of emotional support given by the AI-based system in any work circumstance?

(3) Is the end user anticipated to depend on the AI-based system's performance in order to take action?

(4) Is the end user experiencing manipulative influence from the AI-based system and thus jeopardising their decision-making?

(5) Can the end user independently oversee the AI-based system's performance?

(c) In the case of a level 2 AI-based system, verify and discuss the AI-based system's ConOps with the end users in order to ensure there are no socially unacceptable characteristics present.

(1) Is the end user anticipated to experience professional threats such as, but not limited to, feeling controlled, unfit to perform, that they are lacking the right competencies or that their job is at stake?

(2) Is the end user's well-being and/or mental health anticipated to experience negative impacts?

(3) Is the end user anticipated to experience responsibility issues?

(4) Is the end user informed about if, when, in what circumstances, any type of their personal data will be recorded by the AI-based system?

Additional considerations:

(d) If the training of the ML model is based on personal data, check whether compliance with Regulation (EU) 2016/679 (the GDPR Regulation) is needed.

(e) If an environmental impact is expected from the model training and/or operations, check whether compliance with applicable aviation sustainability regulations is needed.

AMC2 AI.140 AI-based system ethics-based assessment

MITIGATION OF RISK OF UNFAIR BIAS

(a) If a risk of creating or reinforcing unfair bias has been identified, to avoid discrimination or unfair treatment caused by the AI-based system's outputs, the AI-based system should be designed considering:

(1) means of raising awareness among everyone 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 (e.g. an ethics-based policy, procedures, guidance or controls);

(2) removing identified unfair bias through the development assurance processes.

MITIGATION OF RISK OF ATTACHMENT, ADDICTIVE BEHAVIOUR AND/OR MANIPULATION

(b) For level 2 AI-based systems, if a risk is anticipated related to attachment, addictive behaviour and/or manipulation, requirements-based tests should include verification that the end users interacting with the AI-based system can perform oversight.

SOCIETAL ACCEPTANCE OF THE AI-BASED SYSTEM CONOPS

(c) To manage the risk of a level 2 AI-based system's socially unacceptable impact:

(1) the end user(s) should be involved in defining the ConOps in order to identify potential socially unacceptable impacts, particularly in terms of:

(i) professional threats, job extinction and consequent unemployment,

(ii) personal or psychological discomfort and consequent negative impacts on well-being and mental health,

(iii) clearly defined borderlines of responsibility.

(2) end users should be informed whether some personal data is recorded by the system.

DS.AI.150 AI-based system intended behaviour

The AI-based system should be designed so that it performs as intended under the specified OD at the specified level of performance.

AMC1 AI.150 AI-based system intended behaviour

SYSTEM DEVELOPMENT ASSURANCE

A system development assurance process should be planned and implemented including the following steps:

(a) system requirements and design management, including the definition of the system requirements and architecture,

(b) requirements validation, ensuring the validation of all requirements, their traceability to the higher-level tasks definition or higher-level requirements if refined, and the identification of derived requirements and their feedback to the safety assessment processes,

(c) requirements verification, ensuring that the integrated system implementation meets the intended behaviour and that development errors have been addressed to a degree of confidence commensurate with the AL or TQL allocated to the AI constituent of the AI-based system,

(d) configuration management, ensuring the compliance and integrity of all the AI-based system's life-cycle data,

(e) process assurance, ensuring that the required activities have been completed as outlined in plans, or that deviations have been identified and mitigated.

GM1 AI.150 AI-based system intended behaviour

REFERENCE STANDARD FOR AI-BASED SYSTEM DEVELOPMENT ASSURANCE

In the absence of domain-specific standards, the latest version of EUROCAE ED-79() / SAE ARP4754() can be adapted to support the definition of the AI-based system development assurance process.

AMC2 AI.150 AI-based system intended behaviour

AI/ML CONSTITUENTS LEARNING ASSURANCE

(a) Each AI/ML constituent should be designed to perform as intended under the specified ODD, at the specified level of performance, and to provide sufficient generalisation and robustness properties, at a level of confidence commensurate with the AL/TQL that was determined through the risk assessment.

RECOGNISED STANDARD FOR AI/ML CONSTITUENTS LEARNING ASSURANCE

(b) EUROCAE ED-324() / SAE ARP6983() is recognised as an acceptable means of compliance regarding the intended behaviour of AI/ML constituents integrating ML models trained with supervised learning methods. When using ED-324()/ARP6983(), the following should apply:

(1) all of the objectives associated with the AL assigned to the AI/ML constituent should be satisfied, and all of the associated life-cycle data should be developed to demonstrate compliance with the applicable objectives, considering the following clarifications:

(i) the operating environment introduced in the standard should be fully traceable to the AI-based system's OD, that is to say equal to or more refined than the OD;

(ii) the objective associated with the AI-based system performance requirements, metrics, thresholds, targets and tolerances are addressed as part of the risk assessment; see AMC2 AI.130;

(2) activities that satisfy each applicable objective should be planned and executed;

(3) the life-cycle data specified in ED-324()/ARP6983(), applicable tool qualification data and any other data needed to substantiate the satisfaction of the applicable objectives should be made available upon the competent authority's request.

AMC3 AI.150 AI-based system intended behaviour

AI/ML CONSTITUENTS LEARNING ASSURANCE

(a) Each AI/ML constituent should be designed to perform as intended under the specified ODD, at the specified level of performance, and to provide sufficient generalisation and robustness properties at a level of confidence commensurate with the AL/TQL that was determined through the risk assessment.

In the event that EUROCAE ED-324()/SAE ARP6983() is not applied, the following steps (b) to (m) provide a set of abstracted objectives that should be complied with when planning and executing an appropriate AI/ML constituent learning assurance process.

AI/ML CONSTITUENT REQUIREMENTS, ARCHITECTURE AND OPERATIONAL DESIGN DOMAIN

(b) As part of the AI/ML constituent requirements management process, the following steps should be planned and executed:

(1) capture the AI/ML constituent requirements;

(2) define the set of parameters pertaining to the AI/ML constituent ODD, with particular attention to:

(i) the definition of nominal data,

(ii) the identification of edge cases and corner cases data in preparation of the model's stability and robustness verification,

(iii) the identification of singular points,

(iv) the detection and management of outliers;

(3) trace the ODD parameters to the corresponding parameters pertaining to the OD when applicable;

(4) capture the data quality requirements (DQRs) for all data required for training, testing and verification of the AI/ML constituent(s);

(5) capture the requirements on data to be preprocessed and engineered for the inference model in development and for the operations;

(6) for AI constituents allocated an AL4/TQL4 or more stringent, describe a preliminary AI/ML constituent architecture;

(7) validate each of the constituent requirements;

(8) document evidence that all derived requirements generated through the learning assurance processes have been provided to the system processes, including the safety (support) assessment; and

(9) document evidence of the validation of the derived requirements and of the determination of any impact on the safety (support) assessment and system requirements.

DATA MANAGEMENT

(c) As part of the data management process, the following steps should be planned and executed:

(1) identify data sources and collect data in accordance with the defined ODD, while ensuring the defined DQRs are satisfied, in order to drive the selection of the training, validation and test datasets;

(2) if following a supervised learning approach, ensure that all data is labelled and the annotated or labelled data in the dataset satisfies the DQRs;

(3) for AI constituents allocated an AL3/TQL3 or more stringent, define and document the preprocessing operations on the collected data in preparation of the model training;

(4) for AI constituents allocated an AL3/TQL3 or more stringent, define and document the necessary transformations of the preprocessed data from the specified input space into features that are effective for the performance of the selected learning algorithm;

(5) distribute the data into three separate datasets (training, validation and test), which meet the specified DQRs in terms of independence;

(6) if following an unsupervised learning approach, ensure that the annotated or labelled data in the test dataset satisfies the DQRs;

(7) ensure verification of the data, as appropriate, throughout the data management process so that the data management requirements (including the DQRs) are addressed.

LEARNING MANAGEMENT

(d) As part of the learning management process, the following steps should be planned and executed:

(1) describe the ML model architecture;

(2) capture the requirements pertaining to the learning management and training processes;

(3) for AI constituents allocated an AL3/TQL3 or more stringent, document the verification credit sought from the training environment and qualify the environment accordingly;

(4) for AI constituents allocated an AL3/TQL3 or more stringent, provide quantifiable generalisation bounds and ensure that the assumptions used to derive them are valid;

(5) for AI constituents allocated an AL3/TQL3 or more stringent, document any model optimisation performed at the end of the ML model training and provide evidence that model optimisations do not alter the model's behaviour or performance;

(6) for AI constituents allocated an AL3/TQL3 or more stringent, if following a supervised learning approach, account for the bias–variance trade-off or optimisation in the model family selection;

(7) document the result of the ML model training.

LEARNING VERIFICATION

(e) As part of the learning verification process, the following steps should be planned and executed:

(1) for AI constituents allocated an AL3/TQL3 or more stringent, ensure that the estimated bias and variance of the selected model meet the associated learning process management requirements;

(2) evaluate the trained model's performance based on the test dataset, and document the result of the model verification;

(3) perform requirements-based verification of the trained model behaviour;

(4) for AI constituents allocated an AL3/TQL3 or more stringent, provide an analysis of the repeatability of the learning process;

(5) perform and document the verification of the trained model's stability;

(6) perform and document the verification of the trained model's robustness under non-nominal conditions within the ODD (including boundary, edge/corner and singularity cases);

(7) for AI constituents allocated an AL3/TQL3 or more stringent, verify the anticipated generalisation capabilities of the ML model(s) using the test dataset;

(8) capture the description of the resulting ML model.

ML MODEL IMPLEMENTATION

(f) As part of the model implementation process, the following steps should be planned and executed:

(1) capture the requirements pertaining to the ML model implementation process;

(2) validate the ML model description(s) and captured implementation requirements;

(3) document evidence that all the derived requirements generated through the model implementation process have been provided to the system processes, including the safety (support) assessment;

(4) for AI constituents allocated an AL3/TQL3 or more stringent, identify and validate the impact of the post-training model transformations on the behaviour and performance of the model(s), and identify the development environment necessary to perform the model transformations;

(5) plan and execute appropriate development assurance processes to develop the inference model(s) into software and/or hardware items, using the applicable domain-specific acceptable means of compliance (AMC) / guidance material (GM) for software and/or hardware development assurance; in the absence of applicable domain-specific AMC/GM for software and/or hardware development assurance, a tool qualification approach using ED-215/DO-330 at the corresponding TQL may be used;

(6) for AI constituents allocated an AL3/TQL3 or more stringent, verify that any transformation performed during the trained model implementation has no negative impact on the behaviour and performance of the inference model(s);

(7) evaluate the performance of the inference model based on the test dataset, and document the result of the model verification;

(8) perform and document the verification of the stability of the inference model(s);

(9) perform and document the verification of the robustness of the inference model(s) under adverse conditions.

AI/ML CONSTITUENT INTEGRATION AND VERIFICATION PROCESS

(g) As part of the AI/ML constituent integration and verification process, the following steps should be planned and executed:

(1) integrate the inference model with all other items of the AI/ML constituent;

(2) perform requirements-based verification of the AI/ML constituent(s);

(3) perform and document the verification of the robustness of the AI/ML constituent(s) under adverse conditions;

(4) for AI constituents allocated an AL3/TQL3 or more stringent, verify the anticipated generalisation capabilities of the AI/ML constituent.

AI/ML CONSTITUENT VERIFICATION OF VERIFICATION PROCESS

(h) For AI constituents allocated an AL3/TQL3 or more stringent, as part of the AI/ML constituent verification of verification process, the following steps should be planned and executed:

(1) perform a data verification step to confirm the appropriateness of the defined ODD and of the datasets used for the training, validation and verification of the ML model, based on stop criteria negotiated with the competent authority;

(2) confirm that the trained model verification activities are complete, based on criteria negotiated with the competent authority;

(3) confirm that the AI/ML constituent verification activities are complete, based on criteria negotiated with the competent authority;

(4) provide necessary AI development explainability to persons who develop, provide, deploy, maintain, approve or support safety investigation of the product through:

(i) characterisation of the need for AI development explainability to be provided to identified users, which is necessary to help detect unintended behaviour and support the AI's operational explainability, as applicable;

(ii) identification and documentation of the methods at the output level satisfying the specified AI development explainability needs, as defined under the previous point.

REUSE OF ML MODELS OR UPGRADING A DEVELOPMENT BASELINE

(i) When reusing previously approved ML models or upgrading a development baseline to address a higher AL/TQL, an impact assessment of the reuse of a trained ML model should be performed before incorporating the model into an AI/ML constituent, considering:

(1) alignment and compatibility of the intended behaviours of the ML models, including a functional analysis to confirm adequacy for the requirements and architecture of the AI/ML constituent;

(2) alignment or compatibility of the ODDs;

(3) compatibility of the performance of the reused ML model with the performance requirements expected for the new application;

(4) availability of adequate technical documentation (e.g. equivalent documentation depending on the required AL);

(5) if applicable, availability and quantification of service experience using the identical ML model or AI/ML constituent (frozen design) in a compatible ODD context, with substantiation of its sufficiency to address limited verification objectives (stability, robustness, generalisation or certain reviews and analyses); exception is made for those related to requirements-based verification and performance verification in nominal regions of the ODD; and

(6) evaluation of the required AL/TQL.

(j) Based on this impact assessment, the necessary steps should be planned to integrate and verify the ML model into the AI/ML constituent, considering two different cases: off-the-shelf (OTS) ML models, and previously developed ML models, both with and without transfer learning.

USE OF OTS ML MODELS, WITH OR WITHOUT TRANSFER LEARNING

(k) When using OTS ML models, the following steps should be additionally planned and executed:

(1) perform an analysis of the unused functions of the OTS ML model and ensure the deactivation of these unused functions;

(2) ensure safe incorporation of the OTS model, so that it does not contribute to more than an H4-level hazard;

(3) if transfer learning is planned, execute the overall learning assurance steps described in (b) to (h).

PREVIOUSLY DEVELOPED ML MODELS, WITH OR WITHOUT TRANSFER LEARNING

(l) When the impact analysis shows that the ML model does not need to be modified, credit can be taken from a previous approval.

(m) If transfer learning is planned, execute the overall learning assurance steps described in (b) to (h).

GM1 AI.150 AI-based system intended behaviour

AI/ML CONSTITUENT OPERATIONAL DESIGN DOMAIN — DEFINITION

(a) 'Operational design domain' (ODD) is the set of operating conditions under which a given AI constituent is specifically designed to function as intended, including but not limited to environmental, geographical and/or time-of-day restrictions. The ODD defines the set of operating parameters, together with the range and distribution within which the AI constituent is designed to operate. The ODD also considers dependencies between operating parameters in order to refine the ranges between these.

(b) The level of detail captured in the OD at the system level is not commensurate with the level of detail typically needed at the AI constituent level to serve the AI model design processes, particularly the data and learning management steps. This is why the additional notion of the AI constituent ODD has been introduced.

(c) For ML, the ODD provides a framework for the selection, collection and preparation of the data during the learning phase, and the monitoring of the data during operations. A correct and complete definition of the ODD is a prerequisite for adequate-quality datasets involved in the learning assurance process.

AI/ML CONSTITUENT OPERATIONAL DESIGN DOMAIN — CONSIDERATIONS

(d) Additional parameters can be identified and defined for the AI constituent's ODD (e.g. parameters linked to sensors used for the AI model's input data, like brightness and 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.).

(e) Some operating parameters require a semantic approach for their definition, especially in high-dimension use cases, such as computer vision.

(f) Ranges for the parameters in the AI constituent's ODD can be a subset of the ranges at the OD level, limiting the design to an area of the OD where the AI model's performance aligns with the captured requirements (e.g. more restrictive weather conditions for the ODD than for the corresponding OD).

(g) Exceptionally, one or a few ranges for the parameters in the AI constituent's ODD can be an extension of the ranges for the corresponding parameters at the OD level (in order to improve the model's performance in those extended ranges).

(h) As for the OD, the range(s) for one or several operating parameters could depend on the value or range of another parameter.

(i) An analysis of the distribution of each parameter can be established. For certain parameters, especially the ones defined using a semantic approach in high-dimension use cases, the distribution can be approximated or reduced to low-dimensional representations.

(j) In relation to the iterative nature of the process aiming at characterising the ODD, stop criteria can be established, based on the achievement of some of the AI model's or AI constituent's performance requirements.

(k) In the case of unsupervised learning, characterising the ODD might be 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 involves an even more iterative approach than in supervised learning.

PREVIOUSLY DEVELOPED ML MODELS OR AI/ML CONSTITUENTS — DEFINITION

(l) 'Previously developed ML models or AI/ML constituents' refers to ML models or integrated AI/ML constituents already developed for use, either previously approved and used operationally or previously installed in the relevant operational context without being functionally activated while collecting outputs in support, for instance, of evaluating its stability, robustness or generalisation capabilities.

(m) The following depicts the cases considered for previously approved ML models:

(1) change of installation,

(2) change of application,

(3) change of development environment,

(4) change of inference platform,

(5) change of AL/TQL.

DS.AI.160 AI-based system continuous risk assessment

The AI-based system should be designed with means for in-service monitoring and recording to support the continuous risk assessment of AI-based systems, including the following steps:

(a) identify and document in the design the AI-based system in which the AI constituent's inputs and outputs need to be monitored;

(b) identify and document which data needs to be recorded for the purpose of supporting the AI-based system's continuous risk assessment and usage, and accident or incident investigations, as applicable;

(c) provide the means to record operational data as identified in point (b) and retrieve this data;

(d) define metrics, target values, thresholds and evaluation periods to guarantee that design assumptions hold.

AMC1 AI.160(a) AI-based system continuous risk assessment

AI-BASED SYSTEM MONITORING CAPABILITIES

For AI constituents allocated an AL4/TQL4 or more stringent, the AI-based system should be designed with the ability to:

(a) monitor that its inputs are within the specified ODD boundaries in which the AI/ML constituent’s performance is guaranteed;

(b) deliver an indication of the level of confidence in the AI/ML constituent’s output, based on actual measurements or on quantification of the level of uncertainty;

(c) monitor that the AI/ML constituent’s outputs are within the specified operational level of confidence.

AMC1 AI.160(b) AI-based system continuous risk assessment

LIST OF DATA TO BE RECORDED

(a) The list of data to be recorded in support of the points below should be established and documented:

(1) informing persons who develop, provide, deploy, maintain, approve or support safety investigation of the product how to retrieve this data;

(2) explaining, post operations, the AI-based system's behaviour and interactions with the end user(s);

(3) monitoring in-service events to detect potential issues or suboptimal performance trends that might contribute to safety margin erosion or to service performance degradations; and

(4) guaranteeing that design assumptions hold (this typically covers assumptions made about the ODD, e.g. for further assessment of possible distribution shift).

(b) The outputs of all monitoring capabilities identified per AMC1 AI.190(a) should be listed in the data to be recorded.

DATA RECORDING FOR THE PURPOSE OF MONITORING THE AI-BASED SYSTEM'S OPERATIONS

(c) The recorded data should contain sufficient information to detect deviations from the AI-based system's expected behaviour, whether it is operating alone or interacting with an end user. In addition, this information should be sufficient:

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

(2) 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;

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

(d) The means of retrieving the recorded data should be provided to those entitled to access and use the data in a way that facilitates their effective monitoring of the safety of AI-based system operations. This includes:

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

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

(3) the possibility of gathering the recorded data over longer periods of time and the possibility of automatically processing part of this data for trend analyses and statistical studies.

DATA RECORDING FOR THE PURPOSE OF ACCIDENT OR INCIDENT INVESTIGATION

(e) 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:

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

(2) identify when communication or teaming between the AI-based system and the end user was degraded, which may require recording additional communications between the end user and other human–AI team members or with other organisations (including voice communications), or recording additional actions performed by the end user at their workstation (e.g. by means of images), as necessary;

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

(f) The data should be recorded in such a way that it can be retrieved and used after an accident or an incident. This includes:

(1) if the AI-based system is airborne, a crashworthy memory medium on board the aircraft;

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

(3) if the AI-based system is airborne, means to retrieve the data from the memory medium after an incident or accident (e.g. means to locate the accident scene and the memory media, tools to retrieve data from damaged memory media);

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

AMC1 AI.160(c) AI-based system continuous risk assessment

START AND STOP LOGIC FOR DATA-RECORDING CAPABILITIES

(a) 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.

(b) The recording should automatically stop when or after the AI-based system is no longer operating.

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

(d) When defining the metrics, the following should be covered:

(1) the acquisition of safety-relevant data related to accidents and incidents (e.g. near-miss events),

(2) the monitoring of in-service data to detect potential issues or suboptimal performance trends that might contribute to safety margin erosion,

(3) definitions of target values, thresholds and evaluation periods, and

(4) analysis of data to determine the possible root cause and trigger corrective actions.

(e) The safety margin erosion should be evaluated when updating the analysis made during the initial safety assessment with in-service data to ensure that the safety objectives are met throughout the product’s life.

DS.AI.170 Human-centred design considerations for AI-based systems

(a) The AI-based system should be developed to provide end users with understandable, reliable and relevant information, with the appropriate level of detail and with appropriate timing, about how the system produces its results.

(b) The AI-based system should be developed to interact safely with the end user, by building situational representation and providing human–AI coordination mechanisms, efficient modalities of interaction and error/failure management.

AMC1 AI.170(a) Human-centred design considerations for AI-based systems

OPERATIONAL EXPLAINABILITY

(a) Level 1B and level 2 AI-based systems should be developed to provide end users with understandable, reliable and relevant information, with the appropriate level of detail and with appropriate timing, about how an AI-based system produces its results. To this purpose, the following steps should be planned and executed:

(1) Characterise the need for explainability for each of the AI-based system’s outputs; if the need is confirmed, apply steps (2) to (4).

(2) Select appropriate methods from the AI development explainability analysis and document explainability requirements for each relevant output.

(3) Verify compliance with the following AI operational explainability attributes:

(i) The relevance of the explanations should be defined so that the end user receiving the information can use the explanation to assess the appropriateness of a resulting decision/action (see GM1 AI.170(a)).

(ii) The level of abstraction of the explanations should be defined, taking into account the characteristics of the task and the situation. Where a customisation capability is available, the end user should be able to customise the level of abstraction as part of the AI operational explainability (see GM2 AI.170(a)).

(iii) The timing of when explanations will be available to the end user should be defined, taking into account the time criticality of the situation, the needs of the end user and the operational impact. Where applicable, the AI-based system should be designed to enable the end user to get, upon request, an explanation when needed (see GM3 AI.170(a)).

(iv) Ensure the validity of the explanation.

(b) End users should be made aware of the fact that they are interacting with an AI-based system.

GM1 AI.170(a) Human-centred design considerations for AI-based systems

AI OPERATIONAL EXPLAINABILITY ATTRIBUTE: UNDERSTANDABLE AND RELEVANT

(a) 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(s) intended to perform the task.

(b) The explanation of a system’s output is relevant if the end user receiving the information can use it to assess the appropriateness of the decision and action as expected. As an example, an initial set of information that could be conveyed by an explanation might include the following.

(1) 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 end user. This increases the usability and the utility of the explanation.

(2) Historical perspectives. To understand the relevance of the AI-based system proposal, it is important for the end user to get a clear overview of the assumptions and context used by the AI-based system.

(3) Information on the 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.

(4) 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 of the decision.

(5) 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.

GM2 AI.170(a) Human-centred design considerations for AI-based systems

AI OPERATIONAL EXPLAINABILITY ATTRIBUTE: LEVEL OF ABSTRACTION

(a) The level of abstraction corresponds to the degree of detail provided by the explanation. There are different possible arguments to substantiate the explainability. The level of detail of these arguments and the number of arguments provided in an explanation may vary depending on several factors, such as:

(1) the end user’s level of expertise. An experienced end user will not have the same needs as a novice end user in terms of rationale and details provided by the AI-based system to

understand how the system came to its results: a novice might need advice and/or detailed information to be able to follow a proposition coming from the AI-based system;

(2) 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.

(b) The level of abstraction has an impact on the collaboration between the AI-based system and the end users. To enhance this collaboration during operations, there is a possible need to customise the level of detail provided for the explanation. This can be tackled in three ways.

(1) First, the designer could set a default level of abstraction depending on factors identified during the AI's development.

(2) Second, the end users could customise the level of abstraction, possibly through presettings. If the abstraction level is not tailored to the end users' needs or level of experience, the explainability can go against its objective.

(3) Third, 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.

GM3 AI.170(a) Human-centred design considerations for AI-based systems

AI OPERATIONAL EXPLAINABILITY ATTRIBUTE: TIMELINESS

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 time frames: before operations and during operations.

(a) Before operations is also known as latent explainability. The knowledge gained by the end user during training about the way an AI-based system works should be considered to contribute to the end user's ability to decrypt the AI-based system's actions and decisions during operations. To this end, the AI-based system's manual needs to provide knowledge about all possible functionalities, behaviours and results in all relevant situations. This can be considered latent explainability. The end users retrieve this knowledge to build their situation awareness, calculate 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's customisation made by the operators/airlines to answer specific operational needs could be provided to the end users before operations.

(b) During operations, the following trade-offs should be considered.

(1) Before the decision and/or action taken by the AI-based system. Information should be provided before the decision or action in case the outcome of the decision and/or action impacts the conduct of the operation. As an example for airborne operations, if an AI-based system has the capability to lower the undercarriage, it will be necessary to provide this information to the crew (for acknowledgement or not) before the action is

performed, as it will impact the aircraft's performance. Another general reason could be to avoid any startle effect and provide the end user with sufficient notice to react appropriately to the decision and/or action.

(2) During the decision-making process and/or action. Explanation provided during the decision-making process and/or 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 and/or action.
Note: The more information relates to short-term tactical approaches, the more it should be provided before the decision and/or 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 operations.

(3) After the decision-making process and/or action. The AI-based system could be designed to provide the explanation after the decision and/or action to update the situation awareness of the end users, or to allow the end user to request an explanation on demand.

AMC1 AI.170(b) Human-centred design considerations for AI-based systems

AI-BASED SYSTEM SITUATION REPRESENTATION

(a) Level 2 AI-based systems should be designed with the ability to:

(1) build their own individual situation representation; and

(2) enhance the end user's individual situation awareness.

(b) Level 2B AI-based systems should be designed with the ability to enable and support a human–AI shared situation awareness.

DECISION CROSS-CHECK VALIDATION

(c) If a decision is taken by the AI-based system that requires validation based on procedures, the AI-based system should be designed with the ability to request a cross-check validation from the end user.

COMPLEX SITUATIONS UNDER NORMAL OPERATIONS

(d) Level 2B AI-based systems, for complex situations under normal operations, should be designed with the ability to:

(1) identify a suboptimal solution that could have a negative impact on safety and propose, through argumentation, an improved solution; and

(2) process and act upon a proposal rejection from the end user.

COMPLEX SITUATIONS UNDER ABNORMAL OPERATIONS

(e) Level 2B AI-based systems, for complex situations under abnormal operations, should be designed with the ability to:

(1) identify the problem; share the diagnosis, including the root cause, the resolution strategy and the anticipated operational consequences; and

(2) process and act upon arguments shared by the end user.

TIME-CRITICAL SITUATIONS

(f) Level 2B AI-based systems should be designed with the ability to detect poor decision-making by the end user in a time-critical situation, and alert and assist the end user.

HUMAN–AI TEAMING: ALTERNATIVE SOLUTIONS SUGGESTION

(g) Level 2B AI-based systems should be designed with the ability to propose alternative solutions and support their positions.

DYNAMIC TASK ALLOCATION PATTERN

(h) Level 2B AI-based systems, if expected to be involved in dynamic task allocation, should be designed with the ability to implement a modification of the task allocation pattern requested by the end user (instantaneously or in the short term).

GM1 AI.170(b) Human-centred design considerations for AI-based systems

AI-BASED SYSTEM SITUATION REPRESENTATION

(a) Level 2 AI-based systems are designed to build a situation representation by collecting, analysing, consolidating or monitoring data from multiple sources, representing relevant aspects of the situation.

(b) As the AI-based system can analyse multiple systems more rapidly than the end user, the end user can refer to the AI-based system to reinforce their own situation awareness. The end user can build and reinforce their situation awareness by acquiring information from the AI-based system.

(c) ‘Human–AI shared situation awareness’ refers to the collective understanding and perception of a situation, achieved through integrating human and AI-based system capabilities. It involves the ability of both humans and AI-based 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.

COMPLEX SITUATIONS UNDER NORMAL OPERATIONS

(d) Complex situations from the end-user perspective can be those associated with high workload. Stress can result in cognitive tunnelling, wherein 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 and more complex situations being created consequently, thus reducing the safety margins.

(e) To support the end user in complex situations, the AI-based system can be capable of monitoring the situation and determining what the current situation is, as part of its situation representation. The system can also monitor the actions taken by the end user in complex situations and determine what the outcomes of these actions will be.

(f) If the AI-based system's solution differs from the end user's solution, then the AI-based system can propose one or more alternative solutions. The outcome of alternative solutions can be defined in terms of the constraints of the operational system, such as time, workload, route miles flown and time at preferred altitude.

COMPLEX SITUATIONS UNDER ABNORMAL OPERATIONS

(g) The AI-based system is expected to be capable of assisting end users during complex and abnormal operations. The AI-based system can identify the problem, share the diagnosis, identify the root cause, describe a resolution strategy and advise of anticipated consequences.

TIME-CRITICAL SITUATIONS

(h) Time-critical situations are situations that require an immediate reaction — such as high traffic flow density periods (typically morning and evening) for air traffic management / air navigation services — or that require an immediate manoeuvre from a pilot, or emergency situations. The AI-based system can:

(1) detect time-critical situations, through, for instance, phase of flight, rate of system interaction, traffic pattern or voice stress pattern;

(2) contextualise the decision being made by the end user, advise on what the outcomes of the decision will be and suggest alternative actions that might better align with the end goal;

(3) identify, based on its situation representation, a more optimal suggestion that will achieve the goal more effectively;

(4) communicate with the end user to deliver the solution in a manner that fits the time-critical situation the end user is facing;

(5) provide sufficient time to enable the end user to execute the appropriate action.

AMC2 AI.170(b) Human-centred design considerations for AI-based systems

INTERACTION MODALITY: SPOKEN LANGUAGE

(a) For level 2 AI, if spoken procedural language is used, the syntax of the spoken procedural language should be designed so that it can be learned and applied easily by the end user.

(b) For level 2 AI, if spoken (procedural or natural) language is used, the AI-based system should be designed:

(1) with the ability to process end-user verbal communication (e.g. requests, statements, responses and reactions) and provide acknowledgement of the end user's actions;

(2) with the ability to not interfere with other communications or activities on the end user's side; and

(3) so that this modality can be deactivated for the benefit of other modalities.

(c) For level 2B AI, if spoken (procedural or natural) language is used, the AI-based system should be designed with the ability to:

(1) identify through the end-user responses or their action that there was a possible misinterpretation on the part of the end user and notify them;

(2) in the event of confirmed misunderstanding or misinterpretation, provide information on corrective measures;

(3) provide information regarding the associated AI-based system capabilities and limitations;

(4) assess the performance of the dialogue; and

(5) transition between spoken natural language and spoken procedural language, depending on the performance of the dialogue, the situation's context and the task's characteristics.

INTERACTION MODALITY: GESTURE LANGUAGE

(d) For level 2 AI, if gesture language is used, the AI-based system should be designed:

(1) with a gesture language syntax that is intuitively associated with the command that it is supposed to trigger; and

(2) with the ability to disregard unintentional gestures.

(e) For level 2B AI, if gesture language is used, the AI-based system should be designed with the ability to:

(1) recognise the end user's intention; and

(2) acknowledge the end user's intention with appropriate feedback.

AMC3 AI.170(b) Human-centred design considerations for AI-based systems

MULTIMODAL INTERACTION: MODALITY OF INTERACTION FOR HUMAN–AI COLLABORATION

For level 2B AI, the AI-based system should be designed with the ability to:

(a) combine or adapt the interaction modalities depending on the characteristics of the task, the operational event and/or the operational environment; and

(b) automatically adapt the modality of interactions to the end-user states, the situation, the context and/or the perceived end user's preferences.

AMC4 AI.170(b) Human-centred design considerations for AI-based systems

ERROR MANAGEMENT

(a) For level 2B AI, the AI-based system should be designed to:

(1) minimise the likelihood of design-related end-user errors;

(2) minimise the likelihood of human–AI resource-management-related errors;

(3) tolerate end-user errors;

(4) detect the errors made by the end user while they are interacting with the AI-based system; and

(5) efficiently inform the end user once an error is detected.

FAILURE MANAGEMENT

(b) For level 2B AI, the AI-based system should be designed with the ability to:

(1) diagnose the failure and present the pertinent information to the end user;

(2) if necessary, propose a solution to the failure to the end user;

(3) support the end user in the implementation of the solution; and

(4) inform the end user that logs of system failures are kept for subsequent analysis.