Innovative approaches to approval and certification

Deliverable ID:	**D3.2**
Project acronym:	**HUCAN**
Grant:	**101114762**
Call:	**HORIZON-SESAR-2022-DES-ER-0**
Topic:	**HORIZON-SESAR-2022-DES-ER-01-WA1-2**
Consortium coordinator:	**Deep Blue**
Edition date:	**30 August 2024**
Edition:	**02.00**
Status:	**Official**
Classification:	**PU**

Abstract

This deliverable presents the most innovative approaches to the approval and certification process of automated and AI-based technology and analyses them showing pros and cons of each of them, as well as suitability for the aviation domain.

Authoring & approval

Author(s) of the document

Organisation name	Date
Giuseppe Contissa | EUI	21.05.2024
Federico Galli | EUI	21.05.2024
Marco Sanchi | EUI	21.05.2024
Mariken Everdij | NLR	21.05.2024
Henk Hesselink | NLR	21.05.2024
Sybert Strove | NLR	21.05.2024
Jin Choi | NLR	21.05.2024
Paola Lanzi | DBL	21.05.2024
Elisa Spiller | DBL	21.05.2024
Giuseppe Contissa |EUI	22.08.2024
Marco sanchi | EUI	22.08.2024

Reviewed by

Organisation name	Date
Gabriella Gigante |CIRA	24.05.2024
Edoardo Fornaciari | D-Flight	24.05.2024
Mohsan Jameel | DLR	24.05.2024
Mariken Everdij |NLR	23.08.2024
Sybert Strove |NLR	23.08.2024
Paola Lanzi | DBL	29.08.2024
Elisa Spiller | DBL	29.08.2024

Approved for submission to the SESAR 3 JU by[1]

Organisation name	Date
Paola Lanzi | DBL	29.05.2024
Mariken Everdij | NLR	29.05.2024
Giuseppe Contissa | EUI	29.05.2024
Gabriella Gigante | CIRA	29.05.2024
Mohsan Jameel | DLR	29.05.2024
Edoardo Fornaciari | D-Flight	29.05.2024

[1] Representatives of all the beneficiaries involved in the project

Giuseppe Contissa | EUI	22.08.2024
Mariken Everdij | NLR	29.08.2024
Paola Lanzi | DBL	29.08.2024

Rejected by[2]

Organisation name	Date
S3JU	27.07.2024

Document history

Edition	Date	Status	Company Author	Justification
0.1	15.11.2023	Draft	Giuseppe Contissa | EUI Federico Galli | EUI Marco Sanchi | EUI	ToC
0.2	15.03.2024	Draft	Federico Galli | EUI Marco Sanchi | EUI Mariken Everdij | NLR Sybert Strove | NLR Elisa Spiller | DBL	Draft
0.3	21.05.2024	Draft	Federico Galli | EUI Marco Sanchi | EUI Mariken Everdij | NLR Sybert Strove | NLR Jin Choi | NLR Elisa Spiller | DBL	Final for review
0.4	27.04.2024	Draft	Federico Galli | EUI Marco Sanchi | EUI Mariken Everdij | NLR Sybert Strove | NLR Jin Choi | NLR Elisa Spiller | DBL	Final for approval
1.0	31.05.2024	Official	Federico Galli | EUI Marco Sanchi | EUI Mariken Everdij | NLR Sybert Strove | NLR Jin Choi | NLR	Final

[2] Representatives of the beneficiaries involved in the project

			Elisa Spiller | DBL	
1.1	22.08.2024	Draft	Federico Galli | EUI Marco Sanchi | EUI Mariken Everdij | NLR Sybert Strove | NLR	Final for review and approval
2.0	30.08.2024	Official	Federico Galli | EUI Marco Sanchi | EUI Mariken Everdij | NLR Sybert Strove | NLR Paola Lanzi | DBL Elisa Spiller | DBL	Final

HUCAN

HOLISTIC UNIFIED CERTIFICATION APPROACH FOR NOVEL SYSTEMS BASED ON ADVANCED AUTOMATION

HUCAN

This document is part of a project that has received funding from the SESAR 3 Joint Undertaking under grant agreement No 101114762 under European Union's Horizon Europe research and innovation programme.

Table of contents

List of figures

List of tables

1 Introduction

1.1 Purpose of the document

Deliverable D3.1 (*Certification methods and automation: benefits, issues, and challenges*) studied the usefulness of *currently used* aviation certification methods for advanced automation and artificial intelligence (AI). The aim of this **Deliverable 3.2 (Innovative approaches to approval and certification)** is to consider *innovative* approaches to certification, and to explore their suitability for application to advanced automation and AI-powered technologies. This includes benefits, challenges and issues that may emerge for the certification of non-deterministic systems based on AI and ML (machine learning). The result will be taken into account in WP4 for the design of the new holistic certification approach for systems based on advanced automation.

This deliverable feeds into HUCAN task T4.3 (that task requires from D3.2 an overview of innovative approaches for certification and automation, and an overview of their benefits, issues, and challenges), and into WP5 (which validates the WP4 approach in order to ensure its usability, suitability and effectiveness for the specific aim it has, hence requires from D3.2 a list of issues that need improvement).

1.2 Scope of the document

The document focuses on innovative certification approaches in advanced automation and AI. The analysis requires a preliminary clarification of the concept of “advanced automation” and “AI” (as well as their differences) and of “certification approach”. Please refer to the Glossary in Chapter 7 for a more complete overview of terms and their definitions, collected from key literature sources.

1.2.1 Advanced Automation and Artificial Intelligence for Aviation

According to (EASA, 2023), *automation* is defined as ‘The use of control systems and information technologies reducing the need for human input, typically for repetitive tasks.’ *Advanced automation* is defined as ‘The use of a system that, under specified conditions, functions without human intervention.’ Advanced automation in aviation refers to the development of systems that can either replace humans for routine tasks or work alongside humans to augment their capabilities, particularly in tasks involving real-time processing of complex data (*e.g.,* airspace management). Advanced automation follows a series of key points:

- **Focus on non-trivial tasks.** Human expertise remains crucial for decision-making, judgement, and handling unforeseen situations. Automation frees humans from routine or lower-level tasks, allowing them to focus on higher-order thinking and problem-solving.
- **Dynamic and flexible processes.** Advanced automation allows for customization of complex processes on-the-fly. For example, airspace management can become more dynamic based on real-time data, optimising efficiency and safety.
- **Human-Automation Teaming**. This concept emphasises collaboration between humans and automated systems, ensuring humans remain "*in-the-loop*" and can supervise or intervene when necessary, a concept complimentary to Human-AI Teaming.

There is no harmonised definition for *Artificial Intelligence (AI)*, but many sources refer to it as a machine's ability to perform cognitive functions we usually associate with human minds. Even though some references seem to use advanced automation and AI interchangeably, they are not the same, and one is possible without the other. Artificial Intelligence solutions and approaches enhance human-automation relationships by enabling, accelerating, and supporting automation, including advanced forms. This makes the overall process more complex and multi-layered. Additionally, it opens the door to autonomous systems alongside existing automated ones. AI for automation follows the below presented key points:

- **Enabler and Accelerator.** AI, particularly Machine Learning (ML), acts as a powerful tool for advanced automation. By analysing data, AI can support decision-making through automation, leading to improved performance and safety.
- **Adaptive behaviours**. AI systems based on machine learning algorithms are capable of learning from and adapting to new data over time. This adaptive capability allows AI to improve its performance based on real-world feedback and changing conditions. Adaptive behaviours enable AI to respond to unforeseen events, continuously refine decision-making processes, and provide tailored solutions that evolve with the environment and operational demands.
- **Shifting Human Roles.** AI has the potential to achieve high levels of automation, which could transform the role of humans in the workplace. They may transition from solo controllers to "augmented controllers," "co-workers," or "supervisors" working alongside intelligent automation systems.
- **Data Sharing and Network Optimization.** Advanced automation facilitates a shift from local optimisation to a network-wide perspective through data sharing. This is what could allow for a more holistic approach to air traffic management, potentially leading to significant efficiency gains.
- **Full Integration of Crewed and Uncrewed Systems.** As automation advances, the seamless integration of crewed and uncrewed (e.g., drones) aircraft within the airspace becomes a possibility.
- **Intuitive Interfaces.** To keep humans effectively "*in-the-loop*," advanced automation systems require user-friendly interfaces that clearly communicate information and system status. AI solutions, especially considering recent advancements in Natural Language Processing (NLP), could aid in doing so.

While advanced automation and AI offer immense potential for aviation safety and efficiency, several crucial considerations remain, with more complications and possible problems being introduced:

- **Explainability**. Understanding how AI systems reach decisions is critical for building trust and ensuring responsible use. This appears particularly crucial for the domain of ATM and aviation in general, especially for considerations related to the "*human-in-the-loop*" doctrine.
- **Autonomous vs. Automated.** There's a distinction between truly *autonomous* systems (capable of independent decision-making) and *automated* systems (which follow pre-programmed rules). This opens the door to further complications relating to AI autonomy and independent decision-making, as mentioned below.
- **Unforeseen Events**. Advanced automation systems need to be robust enough to handle unforeseen situations that fall outside their programming.
- **Automation Bias.** Over-reliance on automation can lead to overlooking crucial information or ignoring warnings. Effective human oversight remains essential.

- **Trust and Awareness.** Building trust and ensuring human awareness of system capabilities is critical for successful human-automation teaming.
- **Black Box AI.** If AI systems become too complex and their decision-making processes opaque, it can hinder human oversight and troubleshooting.
- **Data Poisoning and Training.** The quality of training data used for AI systems is paramount. Data poisoning, that is, feeding the system with biased or incorrect data, can lead to flawed decision-making.

Advanced automation and AI hold tremendous promise for revolutionising aviation, enhancing safety, efficiency, and airspace management. However, careful consideration must be given to the human role, system explainability, and potential pitfalls to ensure a smooth and successful integration of these technologies.

S3JU has recently provided insights into contextualising different types of AI, aligning them according to various levels of automation. As depicted in the Figure 1-1, adapted from SESAR JU (2024), distinct AI categories can influence diverse human-machine interaction types, contingent upon the attained level of automation in specific cognitive tasks. At all Levels there is full automation for the activities of Perception and Analysis, but there are differences where the Decision-making, the Execution of the action, and the Authority of the human operator are concerned.

- At Level 1A (EASA), AI acts as "human augmentation" with “low automation” (Level 0, S3JU), where human operators retain full decision-making and execution responsibilities.
- At Level 1B (EASA), AI functions as "human assistance" with a focus on "decision support" (Level 1, S3JU) enabling humans to make informed decisions based on overviews of feasible options provided by the system.
- At Level 2A (EASA), AI facilitates "human-AI cooperation" as a "resolution support" system (Level 2, S3JU), where humans evaluate and refine solutions proposed by automation.
- At Level 2B (EASA), AI fosters "human-AI collaboration" at a "conditional automation" level (Level 3, S3JU), allowing humans to assign tasks to either the automation or themselves.
- At Level 3A (EASA), AI operates in a "safeguarded" or "confined" automation mode (Level 4, S3JU), functioning autonomously but supervised by humans upon request or when operating outside its designated domain.
- At Level 3B (EASA), AI operates fully autonomously without human supervision (Level 5, S3JU).

At the top levels, the level of automation may not be formally considered ‘Advanced’, but AI-powered technologies are still applicable. Therefore, in the context of this report, the term ‘Advanced automation and AI-powered technologies’ refers to any of the levels in Figure 1.

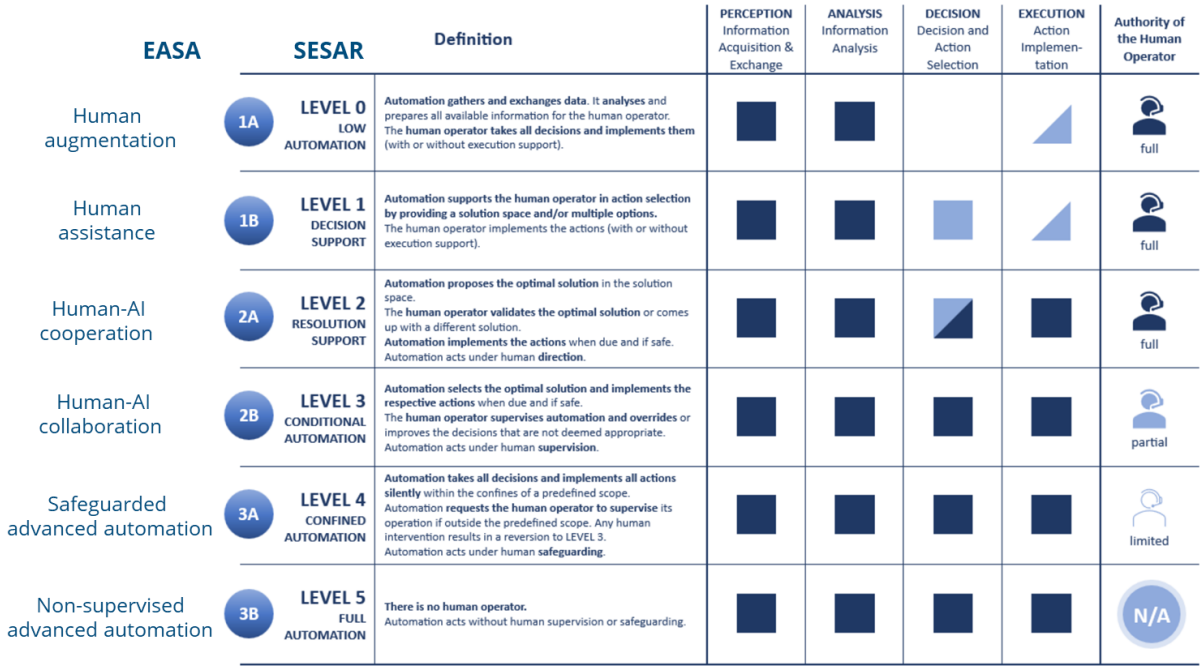

Figure 1. SESAR JU proposed new Levels of Automation Taxonomy and correspondence to EASA AI Levels, adapted from (SESAR JU, 2024)

1.2.2 Certification approach

Due to the variety of key points and complications described above, one of the key challenges with advanced automation and AI is to address their approval and certification. This deliverable aims to review innovative approaches for certification, and explore their suitability for application to advanced automation and AI-powered technologies. This section aims to define what we mean by 'certification approach'.

Based on the analysis carried out, different elements may constitute an "approach" to certification. We list them here:

- **Scope.** Since certification may refer to technology (hardware and/or software), systems, personnel, or organisations, and since the activities to be taken may be different for each, it is important to be clear about the applicable scope of the certification approach. In the domain of automation and AI-based technologies, this includes defining what types of technologies may be covered and whether different levels of automation/autonomy pertaining to the technology are taken into consideration. The scope should also make clear to what extent human factors and human-AI teaming can be addressed.
- **Goals.** In choosing and developing a certification approach, the organisations involved must choose which goals to pursue, *that is*, which values, rights and stances must be realised through the implementation of the certification process. *For example*, in the domain of aviation, safety is a critical and principal goal. Choosing goals mirrors the priorities and hierarchy of values present in a given domain, to be realised through the certification approach.

- **Standards.** Given a specific target, a standard appears as a requirement, threshold, method or process capable of ensuring that it is met. Standards may refer to different targets, such as product characteristics, procedures, management measures and performance indicators, while also potentially including evaluation criteria for their validation. Moreover, metrics and guidelines aiding actors in choosing a certain set of suitable standards under their accountability should be included in the process (see the following sections), as the choice between different standards is not neutral.
- **Actors.** Certification is a multi-agent and complex process which presupposes the cooperation of different bodies, both private and public, such as producers, state authorities, European organisations and qualified third-party organisations. The relationship between all the relevant actors, as well as their powers, duties and rights towards one another must be taken into account in the certification process, especially in light of balancing the impartiality and thirdness (independence from industry or government) of the overall approach. *For example*, a given approach could identify in producers the actor addressee of duties, and in governmental bodies the holders of enforcement powers, while having third-party assessment organisations involved.
- **Certification Process.** The process of certification itself concerns a variety of steps, measures and duties, such as data collection, record-keeping, document tracking, testing, sandboxing and both *pre*- and *post*-market monitoring. The choice of which to include is a principal priority in developing a certification approach, and heavily characterises the process by communicating technical, ethical and socio-juridical stances to all actors. *For example*, pushing for binding categorisation of products and static standards, both chosen *ex-ante*, could improve safety and certainty while hindering innovation, communicating a clear hierarchy of goals.
- **Enforcement.** All of the above must be framed into an enforcement framework capable of ensuring that the duties and obligations arising from the approach are upheld, following the concept of liability as a norm of conduct. Moreover, under a lens of accountability, an approach could leave room for choice regarding the standards, evaluation criteria, management and organisational measures to adopt and so forth, pursuing a stance similar to the one used in the **GDPR** architecture. Finally, enforcement in the strict sense must be realised by including administrative powers of control, intervention and fining, addressed to specific and impartial actors.
- **Documentation**. Emerges as a critical aspect tied to all key phases and choices of a certification approach, and it acquires more relevance the more discretion is left to actors and stakeholders in the context of accountability. Choosing which documents must be kept, their technical depth and the governance of data relating to certification is a constituting element and essential set of choices tied to certification approaches.
- **Harmonisation.** Once an approach has been developed, it should be *past* and *future*-proof. By this we mean it should take into account existing certification measures and be able to ensure a smooth and fluid transition from the old approach to the new, while additionally being able to withstand innovation and change in the domain of reference, reducing friction in either sense to a minimum. Drawing from a variety of domains and certification approaches (as seen in Section 2.11) could support harmonisation by offering varied and different insights.

Given the above, we will employ a general working definition of a "**certification approach**", i.e. the combination of a multitude of legal, technical and social processes capable of setting adequate certification standards and ensuring that they are met, with the goal of upholding socio-technical, legal and ethical values, such as safety, robustness, privacy, human agency and explainability.

1.3 Structure of the document

This document is structured as follows:

- Chapter 2 gives an overview of emerging and innovative approaches to the approval and certification process of automated and AI-based technology.
- Chapter 3 develops criteria and analyses the innovative approaches against the criteria identified above.
- Chapter 4 discusses the suitability of the approaches for the aviation domain.
- Chapter 5 provides references to literature and other documents.
- Chapter 6 provides a list of acronyms used.
- Chapter 7 provides an extensive glossary of terms, and also explains why it is included in this document.

2 Emerging Certification Approaches for Advanced Automation, including Aviation

2.1 Introduction

The aim of this chapter is to give an overview of the emerging and innovative **approaches** to the approval and certification process of automated and AI-based technology. We will adopt the notion of "certification approach" presented in Chapter 1, with the caveat that not all elements addressed under the definition, are necessarily present in the below overview.

Figure 2-1 provides an overview of regulations, Acceptable Means of Compliance (**AMC**) and Guidance Material (**GM**), industry standards, supporting methods of AMC & GM and standards, and research roadmaps for the development of supporting methods. The figure is based on the rulemaking concept for AI as conceived in the EASA AI Roadmap (EASA, 2023), which includes foreseen feedback for the development of requirements to the EU AI Act (EU, 2021) as well as to domain-specific regulations.

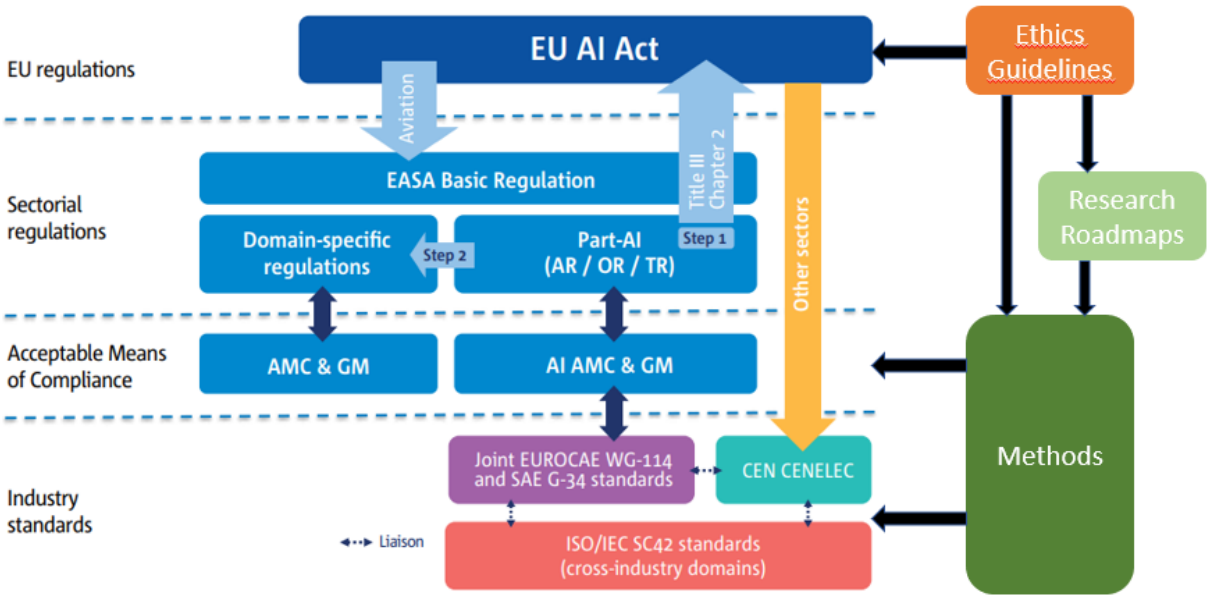

Figure 2. Anticipated regulatory structure for EASA AI, adapted from (EASA, 2023)

The innovative approaches collected in this Chapter include ethics guidelines, the EU AI Act, various research roadmaps, and various methods that have been developed in (emerging) standards. In particular, the remainder of this chapter is structured as follows:

- Section 2.2 describes ethics guidelines for trustworthy AI as developed by a High-level Expert Group for the European Commission. These guidelines provided input for the EU AI Act and for the EASA AI roadmap and developing guidance material.
- Section 2.3 describes the EU AI Act.
- Section 2.4 describes the EASA AI roadmap and the developing guidance material for machine learning applications (developing emerging methods for AMC & GM) for future acceptable means of compliance).

- Section 2.5 describes an extensive research roadmap of the National Research Council (2014) for increasingly autonomous operations in civil aviation.
- Section 2.6 describes the development of a roadmap and methods for AI safety assurance by the FAA.
- Section 2.7 describes a roadmap and methods for supporting suitable interaction between humans and AI-based systems.
- Section 2.8 describes the development of standards for AI certification by EUROCAE WG-114 and SAE G-34.
- Section 2.9 describes steps towards the development of methods for testing and evaluating AI-enabled systems by the US Air Force.
- Section 2.10 describes cross-industry standards of ISO/IEC for software and AI.
- Section 2.11 describes cross-industry standards of IEEE for AI and autonomous systems.
- Section 2.12 describes cross-industry safety assurance objectives for autonomous systems, as developed by the Safety Critical Systems Club.
- Section 2.13 describes methods for process-based certification.
- Section 2.14 describes methods for certification in road transport.

2.2 European Commission Ethics guidelines for Trustworthy AI

The European Commission set up a High-level Expert Group on AI for the development of ethics guidelines for **Trustworthy AI** (*High-level Expert Group on AI, 2019*). According to this group, Trustworthy AI has three components, which should be met throughout the system's entire life cycle:

1. It should be **lawful**, complying with all applicable laws and regulations;
2. It should be **ethical**, ensuring adherence to ethical principles and values; and
3. It should be **robust**, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.

Four ethical principles are distinguished for AI systems:

1. **Respect for human autonomy.** Humans interacting with AI systems must be able to keep full and effective self-determination over themselves. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice.
2. **Prevention of harm.** AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity. AI systems and the environments in which they operate must be safe and secure.
3. **Fairness.** The development, deployment and use of AI systems must be fair. It should ensure equal and just distribution of both benefits and costs, and ensure that individuals and groups are free from unfair bias, discrimination and stigmatisation. In support, the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable.
4. **Explicability.** Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected.

These ethical principles are translated in the following requirements to achieve Trustworthy AI:

1. **Human agency and oversight.** AI systems should support human autonomy and decision-making. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights, and allow for human oversight. Oversight may be achieved through governance mechanisms such as human-in-the-loop (human intervention in every decision cycle of the system), human-on-the-loop (monitoring the system's operation), or human-in-command (overseeing the overall activity of the AI system and the ability to decide when and how to use the system in any particular situation). Oversight mechanisms can be required to support other safety and control measures to varying degrees, depending on the AI system's application area and potential risk. All other things being equal, the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required.
2. **Technical robustness and safety.** Technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. This should also apply to potential changes in their operating environment or the presence of other agents (human and artificial) that may interact with the system in an adversarial manner. The physical and mental integrity of humans should be ensured. AI systems must be secure and resilient to attack. AI systems should have safeguards that enable a fall-back plan in case of problems. AI systems must provide accurate results and when occasional inaccurate predictions cannot be avoided, it is important that the system can indicate how likely these errors are. Results of AI systems must be reproducible (effectively ruling out non-deterministic AI).
3. **Privacy and governance.** AI systems must guarantee privacy and data protection throughout a system's entire lifecycle. The quality and integrity of the data used for training of AI systems must be assured, such that it does not contain biases, inaccuracies or errors.
4. **Transparency.** The data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability and an increase in transparency. Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available. AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system.
5. **Diversity, non-discrimination and fairness**. Unfair biases should be avoided, as data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historical bias, incompleteness and bad governance models. The way in which AI systems are developed (e.g. algorithms' programming) may also suffer from unfair bias. Particularly in business-to-consumer domains, systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Stakeholder participation is needed throughout the life cycle of an AI system.
6. **Societal and environmental well-being**. In line with the principles of fairness and prevention of harm, the broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system's life cycle. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI

solutions addressing areas of global concern, such as for instance the Sustainable Development Goals. Ideally, AI systems should be used to benefit all human beings, including future generations.

7. **Accountability**. Mechanisms must be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use. This includes auditability of the algorithms, data and design processes. The ability to report on actions or decisions that contribute to the system outcome, and to respond to the consequences of such an outcome, must be ensured. Trade-offs should be addressed in a rational and methodological manner, entailing that relevant interests and values implicated by the AI system should be identified and that, if a conflict arises, trade-offs should be explicitly acknowledged and evaluated in terms of their risk to ethical principles, including fundamental rights.

To implement these requirements, both technical and non-technical methods could be employed. These encompass all stages of an AI system's life cycle. An evaluation of the methods employed to implement the requirements, as well as reporting and justifying changes to the implementation processes, should occur on an ongoing basis. In particular, technical methods towards Trustworthy AI include the following:

- **Architectures.** Architectures for Trustworthy AI should be anchored by procedures reflecting the above requirements, e.g. by "white list" rules (behaviours or states) that the system should always follow and "*black list*" restrictions on behaviours or states that the system should never transgress. Monitoring of the system's compliance with these restrictions during operations may be achieved by a separate process.
- **Ethics and X-by-design.** Methods to ensure values-by-design (e.g. *privacy-by-design*) provide precise and explicit links between the abstract principles which the system is required to respect and the specific implementation decisions. To earn trust, AI needs to be secure in its processes, data and outcomes, and should be designed to be robust to adversarial data and attacks.
- **Explanation methods.** A whole field of research, Explainable AI (XAI) tries to address this issue to better understand the system's underlying mechanisms and find solutions.
- **Testing and validating.** Due to the non-deterministic and context-specific nature of AI systems, traditional testing is not enough. Failures of the concepts and representations used by the system may only manifest when a programme is applied to sufficiently realistic data. Consequently, to verify and validate processing of data, the underlying model must be carefully monitored during both training and deployment for its stability, robustness and operation within well-understood and predictable bounds. It must be ensured that the outcome of the planning process is consistent with the input, and that the decisions are made in a way allowing validation of the underlying process.
- **Quality of service indicators.** These indicators could include measures to evaluate the testing and training of algorithms as well as traditional software metrics of functionality, performance, usability, reliability, security and maintainability.

Continuing, non-technical methods towards Trustworthy AI include the following:

1. Regulation,
2. Codes of conduct,
3. Standardisation,

4. Certification,
5. Accountability via governance frameworks,
6. Education and awareness to foster an ethical mind-set,
7. Stakeholder participation and social dialogue,
8. Diversity and inclusive design teams.

2.3 The AI Act

The **Artificial Intelligence Act (AIA)** (EU, 2021) is the first-ever legal framework specific to the challenges of AI development, deployment, and use across the European Union. Proposed in April 2021 by the European Commission, it aims to foster trustworthy, explainable and human-centric AI within the EU, ensuring the creation of systems as tools capable of upholding safety, health and fundamental rights, while mitigating the risks posed by state-of-the-art AI models, such as generative and foundation systems. In July 2024, the text was published in the Official Journal of the EU and entered into force on August 1, 2024[3]. It will gradually become enforceable (e.g., rules on prohibited practices will apply after 6 months) and will be fully applicable approximately in August 2026.

The AI Act follows a **risk-based approach** by categorising AI systems according to their risks to safety, health, and fundamental rights, and establishing a specific legal regime for each class. Four categories of risk, in particular, are identified: **1)** AI systems with unacceptable risks; **2)** high-risk AI systems; **3)** low-risk AI systems; **4)** no-risk AI systems.

Certain AI systems that present **unacceptable risks** are prohibited. This category includes sensitive AI applications, which, according to EU institutions, are most blatantly at odds with European core values. Generally, this category may include AI systems exploiting vulnerabilities, causing harm, or infringing on privacy through manipulative behaviours, social scoring, biometric categorisation, untargeted facial scraping, and emotion recognition in sensitive environments[4].

In the **high-risk category**, a wide range of AI systems are permitted but subject to a set of technical requirements for access to the EU market. This is the category to which the majority of the AI Act provisions are addressed. We shall see in Section 2.3.2 that these provisions are indirectly relevant for certain AI systems in the aviation sector. Therefore, in this section, we shall particularly focus on high-risk AI systems.

AI systems that present a **limited risk** are regulated by Article 50 of the AI Act and subject only to transparency measures, which basically translate into the duty of the provider or deployers to inform the person exposed or affected by the systems. Only four types of systems are included in this

[3] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), OJ L, 2024/1689, 12.7.2024.

[4] See: AI Act, Article 5. Generally, these include: AI systems manipulating behaviours or exploiting vulnerable people, causing physical or psychological harms; AI systems used by or on behalf of public authorities for social scoring; biometric categorisation systems that use sensitive data; untargeted scraping of facial images from the internet or CCTV footage to create facial recognition databases; emotion recognition in the workplace and educational institutions.

category: systems intended to interact with natural persons (e.g. chatbots), emotion recognition and biometric categorisation systems, content generative-AI systems and deep fakes.

Finally, AI systems that are not included in one of the previous categories are implicitly considered as **no-risk**, thus are not covered by any requirements. For those systems, Article 95 only establishes that the Commission, through the AI Office, and the Member States shall encourage the drafting of codes of conduct, which are supposed to lead AI providers to voluntarily apply mandatory requirements set for high-risk AI systems.

2.3.1 Definition of "AI"

The Regulation portrays a wide **definition of AI**, intended as "a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (Article 3(1)). This definition replaced the technology-driven definition originally provided by the Commission[5] and aligned with the definition provided by the OECD[6].

Recital 12 of the AI Act clarifies that "AI systems are designed to operate with varying levels of autonomy, meaning that they have some degree of independence of actions from human involvement and of capabilities to operate without human intervention". Apart from that, the AI Act does not clarify the minimum degree of independence relevant for a system to be classified as an AI system, nor does it connect to specific capability functions, blurring the lines between the concept of "AI" and "automation", including "advanced automation".

The OECD's explanation of "autonomy" is more specific, as it should be understood as "the degree to which a system can learn or act without human involvement following the delegation of autonomy and process automation by humans. Human supervision can occur at any stage of an AI system's life cycle, such as during AI system design, data collection and processing, development, verification, validation, deployment, or operation and monitoring. Some AI systems can generate outputs without specific instructions from a human". Autonomy is, therefore, reconnected to the system's capability of learning and/or acting in the environment without or with limited human involvement. Such autonomy may derive from an act of human autonomy delegation but also "process automation". In this case,

[5] "'Artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with". Annex I included the following techniques and approaches: "(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning; (b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
(c) Statistical approaches, Bayesian estimation, search and optimization methods".

[6] OECD (2024), https://oecd.ai/en/wonk/ai-system-definition-update. The OECD's definition reads as following: "An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment".

too, the definition seems to be broad enough to also include automation, at least advanced automation levels.

2.3.2 High-risk systems: Qualification and Relevance for the Aviation Sector

The Regulation mainly addresses so-called "**high-risk systems**" by mandating technical requirements and distributing responsibilities for their implementation along the value chain, especially among providers.

According to Article 6, an AI system is considered high-risk if:

1. it is intended to be used as a **safety component of a product**, or the AI system is itself a **product**, covered by the Union harmonisation legislation listed in **Annex I** (art. 6(1)(a)) and, pursuant that legislation, is required to undergo a **third-party conformity assessment**, with a view to the placing on the market or the putting into service (art. 6(1)(b)); or
2. it is referred to in **Annex III** (art. 6(2)), covering the following high-risk areas: biometric identification and categorisation of individuals; management and operation of critical infrastructures; education and vocational training; employment, labour management and access to self-employment; access to and enjoyment of essential private services and public services and benefits; law enforcement; management of migration, asylum and border control; administration of justice and democratic processes.

The first of the two conditions bears great relevance to the **aviation sector.** In fact, Annex I is further divided into two sections:

- **Section A** lists a number of product safety pieces of legislation, commonly referred to under the so-called "New Legislative Framework". The latter include, for example, Directive 2009/48/EC on the safety of toys and Regulation (EU) 2017/745 on medical devices.
- **Section B** lists other Union harmonisation pieces of legislation, which represent the "old approach" to product legislation. These include, among others, two important regulations in the aviation sector, namely (i) **Regulation (EC) No 300/2008** common rules in the field of civil aviation security and (ii) **Regulation (EU) 2018/1139** of the European Parliament and of the Council of 4 July 2018 on common rules in the field of civil aviation and establishing a European Union Aviation Safety Agency (so-called "**Basic Regulation**" (BR))

The distinction between Sections A and B in Annex I is relevant for the scope of application of the AI Act. In fact, **Article 2(2)** establishes "*for AI systems classified as high-risk AI systems in accordance with Article 6(1) and (2) related to products covered by the Union harmonisation legislation listed in Section B of Annex I, only Article 112 applies*. *Article 57 applies only in so far as the requirements for high-risk AI systems under this Regulation have been integrated in that Union harmonisation legislation*." Article 112 refers to the periodic evaluation and review of the AI Act by the Commission, which also includes the possibility of adopting implementing or delegated acts concerning sectoral Union harmonisation legislation listed in Section B of Annex I. Article 57 refers to the possibility of establishing an AI

regulatory sandbox.[7] Therefore, Article 2(2) implies that, despite being classified as high-risk, AI systems adopted under the Basic Regulation are not required to comply with essential requirements established in the AI Act.

Article 108, however, introduces some important amendments to the Basic Regulation, which requires the European Commission to take into consideration the essential requirements of high-risk AI systems, when:

- adopting implementing and delegated acts regarding **airworthiness** (Articles 17 and 19 of the BR as amended by the AIA);
- adopting implementing and delegated acts as regards **ATM/ANS providers and organisations** involved in the design, production or maintenance of ATM/ANS systems and ATM/ANS constituents (Article 43 and 47 of the BR as amended by the AIA);
- adopting implementing and delegated acts regarding **unmanned aircraft** (Articles 57-58 of the BR as amended by the AIA).

To sum-up, the AI Act is therefore relevant to the aviation sector in that:

- It classifies AI systems, which are adopted within the scope of application of the Basic Regulation and must undergo a third-party conformity assessment, as "high-risk AI systems" for the purpose of AI regulation.
- It excludes AI systems from its scope of direct application, except for the provision giving the power to the Commission to adopt implementing and delegated acts for certain high-risk systems.
- It amends the Basic Regulation by requiring the Commission, when adopting, implementing or delegating acts pursuant the different areas of application of the BR, to take into account the mandatory requirements for high-risk AI systems laid down in the AI Act.

We can conclude that essential requirements established in the AI Act for high-risk systems will be relevant in the future for the purpose of adapting the certification framework established in the BR to integrate AI and advanced automation in the aviation sector.

2.3.3 High-risk systems: Essential Requirements

Articles 9 to 15 of the AI Act set out essential requirements for high-risk AI systems, which ensure their compliance with the ethos of the regulation, taking into account their intended purposes as well as the generally acknowledged state of the art on AI and AI-related technologies. In particular, the essential requirements appear as follows:

- **Risk management system (Article 9)**. The AI system has to be complemented by a risk-management system (Article 11), which should allow the provider of the high-risk system to assess the specific risks of the system during the whole system's lifecycle and adopt relevant risk management measures. Such measures should take into consideration the generally

[7] This is defined in Article 3(55) as "a controlled framework set up by a competent authority which offers providers or prospective providers of AI systems the possibility to develop, train, validate and test, where appropriate in real-world conditions, an innovative AI system, pursuant to a sandbox plan for a limited time under regulatory supervision".

acknowledged state of the art, including as reflected in relevant harmonised standards or common specifications. In particular, measures should be adopted so that the overall residual risk is judged acceptable.

- **Data and data governance (Article 10)**. The AI system must rely on appropriate data governance and management practices, which ensure the data is high-quality (Article 10). This requirement mainly refers to training, validation, and testing datasets when data-driven AI models are developed. Minimum requirements for data management are related to: (i) data collection processes, the origin of data and the original purpose of the collection; (ii) relevant pre-processing operations (e.g., annotation, cleaning, etc.); (iii) the assumption with respect to the information the data are supposed to represent; (iv) bias detection, especially when they might impact negatively on fundamental rights. Datasets should be sufficiently representative and, to the best extent possible, free of errors and complete in view of the system's intended purpose and the context of use.
- **Technical documentation (Article 11)**. The AI system must be transparent. Transparency is assessed at three levels, the first being the production of technical documentation (Article 11). Technical documentation is key in the accountability mechanism as it should demonstrate that the system has complied with the essential requirements set out in Articles 9-15. Moreover, it provides the conformity assessment bodies and national supervisory authorities with all the necessary information to assess the compliance of the AI system with those requirements. The content of the technical documentation is contained in Annex IV of the Proposal and includes: (i) a general description of the AI system including; (ii) a detailed description of the elements of the AI system and of the process for its development; (iii) detailed information about the monitoring, functioning and control of the AI system; (iv) a detailed description of the risk management system in accordance with Article 9; (v) a description of any change made to the system through its lifecycle; (vi) a list of the harmonised standards applied in full or in part or, where no such standards have been applied, a detailed description of the solutions adopted to meet the essential requirements; (vii) a copy of the EU declaration of conformity; (viii) a detailed description of the system in place to evaluate the AI system performance in the post-market phase.
- **Record-keeping (Article 12)**. The second level of transparency relates to traceability. The provider shall design AI systems so that all operations can be automatically recorded over the system's lifetime. Traceability requirements are pivotal in monitoring the performance of the AI system, especially after it is put onto the market. This requirement connects to the provider's obligation set out in Article 20 on automatically generated logs.
- **Transparency and provision of information to deployers (Article 13)**. The last level of transparency refers to the interpretability requirement. Interpretability is intended as a sufficient level of transparency in the way the system operates to allow deployers to understand the system's output and use it appropriately. For this to happen, the provider must accompany the provision of the AI system with instructions of use that is understandable and legible to deployers. Such information includes the identity of the provider, the intended purpose of the system and its limitations, the technical capabilities to provide an explanation, the level of accuracy, robustness, cybersecurity, human oversight measures, any necessary maintenance, etc.

- **Human oversight (Article 14)**. The AI system should enable human oversight through technical or organisational measures (Article 14). Technical measures include the internal design of the system and the use of appropriate human-machine interfaces. Organisational measures mainly refer to the appropriate competence and training of natural persons to whom oversight is assigned to during the deployment. Such measures should allow the human controller to properly understand the relevant capacities and limitations of the system and monitor its operation, to avoid over-reliance on the system's output (sc. automation bias), to correctly interpret the output, to override the system's decision, to intervene on the operation, and, in certain cases, impede that a decision is taken based on the output of the system.
- **Accuracy, robustness and cybersecurity (Article 15)**. Finally, AI systems shall ensure an appropriate level of accuracy, robustness, and cybersecurity. The relevant accuracy metrics shall be declared in the accompanying instructions of use. The systems shall be as resilient as regards possible errors, faults, or inconsistencies, and when based on reinforcement learning, it should avoid "feedback loops". Cybersecurity entails appropriate measures to prevent data poisoning, model poisoning, model evasion, and confidentiality attacks.

2.3.4 High-risk systems: Fundamental Rights Impact Assessment (FRIA)

The AI Act grounds the usage and deployment of high-risk systems, a fundamental provision in the regulation, on the Fundamental Rights Impact Assessment (**FRIA**), placing specific requirements on the deployers and operators of AI models of the highest risk tier.

The **FRIA** goes beyond technical compliance measures and aims to identify potential harms to fundamental rights caused by high-risk AI. Since AI providers might not foresee all deployment scenarios and biases within the system, the FRIA acts as a justification and accountability tool. It forces organisations to carefully consider the reasons behind deploying the AI, where it will be used, and how it will function. However, conducting a FRIA presents a challenge, and not all deployers of high-risk AI systems will have the resources to fully assess the risks involved.

Article 27 of the AI Act follows all of the above by stating how prior to deploying a high-risk AI system referred to in Article 6(2) into use, deployers and operators must evaluate the impact on fundamental rights such a system could bring about. The assessment, according to the letter of **Article 27**, consists of:

- A description of the deployer's processes in which the high-risk AI system will be used in line with its intended purpose;
- A description of the period of time within which, and the frequency with which, each high-risk AI system is intended to be used;
- The categories of natural persons and groups likely to be affected by its use in the specific context;
- The specific risks of harm likely to have an impact on the categories of persons or groups of persons identified pursuant point (c) of this paragraph, taking into account the information given by the provider pursuant to Article 13;
- A description of the implementation of human oversight measures, according to the instructions for use;
- The measures to be taken where those risks materialise, including the arrangements for internal governance and complaint mechanisms.

Provisions ensure the relevance of the FRIA to possible changes in the nature and usage of the AI system, laying down obligations on deployers to take all necessary measures to update relevant information in case any of the precedent FRIA elements are no longer up to date.

All of the above is notified to the Market Surveillance authority by the addressee of the FRIA obligation, which the Act states as "*bodies governed by public law, or are private entities providing public services, and deployers high-risk AI systems referred to in points 5 (b) and (c) of Annex III*", limiting to specific stakeholders the application of the FRIA.

By understanding the FRIA requirement and its complexities, deployers can begin to prepare for its implementation and ensure the responsible use of high-risk AI systems within the framework of the EU AI Act.

2.3.5 Conformity Assessment, Certification and Standards

High-risk AI systems are presumed to meet the AI Act requirements if they comply with harmonised standards that are published in the Official Journal of the European Union. These standards are being developed upon request by the Commission[8], which will ensure they are clear, consistent with existing EU legislation, and effectively guarantee compliance with the Act.

The Commission will also request standards for reporting and documentation processes. These processes aim to improve the resource efficiency of AI systems throughout their entire lifecycle, as well as coordinate with record-keeping provisions. This includes standards that specifically target reducing a high-risk AI system's consumption of energy and other resources. Additionally, standards for the energy-efficient development of general-purpose AI models will be requested. To ensure these standards are well-rounded, the Commission will consult with the Board, relevant stakeholders, and an advisory forum before issuing a formal standardisation request to European standardisation organisations (ESOs). The request issued ensures AI systems or models placed on the market or put into service within the Union meet the relevant requirements laid out in the EU AI Act.

The Commission will hold the European standardisation organisations accountable, requesting them to provide evidence that they've made their best efforts to fulfil these objectives. This aligns with Article 24 of Regulation (EU) No 1025/2012.

Lastly, it appears crucial that all above-cited stakeholders cooperate during the standardisation process. In this regard, participants are encouraged to promote investment, innovation, and the overall competitiveness and growth of the EU market. Additionally, they should contribute to strengthening global cooperation on standardisation and consider existing international standards in the field of AI, as long as these standards align with EU values, fundamental rights, and interests. Finally, the process should enhance multi-stakeholder governance, ensuring a balanced representation of interests and the effective participation of all relevant stakeholders, as outlined in Articles 5, 6, and 7 of Regulation (EU) No 1025/2012.

Continuing, the above-mentioned standards serve a key purpose in the conformity assessment procedure, and influence it directly. In fact, when a provider has applied harmonised standards referred to in Article 40, or common specifications referred to in Article 41, to prove the compliance

[8] Cite Decision of the Commission mandating ESO

of a high-risk AI system with the requirements set out in Chapter 2, the provider holds the right to choose a conformity procedure between the following options:

- **Internal Control (Annex VI):** A streamlined approach does not necessitate the involvement of a notified body.
- **Notified Body Assessment (Annex VII):** A more comprehensive option entails a notified body evaluating the quality management system and technical documentation.

Providers must nevertheless opt for the notified body assessment procedure if harmonised standards are absent, if they have not been applied wholly or partially, when common technical specifications are not used or when harmonised standards present limitations in critical portions.

Providers generally have the freedom to choose any notified body. However, an exception exists for specific systems, which include the following: law enforcement, immigration, asylum, or EU institutional applications; which all require the market surveillance authority to act as the notified body.

In the event substantial modifications to the AI model take place, a mechanism of Re-Assessment, similar to the updating of FRIA changing elements, is implemented according to the following:

- **Triggering a New Assessment**. Any significant alterations to a high-risk AI system necessitate a new conformity assessment, regardless of whether the modified system is intended for further distribution or remains in use by the current deployer.
- **Predetermined Changes as Exceptions**. Changes and performance updates to the AI system that were anticipated by the provider at the time of the initial conformity assessment and documented within the technical documentation (Annex IV, point 2(f)) are exempt from being considered substantial modifications.

On a final note, the Commission holds the authority to adopt delegated acts in accordance with Article 97 to update Annexes VI and VII to reflect advancements in technology. The Commission can also enact delegated acts (under Article 97) to broaden the use of the notified body assessment (Annex VII) to encompass high-risk AI systems currently under the internal control procedure (Annex VI). This decision will be based on factors like the effectiveness of internal control in mitigating risks and the accessibility of resources among notified bodies.

2.3.6 Enforcement and Monitoring System

The governance framework developed in the AI Act rests on two principal provisions, detailing norms for post-market monitoring and sharing information about incidents regarding the deployment of AI models.

On the first end, providers must establish documented post-market monitoring systems tailored to the specific AI technology and the risks associated with the high-risk AI system. These measures include the following:

- **Data Collection and Analysis:** The monitoring system should actively gather, record, and analyse relevant data throughout the AI system's lifecycle. This data can come from deployers or other sources and should allow the provider to assess the system's ongoing compliance with the EU AI Act's requirements (Chapter III, Section 2).

- **Interaction Analysis (where applicable):** The monitoring process should include analysing how the AI system interacts with other AI systems, if relevant.
- **Data Privacy for Law Enforcement:** This obligation excludes collecting sensitive operational data from deployers who law enforcement authorities are.
- **Post-Market Monitoring Plan:** The monitoring system must be based on a documented post-market monitoring plan. This plan becomes part of the technical documentation required by Annex IV.
- **Template and Implementation Timeline:** The Commission will establish a template outlining the required elements for the post-market monitoring plan through an implementing act. This act will be adopted within six months before the EU AI Act comes into effect, following the examination procedure outlined in Article 98(2).

Finally, in the case of existing monitoring systems, providers of high-risk AI systems already covered by existing EU harmonisation legislation (listed in Annex I, Section A) can choose to integrate the necessary monitoring elements into their existing systems and plans, provided it achieves an equivalent level of protection and avoids duplication of effort. This option also applies to high-risk AI systems (Annex III, Point 5) placed on the market or used by financial institutions subject to relevant Union financial services law requirements regarding internal governance.

In relation to the sharing of critical information about incidents, providers of high-risk AI systems placed on the EU market are obliged to report any serious incident to the market surveillance authorities of the member state where the incident occurred. Reports must be submitted Immediately after establishing a causal link or reasonable likelihood of a link between the AI system and the incident, or no later than 15 days after the provider or deployer becomes aware of the incident, considering the incident's severity. Exceptions with shorter deadlines exist for widespread infringements or specific serious incidents (as defined in Article 3, point 44(b)), which must be reported immediately, or no later than 2 days after awareness.

To ensure timely reporting, incomplete initial reports can be submitted, followed by a complete report later. On this note, providers must promptly investigate serious incidents, assess the risks, and take corrective actions. They are required to cooperate with authorities during investigations and to avoid altering the AI system in a way that could hinder the evaluation of the incident's cause.

Finally, the AI Act implements a form of expedite, reduced reporting in the following instances:

- Providers subject to **existing EU legislation with equivalent reporting obligations** only need to report incidents involving manipulation, deception, or unfair bias (Article 3, point 44(c)). (Annex III)
- Providers of **high-risk AI systems** that are safety components of devices or devices themselves **covered by Regulations (EU) 2017/745 and (EU) 2017/746** also only need to report manipulation, deception, or unfair bias incidents (Article 3, point 44(c)). They report to the national competent authority designated for such incidents in the member state where the incident occurred.

Concluding, the national competent authorities must notify the Commission of any serious incident, regardless of any actions taken, in accordance with Regulation (EU) 2019/1020.

2.4 EASA AI roadmap and guidance for ML applications

2.4.1 AI roadmap

EASA published an AI Roadmap (EASA, 2023) to discuss the implication of AI on the aviation sector and identify high-level objectives to be met. It builds upon the EU ethics guidelines for trustworthy AI *and refers to the AI Act as a relevant benchmark for compliance*. The objectives of the roadmap are to develop a human-centric AI trustworthiness framework, to make EASA a leading certification oversight authority for AI, to support European aviation leadership in AI, to contribute to an efficient European AI research agenda, and finally, to actively support EU AI strategies and initiatives.

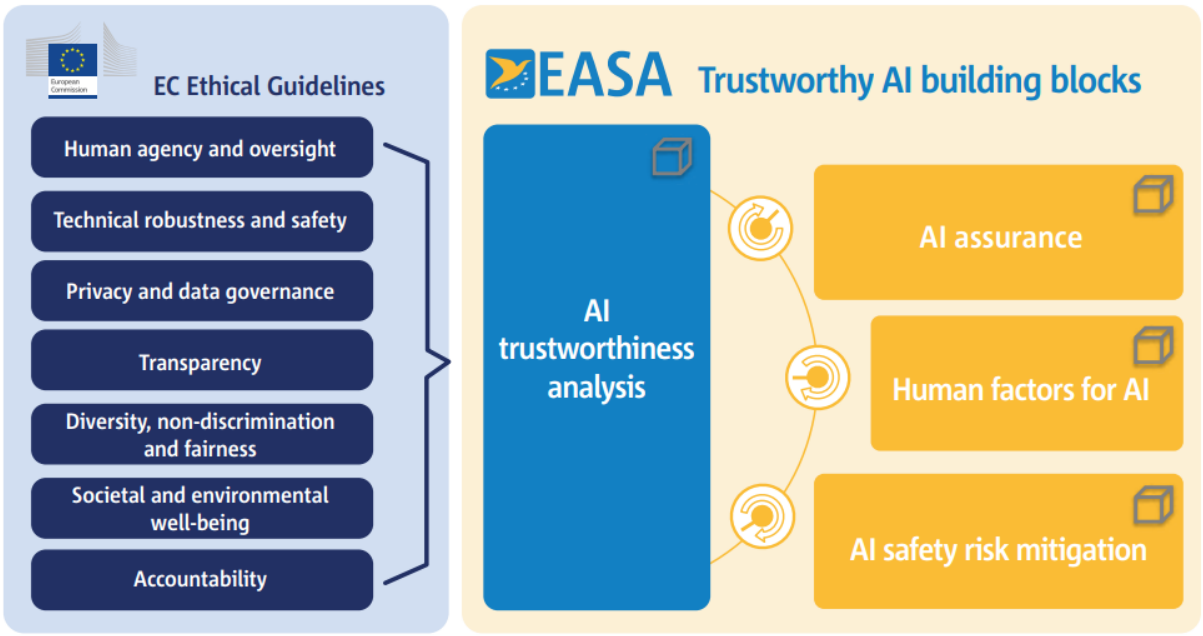

Figure 3. EASA AI trustworthiness building blocks, from (EASA, 2023)

The main building blocks towards AI trustworthiness in aviation are highlighted in Figure 2-2 (EASA, 2023, 2024).

1. **AI trustworthiness analysis.** The trustworthiness analysis building block encompasses different assessments, including ethical aspects, safety and security. Three levels are considered for human oversight over AI applications (see Figure 2-3): **(1)** assistance to humans, including **1a)** human augmentation, and **1b)** human cognitive assistance in decision-making and action selection; **(2)** human-AI teaming, including **2a)** human and AI-based system cooperation, and **2b)** human and AI-based collaboration; **(3)** advanced automation, including **3a)** the AI-based system performs decisions and actions that are overridable by the human, and **3b)** the AI-based system performs non-overridable decisions and actions (e.g. to support safety upon loss of human oversight).
2. **AI assurance.** Firstly, this includes learning assurance, which covers the paradigm shift from programming to learning, as the existing development assurance methods are not adapted to cover learning processes specific to AI/ML. Secondly, it includes the development of post-hoc explainability approaches to provide human users (e.g. *developers, auditors*) with

understandable, reliable and relevant information with the appropriate level of detail on how an AI/ML application produces its results. Thirdly, it includes data recording capabilities for continuous monitoring of the safety of an AI-based system, and for incident or accident investigation.

3. **Human factors for AI**. This concerns the specific human factors needs that are linked with the introduction of AI. Among other aspects, AI operational explainability deals with the capability to provide the humans end users with understandable, reliable and relevant information with the appropriate level of detail and with appropriate timing on how an AI/ML application produces its results. It also includes human-AI teaming to ensure adequate cooperation or collaboration between human end users and AI-based systems to achieve certain goals
4. **AI safety risk mitigation**. AI safety risk mitigation is based on the anticipation that the '*AI black box*' may not always be opened to a sufficient extent and that the associated residual risk may need to be addressed to deal with the inherent uncertainty of AI.

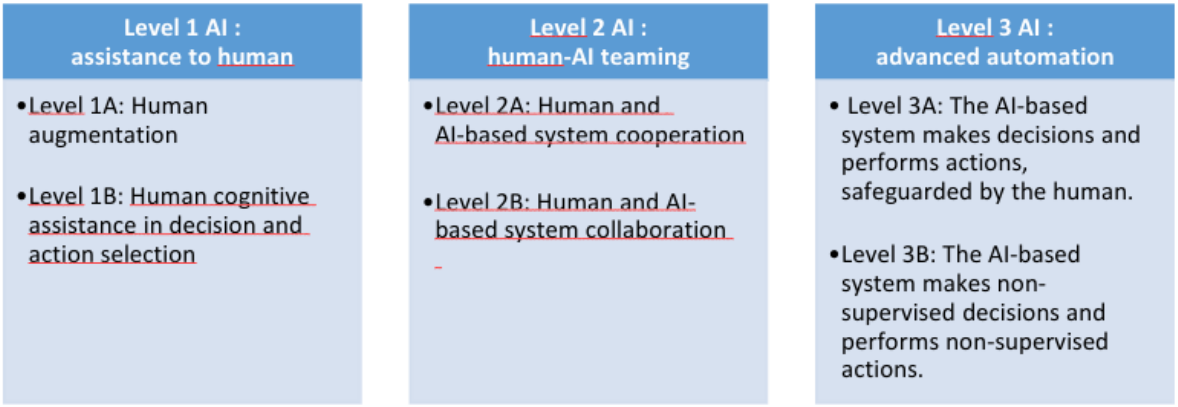

Figure 4. Classification of AI applications (EASA, 2024)

Guidance for Level 1 & 2 machine learning applications (assistance to human, and human-AI teaming) has been published in (EASA, 2024). It covers supervised and unsupervised offline learning, but excludes reinforcement learning and online learning processes. It specifies objectives and anticipated means of compliance (MOC) for these objectives along the AI trustworthy AI building blocks.

2.4.2 AI trustworthiness analysis

The AI trustworthiness analysis blocks include the characterisation of the AI application and assessment of safety, security and ethics.

The following elements need to be defined to characterise an AI application:

- Identifying end users intended to interact with the AI-based system, the associated high-level tasks and the AI-based system definition.
- Defining and documenting the ConOps for the AI-based system, including the task allocation between the end user(s) and the AI-based system. It includes a description of the operational domain, which specifies the conditions under which the AI-based system is intended to function.

- Identifying, proposing a break-down of the high-level functions into sub-functions, allocating the sub-functions to the subsystems, and AI/ML constituents.
- Classification of the AI application, using the EASA taxonomy: Level 1A (human augmentation) to Level 3B (non-supervised advanced automation).

Two types of safety assessment are applied: an initial safety assessment, which is achieved during the development phase, and a continuous safety assessment, which is applied in the operational phase. The safety assessments build on existing aviation safety methodologies. Elements of an initial safety assessment include:

- A functional hazard assessment (FHA) for the system level functions in the ConOps, including analysis of failure conditions of AI-based systems.
- Allocation of assurance levels (e.g. DAL, SWAL) for the system functions, which follow from the likelihood of the failure condition and the severity of the consequences. These assurance levels determine the extent to which objectives in (EASA, 2024) should be satisfied (higher assurance levels require more stringent satisfaction).
- Definition of AI/ML performance metrics in support of verification of potential quantitative safety requirements.
- Analysis and mitigation of the effect of the exposure to input data outside of the operational domain of the AI-based system. Potential mitigation includes monitoring of input data and enabling other systems to ensure safe operation.
- Identification, assessment and mitigation of sources of (aleatory or epistemic) uncertainties.
- Verification that the implementation satisfies the safety objectives.

The purpose of the continuous safety assessment is to ensure that the certified/approved systems are in a condition for safe operation during their operating life. To enable such safety management functions, metrics, target values, thresholds and evaluation periods should be defined, and relevant data needs to be recorded.

Information security considerations for ML applications include the following.

- For each AI-based (sub)system and its data sets, the applicant should identify those information security risks with an impact on safety, identifying and addressing specific threats introduced by AI/ML usage.
- The applicant should document a mitigation approach to address the identified AI/ML-specific security risk.
- Systems embedding an AI/ML constituent should be designed with the objective of being resilient and capable of failing safely and securely if attacked by unforeseen and novel information security threats.

An ethics-based trustworthiness assessment should be performed, for instance using the seven gears of the Assessment List for Trustworthy AI (ALTAI) (High-level Expert Group on AI, 2020). This includes the following topics.

- To ensure that the AI-based system bears no risk of creating overreliance, attachment, stimulating addictive behaviour, or manipulating the end user's behaviour.
- To comply with national and EU data protection regulations (e.g. GDPR), i.e. involve their Data Protection Officer, consult with their National Data Protection Authority, etc.

- To ensure that the creation or reinforcement of unfair bias in the AI-based system, regarding both the data sets and the trained models, is avoided, as far as such unfair bias could have a negative impact on performance and safety.
- To ensure that end users are made aware of the fact that they interact with an AI-based system, and, if applicable, whether some personal data is recorded by the system.
- To perform an environmental impact analysis, identifying and assessing potential negative impacts of the AI-based system on the environment and human health throughout its life cycle (development, deployment, use, end of life), and define measures to reduce or mitigate these impacts.
- To identify the need for new skills for users and end users to interact with and operate the AI-based system, and mitigate possible training gaps.
- To perform an assessment of the risk of deskilling of the users and end users and mitigate the identified risk through a training needs analysis and a consequent training activity.

2.4.3 AI assurance

The AI assurance block provides system-centric guidance for learning assurance and for development and post-ops explainability.

2.4.3.1 Learning assurance

Learning assurance aims at providing assurance on the intended behaviour of the AI-based system at an appropriate level of performance, and at ensuring that the resulting trained models possess sufficient generalisation and robustness capabilities. The learning assurance process follows a W-shaped cycle as shown in Figure 2-4 and it consists of the following elements (EASA, 2024).

- **Requirements management.** This step captures requirements allocated to the AI/ML system, addressing safety (e.g. *performance, reliability*), information security, interfaces, etc. It also defines the ranges of input data in the so-called operational design domain, including nominal data, edge cases, outliers, data formats, etc. The requirements and the AI/ML constituent architecture need to be independently reviewed.
- **Data management.** The data management process covers the identification of the datasets used for training and evaluation, and the dataset preparation (including collection, labelling and processing). It also addresses the validation objective of completeness and correctness of the datasets with respect to the product/system requirements and to the ConOps, as well as considerations on the quality of the datasets. Finally, it should cover objectives on the independence between datasets and an evaluation of the bias and variance inherent to the data.
- **Learning process management.** The learning process management considers the selection and validation of key elements of the training phase such as the training algorithm, the activation function, the loss function, the initialisation strategy, and the training hyperparameters. Another consideration is on the training environment, including the host hardware and software frameworks. The metrics that will be used for the various validation and verification steps should be selected (derived from the requirements) and justified.
- **Model training.** The model training consists primarily of executing the training algorithm in the conditions defined in the previous step, using the training dataset originating from the data management process step. Once trained, the model performance, bias and variance are evaluated, using the validation dataset.

- **Learning process verification & integration.** The learning process verification evaluates the trained model performance on the test dataset, including evaluation of the bias and variance of the trained model. The training phase and its verification can be repeated iteratively until the trained model reaches the expected performance. Any shortcoming in the model quality can lead to iterating again on the data management process step, by correcting or augmenting the dataset. A requirements-based verification of the inference model as integrated in the AI/ML system should be achieved.
- **Model implementation.** The model implementation consists of transforming the training model into an executable model that can run on a target hardware. The environment (e.g. software tools) necessary to perform this transformation should be identified, and any associated assumptions, limitations, or optimisations should be captured and validated. The inference hardware should be identified, and peculiarities associated with the learning process should be managed (e.g. specificities due to GPU usage, memory/cache management, real-time architecture).
- **Inference model verification.** The inference model verification aims at verifying that the inference model behaves adequately compared to the trained model, by evaluating the model performance with the test dataset and explaining any differences in the evaluation metric compared to the one used in the training phase verification.
- **Data and learning verification of verification.** This process step aims at verifying that all the data management and learning process steps have been performed correctly and completely. It closes the data management life cycle, by verifying that data sets were adequately managed, considering that the verification of the data sets can be achieved only once the inference model has been satisfactorily verified on the target platform. It verifies that the trained model has been satisfactorily verified, including the necessary coverage analyses.
- **AI/ML constituent requirements verification.** This addresses the verification of the AI/ML constituent fully integrated in the overall system by traditional assurance methodologies (like ED-79B).

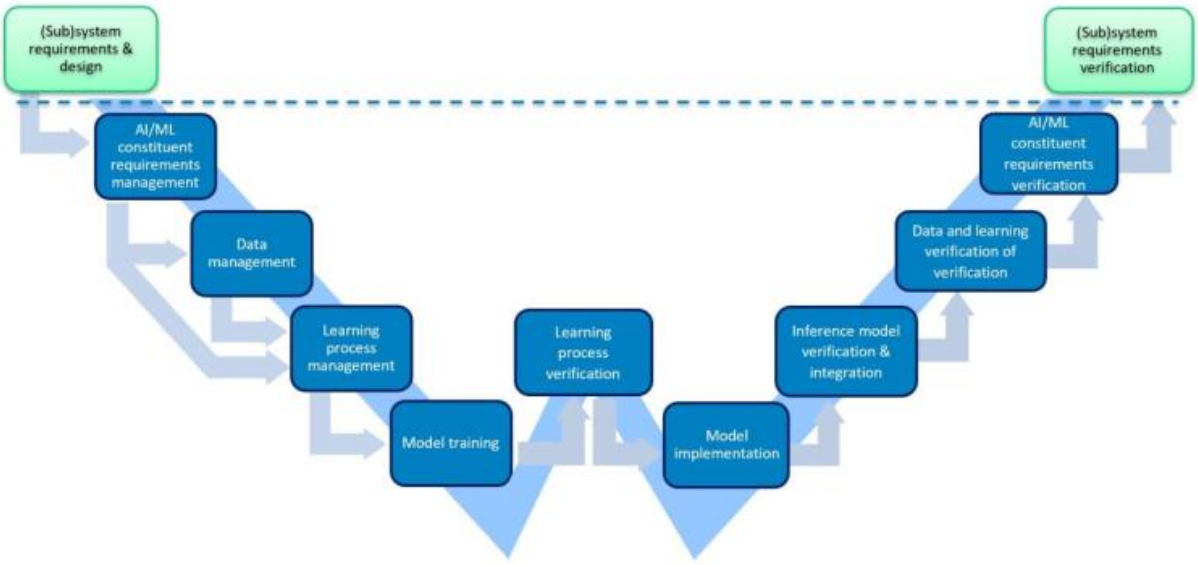

Figure 5.Learning assurance processes (processes below the dashed line), source (EASA, 2024)

In (EASA, 2020b) and (Balduzzi et al., 2021), a visual landing guidance system employing a convolutional neural network processing RGB camera data to detect a runway is used as a detailed example of learning assurance processes. The functional hazard assessment (FHA) of the application uses a functional decomposition of the visual landing guidance system, it identifies associated failure conditions, and it assesses severity levels. Safety objectives are allocated to the failure conditions identified in the FHA. Some architectural means are indicated as possible ways to meet the safety objectives, such as runtime monitoring functions and different instances of independent machine learning models.

Results of the ForMuLa project "*Formal Methods Use for Learning Assurance" regarding the adoption of formal methods in the design assurance process of machine learning*-enabled systems are presented in (EASA and Collins Aerospace, 2023). Formal methods are mathematically rigorous techniques for the specification, development, analysis, and verification of software and hardware systems. The mathematical basis of formal methods consists of formal logic, discrete mathematics, and computer-readable languages. In the report, a range of formal methods and their use in the W-cycle diagram for learning assurance are provided. Illustrations of their application are provided for a use case for ML-based prediction of remaining useful life in prognostic and health management.

The Machine learning Application Approval (MLEAP) project (MLEAP Consortium, 2023) has done research on methods for learning assurance processes in Figure 2-4 in support of the EASA AI roadmap. A general framework for learning assurance is proposed, shown in Figure 2-5, which supports controlling the complexity and capacity of the models depending on the scope of the task under development, and the volume and nature of input data, while measuring the level of generalisation reached by a training session. Furthermore, approaches are presented to ensure the stability and robustness of machine learning models.

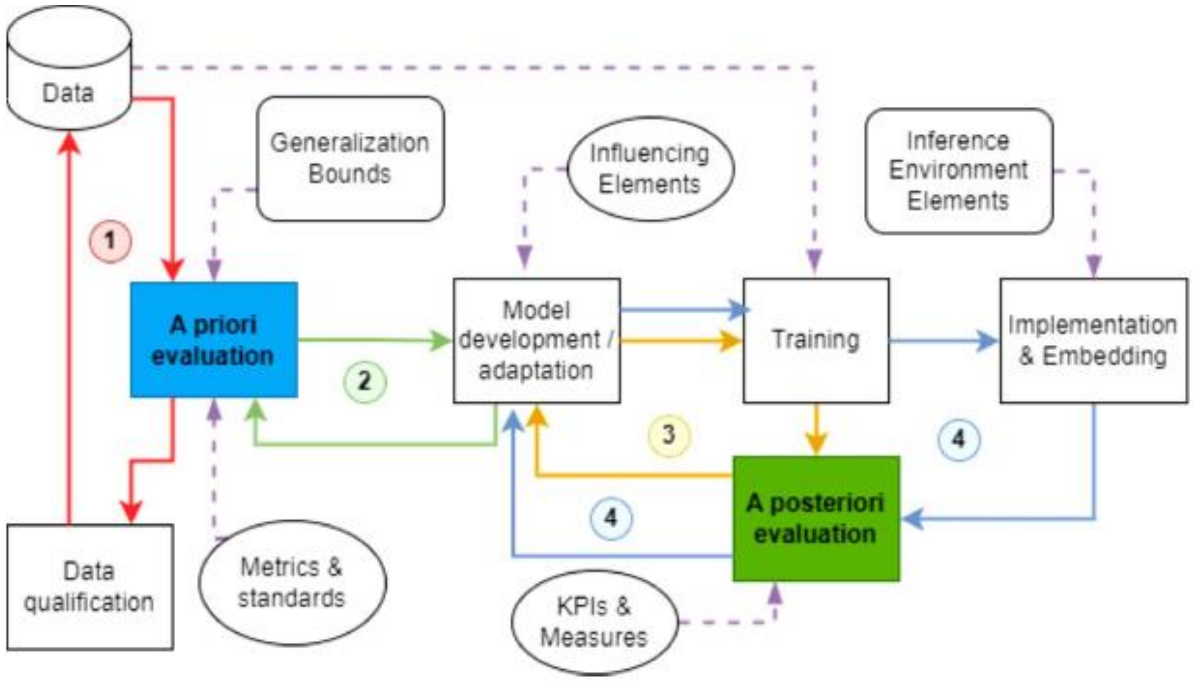

Figure 6. General framework to support learning assurance processes, proposed in (MLEAP Consortium, 2023)

2.4.3.2 Development and post-ops explainability

AI explainability is the capability to provide humans with understandable, reliable, and relevant information on how an AI/ML application is coming to its results, provided with the appropriate level of detail and at an appropriate time. The target audience for development and post-ops explainability includes engineers, certification authorities and safety investigators to support the development, and learning from occurrences. Examples of objectives in support of this type of explainability include:

- To identify and document the methods at AI/ML item and/or output level satisfying the specified AI explainability needs.
- To deliver an indication of the level of confidence in the AI/ML constituent output, based on actual measurements or on quantification of the level of uncertainty.
- To design the AI-based system with the ability to monitor its input and outputs to ensure that they are within the specified bounds.
- To provide the means to record operational data that is necessary to explain, post operations, the behaviour of the AI-based system and its interactions with the end user, as well as the means to retrieve this data.

2.4.4 Human factors for AI

It is explained in (EASA, 2024) that elements from existing human factors requirements and guidance are applicable for AI-based installed systems and equipment for use by the end users. For instance, **AMC 25.1302** provides design guidance and principles as well as human factors methods for flight deck design. Complementary guidance is provided for AI-based systems.

AI operational explainability is provided during operations and informs end users like flight crews, air traffic controllers and maintenance staff. It supports them in understanding decisions, predicting AI behaviour, building trust, and a suitable human-machine interface. The level of explainability depends on the AI level of the AI system, ranging from no change with respect to current systems for Level 1A (human augmentation) to very new explainability regimes for Level 2B (Human-AI Teaming: collaboration). Examples of objectives for operational AI explainability include:

- To ensure that the AI-based system presents explanations to the end user in a clear and unambiguous form.
- To define relevant explainability so that the receiver of the information can use the explanation to assess the appropriateness of the decision/action as expected.
- To allow the end user to customise the level of abstraction as part of the operational explainability.
- To define the timing when the explainability will be available to the end user taking into account the time criticality of the situation, the needs of the end user, and the operational impact.
- To enable the end user to get upon request explanation or additional details on the explanation when needed.

Human AI-based system teams (**HAT**) use cooperation or collaboration between humans and AI-based systems. Cooperation is a process in which the AI-based system works to help the end user accomplish

their own objective and goal. The AI-based system will work according to a predefined task allocation pattern with informative feedback on the decision and/or action implementation. Collaboration is a process in which the human and the AI-based system work together and jointly to achieve a common goal (or work individually on a defined goal) and solve a problem through a joint constructive approach. Collaboration implies the capability to share situation awareness and to readjust strategies and task allocation in real-time. Examples of objectives in **HAT** include the following.

- To design the AI-based system with the ability to build its own individual situation representation.
- To design the AI-based system with the ability to reinforce the end-user individual situation awareness.
- To design the AI-based system with the ability to enable and support shared situation awareness.
- To design the AI-based system with the ability to identify a suboptimal strategy and propose through argumentation an improved solution.
- To design the AI-based system with the ability to identify the problem in complex situations under abnormal operations, and to share the diagnosis including the root cause, the resolution strategy and the anticipated operational consequences.
- To design the AI-based system with the ability to detect poor decision-making by the end user in a time-critical situation, alert and assist the end user.

Design guidance objectives are provided for new modes of human-machine interaction through voice, gesture, or other. For instance, if spoken natural language is used, the AI-based system should have the ability to process end-user requests, responses and reactions, and provide an indication of acknowledgement of the user's intentions, it should notify the end user that they possibly misunderstood information, and the AI-based system should have the ability to resolve the issue. If gesture language is used, the gesture language syntax should be intuitively associated with the command that it is supposed to trigger and the system should have the ability to disregard non-intentional gestures.

It is expected that the introduction of AI-based systems may contribute to human errors in various ways (EASA, 2024), *e.g.* increased likelihood of errors if an AI system fails due to over-reliance on the system, more judgement errors because of complex information streams by the AI-based system that cannot be fully overseen by a user, unexpected failure modes of an AI system which cannot be effectively handled by users, lack of transparency of the AI-based system leading to errors in decision-making or action implementation. Several objectives are defined to reduce the risk of human errors, stating that the AI-based system should minimise the likelihood of design-related end-user errors and human AI resource management errors, that it should be tolerant to end-user errors and that the system should allow the user to detect errors and cope with them efficiently.

Finally some objectives are provided for failure management, such as enabling failure diagnosis and presenting the pertinent information to the user, presenting a possible solution of the failure condition to the end user and supporting the user in implementing such a solution.

2.4.5 AI safety risk mitigation

Once activities associated with all other building blocks are defined, the applicant should determine whether the coverage of the objectives associated with the explainability and learning assurance

building blocks is sufficient or whether an additional dedicated layer of protection, i.e. safety risk mitigation, would be necessary to mitigate the residual risks to an acceptable level. If residual risks are too high, effective safety risk mitigation means should be defined, e.g. by real-time monitoring of the output of the AI/ML constituent and passivation of the AI-based system with recovery through a traditional backup system.

2.4.6 Organisations

High-level objectives providing guidance to organisations for the introduction of AI-based systems include:

- The organisation should review its processes and adapt them to the introduction of AI technology.
- Implement a data-driven '*AI continuous safety assessment system*' based on operational data and in-service events.
- The organisation should establish means to continuously assess ethics-based aspects for the trustworthiness of an AI-based system.
- The organisation should ensure that the safety-related AI-based systems are auditable by internal and external parties, including the approving authorities.
- The organisation should adapt the continuous risk management process to accommodate the specificities of AI, including interaction with all relevant stakeholders.
- The organisations operating AI-based systems should ensure that end users' licensing and certificates account for the specificities of AI, including interaction with all relevant stakeholders.

2.5 Research roadmaps for increasingly autonomous operations

The Committee on Autonomy Research for Civil Aviation of the National Research Council of the National Academies (USA) published an early research agenda study for autonomy research in civil aviation (National Research Council, 2014), which contains the following topics.

1. **Behaviour of Adaptive/Nondeterministic Systems.** Develop methodologies to characterise and bound the behaviour of adaptive/nondeterministic systems over their complete life cycle.
    (a) Develop mathematical models for describing adaptive/non-deterministic processes as applied to humans and machines.
    (b) Develop performance criteria, such as stability, robustness, and resilience, for the analysis and synthesis of adaptive/nondeterministic behaviours.
    (c) Develop methodologies beyond input-output testing for characterising the behaviour of IASs (Increasingly Autonomous Systems).
    (d) Determine the roles that humans play in limiting the behaviour of adaptive/nondeterministic systems and how IASs can take over those roles.

2. **Operation Without Continuous Human Oversight**. Develop the system architectures and technologies that would enable increasingly sophisticated IASs and unmanned aircraft to operate for extended periods of time without real-time human cognisance and control.
    (a) Investigate human roles, including temporal requirements for supervision, as a function of the mission, capabilities, and limitations of IASs.

(b) Develop IASs that respond safely to the degradation or failure of aircraft systems.
(c) Develop IASs to identify and mitigate high-risk situations induced by the mission, the environment, or other elements of the National Airspace System (NAS).
(d) Develop detect-and-avoid IASs that do not need continuous human oversight.
(e) Investigate airspace structures that could support UAS (Unmanned Aerial Systems) operations in confined or pre-approved operating areas using methods such as geofencing.

3. **Modelling and Simulation.** Develop the theoretical basis and methodologies for using modelling and simulation to accelerate the development and maturation of advanced IASs and aircraft.
   (a) Develop theories and methodologies that will enable modelling and simulation to serve as embedded components within adaptive/non-deterministic systems.
   (b) Develop theories and methodologies for using modelling and simulation to coach adaptive IASs and human operators during training exercises.
   (c) Develop theories and methodologies for using modelling and simulation to create trust and confidence in the performance of IASs.
   (d) Develop theories and methodologies for using modelling and simulation to assist with accident and incident investigations associated with IASs.
   (e) Develop theories and methodologies for using modelling and simulation to assess the robustness and resiliency of IASs to intentional and unintentional cybersecurity vulnerabilities.
   (f) Develop theories and methodologies for using modelling and simulation to perform comparative safety risk analyses of IASs.
   (g) Create and regularly update standardised interfaces and processes for developing modelling and simulation components for eventual integration.
   (h) Develop standardised modules for common elements of the future system, such as aircraft performance, airspace, environmental circumstances, and human performance.
   (i) Develop standards and methodologies for accrediting IAS models and simulations.

4. **Verification, Validation, and Certification (VV&C).** Develop standards and processes for the verification, validation, and certification of IASs, and determine their implications for design.
   (a) Characterise and define requirements for intelligent software and systems.
   (b) Improve the fidelity of the VV&C test environment.
   (c) Develop, assess, and propose new certification standards.
   (d) Define new design requirements and methodologies for IASs.
   (e) Understand the impact that airspace system complexity has on IAS design and on VV&C.
   (f) Develop VV&C methods for products created using non-traditional methodologies and technologies.

5. **Non-traditional Methodologies and Technologies.** Develop methodologies for accepting technologies not traditionally used in civil aviation (e.g., open-source software and consumer electronic products) in IASs.
   (a) Develop modular architectures and protocols that support the use of open-source products for non-safety critical applications.

(b) Develop and mature non-traditional software languages for IAS applications.
(c) Develop paths for migrating open-source, intelligent software to safety-critical applications and unrestricted flight operations.
(d) Define new operational categories that would enable or accelerate experimentation, flight testing, and deployment of non-traditional technologies.

6. **Roles of Personnel and Systems.** Determine how the roles of key personnel and systems, as well as related human-machine interfaces, should evolve to enable the operation of advanced IASs.
   (a) Develop human-machine interface tools and methodologies to support the operation of advanced IASs during normal and atypical operations.
   (b) Develop tools and methodologies to ensure effective communication among IASs and other elements of the NAS.
   (c) Define the rationale and criteria for assigning roles to key personnel and IASs and assess their ability to perform those roles under realistic operating conditions.
   (d) Develop intuitive human-machine integration technologies to support real-time decision-making, particularly in high-stress, dynamic situations.
   (e) Develop methods and technologies to enable situational awareness that supports the integration of IASs.

7. **Safety and Efficiency.** Determine how IASs could enhance the safety and efficiency of civil aviation.
   (a) Analyse accident and incident records to determine where IASs may have prevented or mitigated the severity of specific accidents or classes of accidents.
   (b) Develop and analytically test methodologies to determine how the introduction of IASs in flight operations, ramp operations by aircraft and ground support equipment, ATM systems, airline operation control centres, and so on might improve safety and efficiency.
   (c) Investigate airspace structures and operating procedures to ensure safe and efficient operations of legacy and IASs in the NAS.

8. **Stakeholder Trust.** Develop processes to engender broad stakeholder trust in IASs for civil aviation.
   (a) Identify the objective attributes of trustworthiness and develop measures of trust that can be tailored to a range of applications, circumstances, and relevant stakeholders.
   (b) Develop a systematic methodology for introducing IAS functionality that matches authority and responsibility with earned levels of trust.
   (c) Determine the way in which trust-related information is communicated.
   (d) Develop approaches for establishing trust in IASs.

A recent roadmap for autonomy verification and validation, including a considerable overview of current progress, is provided in a study by NASA in cooperation with representatives of the aerospace industry and academia (Brat et al., 2023). It includes the following topics:

- **Compositional verification techniques.** Compositional verification refers to the theories, technologies, and tools to verify subsystem components by parts. Compared to verification that produces guarantees for the whole system, compositional verification scales with system

complexity and allows the creation of modular and verified subsystem components. New methods are needed to support the inclusion of learned components.

- **Hybrid systems V&V techniques.** Enhanced methods are needed for the modelling and simulation of hybrid systems, which are represented by combinations of continuous and discrete state spaces.
- **Machine learning V&V.** There is a need for the development of acceptable assurance techniques, including simulation, testing, and verification.
- **Safety in human-autonomy interaction.** Human-autonomy interaction is mainly discussed for the robot industry. It lacks key insights of human-autonomy issues in aerospace applications.
- **Runtime assurance.** Run-time assurance (RTA) architectures are a method to address the residual challenges posed by complex algorithms by monitoring systems' behaviour during operation. RTA architectures add high-assurance components to a system design to ensure that the components containing complex behaviours (or difficult-to-certify algorithms such as ML) cannot cause unsafe or unintended system behaviours. The high-assurance components include run-time monitors, safety backup components, and a switch that manages whether the complex component or a safety backup is being used. New methods are needed to show that interactions among monitors and mitigation actions are not in conflict. These methods will need to be integrated with current standards and certification practices for design and safety analysis.
- **Autonomy for contingency planning.** Contingency management systems play a critical role in Aerospace system safety. The first step is to *recognise* the situation, and then the system must *appropriately respond* to this situation. The response could be as straightforward as switching to a redundant backup system or as complex as planning and executing an urgent or emergency landing. Contingency management systems can be prepared to respond to well-characterised anomalies such as adverse weather, control actuators, and sensor failures with deterministic real-time response protocols that can be verified, validated, and certified. The first step to developing autonomy for contingency management is to understand and capture models of risks, hazards, and mitigation strategies. The next step is to capture checklists, implement watchdogs, and prove data link technology in a manner that addresses the "easy" contingency management responses, enables software monitoring software (in lieu of human monitoring), and that prepares for the high-bandwidth low-latency information pipeline required for diverse traffic in a shared airspace to coordinate and accommodate the requests of a distressed aircraft without delay.
- **Dynamic assurance**. Dynamic certification of autonomous systems requires the involvement and distillation of knowledge from diverse stakeholders, from engineers to architects to regulators. It must start at the beginning of the life cycle, where decisions are most effective and cost of design changes are lowest. Iterative assessment for dynamic certification depends on the use and context in which autonomous systems are expected to deploy and must account for both technical and social requirements. Multiple rounds of testing and interrogating requirements via formal methods provide insights into the uncertainty present in varying deployment scenarios, where they must perform efficiently and safely (Bakirtzis et al., 2023).

2.6 FAA roadmap and methods for AI Safety Assurance

The FAA is developing a roadmap for AI safety assurance and has organised technical interchange meetings (TIMs) in September 2023 and January 2024 for coordination with stakeholders. This section provides some insights from presentations during these TIMs. These presentations were used to stimulate discussion and show intermediate results.

The focus of the AI safety assurance roadmap is on systems that use learned algorithms or learning during their operational life, rather than rule-based AI systems. It states that in principle, existing civil aviation safety requirements, processes and methods will be used for AI-based systems, except where they are found to be inadequate. It is considered that the personification of AI systems should be avoided and that it must always be clear to humans whether they are dealing with an AI system. In the development of the AI safety assurance methods an incremental approach will be adopted, which is updated with experience, starting with lower-risk applications. It is considered that the ethical use of AI is outside of the scope of safety assurance.

The key pillar for assuring the safety of AI is to address whether a learned model performs as expected and how any aberrant behaviours are mitigated in the system. In a study for the FAA, Paul et al. (2023) proposed the use of an Overarching Properties-Driven Approach (Holloway, 2019) for assurance of ML-based aerospace systems. Three overarching properties have been defined that are sufficient to establish the suitability of a product for installation on an aircraft:

- **Intent**: The defined intended behaviour is correct and complete with respect to the desired behaviour
- **Correctness**: The implementation is correct with respect to its defined intended behaviour, under foreseeable operating conditions, and
- **Innocuity**: Any part of the implementation that is not required by the defined intended behaviour has no unacceptable impact.

Desired behaviour denotes the needs and constraints expressed by the stakeholders, *defined intended behaviour* is the physical representation of the desired behaviour (e.g., a set of requirements), *implementation* is the hardware or software element or combination of items for which approval is being sought, *foreseeable operating conditions* are the external and internal conditions in which the element to be evaluated is to be used, encompassing all known normal and abnormal conditions, and *unacceptable impact* is any impact that can lead to a direct or indirect undesirable effect on an aircraft or its components. To warrant approval for a hardware or software element, it is necessary to show with evidence that the element meets all three overarching properties. In (Paul et al., 2023) the overarching properties approach is applied to a use case of a Recorder Independent Power Supply (RIPS) system, which provides several minutes of backup power to the data recorder when an aircraft loses access to standard power supply. In the use case neural network-based predictions are used to increase the time between maintenance actions of the RIPS battery. Building on a functional hazard assessment and using an argument structure, overarching properties were justified for this DAL-D use case.

At the Highly Automated Safety Center of Excellence (HASS) of the US Department of Transportation a framework for runtime assurance of AI/ML is developed, which should complement verification and validation methods in the design phase. Components for runtime assurance include runtime monitors and recovery functions, which provide dynamic consistency checking and enable recovery functionality

if particular safety boundaries are crossed by the AI/ML-based system. An industry standard for run time assurance for aircraft systems is (Nagarajan et al., 2021). The overall AI assurance framework incorporates:

- **AI Design time assurance (DTA).** DTA activities include analysis, verification, validation, testing and simulation of AI models and data in the development, as well as use of the assurance results in structured argument to demonstrate predefined safety and performance goals are achieved. DTA covers AI development activities from data collection/generation, model design, training, optimization, compiling to inference.
- **AI Operation time assurance (OTA)**. OTA deals with operational constraints, uncertainties, faults/failures, and any unexpected situations that cannot be addressed by DTA. OTA monitors AI components and their enabled systems to detect any violations of operation time requirements; with safety guards, OTA ensures only safe functions are performed. OTA includes runtime verification for deployment, integration, online data, inference models, as well as safety arguments regarding expected safety and performance goals based on OTA results and operational conditions. Safety goals are usually directly derived from the analysis of safety hazards and risks identified from AI safety assessment activities.
- **AI contingency management (CM).** CM mainly addresses system failures and situations that can't be addressed by AI and its assurance, as a very last measure to guarantee system and operation safety with potentially degraded service or end of service, e.g., emergency stop or contingency landing.

In a presentation on the AI/ML assurance, dr. Natasha Neogi from NASA referred extensively to the book "*Software for Dependable Systems: Sufficient Evidence?*" (National Research Council, 2007). It specifies three E's for certifying dependable software: explicit claims, evidence, and expertise. A claim on the dependability of the system must articulate precisely the properties the system is expected to exhibit and the assumptions about the system's environment upon which the claim is contingent. For an AI/ML component this means that the operational context, human interactions and environmental conditions should be well specified. Concrete evidence is usually a combination of testing and analysis, including modelling and simulation. Expertise means that developers should be familiar with best practices and deviate from them only when needed, but also that experts can wisely tailor their approach to assuring novel elements with respect to methods, languages, tools, and processes.

An aircraft AI/ML learning assurance approach was presented by Norm Fenlason of MITRE. The approach is based on verification uncertainty and risk, describing the potential for negative impact on system safety should a verification activity not reveal system behavioural shortfalls and subsequent unacceptable conditions. Such verification risks are handled in the verification assurance by design-level risk mitigations, e.g. design modifications to account for verification risk, in system-level risk mitigation, e.g. runtime assurance, and operational limitations, e.g. constraints on usage.

2.7 NASEM and HFES certification frameworks for human-AI teaming

A consensus study report was published by the National Academies of Sciences, Engineering, and Medicine about state of the art and research needs for human-AI teaming (NASEM, 2022). The report focuses on human-AI interaction in military operations, but the identified state of the art and research needs are also relevant for civil operations, such as ATC and flight operations. The study identified 57 research objectives for near-term, mid-term, and far-term studies for the following topics.

- **Human-AI team models.** Improved computational models of human-AI teams are called for, that consider the interrelated, dynamically evolving, distributed, and adaptive collaborative tasks and conditions in operations. Improved metrics for human-AI teaming are needed that consider the team's ability to manage interdependencies and dynamic role assignments, that reduce uncertainty, and that improve the ability of the AI system to deliver capabilities that are in line with expectations of operators.
- **Human-AI team processes**. Supporting humans and AI systems as teammates relies on a carefully designed system with the capability for both taskwork and teamwork. Along this line, research is needed to improve team effectiveness in long-term, distributed, and agile human-AI teams through improved team assembly, goal alignment, communication, coordination, social intelligence, and the development of a new human-AI language.
- **Situation awareness.** Human situation awareness (**SA**) is critical for effective operations, including the oversight of AI systems. Methods for improving human SA of AI systems need to be developed that consider diverse types of applications, timescales of operations, and the changing capabilities associated with machine learning (**ML**)-based AI systems. In addition, research directed at creating shared SA within the human-AI team deserves attention. The degree to which AI systems need to have both self-awareness and awareness of their human teammates needs to be explored, to determine the benefit for overall team performance. Finally, future AI systems will need to possess integrated situation models to appropriately understand the current situation and to project future situations for decision-making. AI models of the dynamic task environment will be needed that can work with humans to align or deconflict goals and to synchronise situation models, decisions, function allocations, task prioritisation, and plans to achieve coordinated and approved actions.
- **AI transparency and explainability**. Display transparency considers the provision of a real-time understanding of the actions of an AI system as a part of situation awareness (**SA**). Explainability considers the provision of information in a backwards-looking manner based on the logic, process, factors, or reasoning upon which the system's actions or recommendations are based. In dynamic, time-constrained operations, explanations will primarily contribute to the development of improved mental models that can improve SA in the future, and decision-making will be primarily reliant on real-time display transparency. In situations that involve sufficient time for reviewing and processing explanations, both display transparency and explainability may directly impact decision-making. System transparency and explainability are key mechanisms for improving SA, trust, and performance in human-AI teams. Methods for supporting transparency and explainability in future human-AI teams need to consider the appropriate types of information, methods for displaying that information, and timeliness of information presentation, particularly as these factors relate to dynamically changing AI systems. Methods for tailoring and adapting transparency and explainability information would benefit from further exploration, as would the advantages of bi-directional explanation in human-AI teams.
- **Human-AI team interaction**. Interaction mechanisms and strategies within the human-AI team are critical to team effectiveness, including the ability to support flexible assignments of levels of automation (**LOAs**) across functions over time. Research is needed to determine improved methods for supporting collaboration between humans and AI systems in shared functions, to support human operators working with AI systems at multiple LOAs, and to determine methods for maintaining or regaining SA when working with AI systems at high LOAs (i.e., *on-the-loop control*). Research is also needed to determine new requirements to support dynamic

functional assignments across human-AI teams, and to determine the best methods for supporting dynamic transitions in LOAs over time, including when such transitions should occur, who should activate them, and how they should occur, to maintain optimal human-AI team performance.

- **Trust**. Trust in AI is a foundational factor associated with use of AI systems. It would benefit future research to better document the decision context and goals involved in the teaming environment, to advance understanding of how broader socio-technical factors affect trust in human-AI teams. Interaction structures that extend beyond supervisory control arrangements would also benefit from further study, particularly to understand the effect of AI directability on the trust relationship. Dynamic models of trust are needed to capture how trust evolves and affects performance outcomes in various human-AI team contexts.
- **Bias**. The potential for bias in AI systems can be introduced through the development of its algorithms as well as through systemic biases in training sets, among other factors. Further, humans can suffer from several well-known decision biases. Human decision-making can be directly affected by the accuracy of the AI system, creating a human-AI team bias. Research is needed to better understand the interdependencies between human and AI decision-making biases, how these evolve over time, and methods for detecting and preventing bias with ML-based AI. Research is also needed to detect and prevent potential adversarial attacks that may attempt to take advantage of these biases.
- **Training**. Training of the human-AI team will be needed to develop the appropriate team constructs and skills necessary for effective performance. Directed research is needed to determine what, when, why, and how to best train human-AI teams, taking into consideration various team compositions and sizes. Training may be needed to better calibrate human expectations of AI teammates and to foster appropriate levels of trust. Specific platforms will be necessary to develop and test human-AI teamwork procedures.
- **Human-system integration (HSI)** *processes and measures.* Good HSI practice will be key to the design, development, and testing of new AI systems, particularly with respect to system development based on agile practices. New HSI design and testing methods for effective human-AI teams will also be needed, including an improved ability to determine requirements for human-AI teams, particularly those that involve ML. Improved approaches for multidisciplinary AI development teams are needed that include human factors engineers, sociotechnical researchers, systems engineers, and computer scientists. New teams, methods, and tools centred around AI lifecycle testing and auditability, as well as AI cyber vulnerability, will also be needed. Methods for testing and verification of evolving AI systems need to be developed to detect AI system blind spots and edge cases and to consider brittleness. New human-AI testbeds to support research and development activities by these new teams will also be important. Finally, improved metrics for human-AI teaming may be needed, specifically regarding matters of trust, mental models, and explanation quality.

Building on Bainbridges seminal paper "*Ironies of automation*" (Bainbridge, 1983), Endsley (2023) discusses five ironies of artificial intelligence:

- Artificial intelligence is still not that intelligent;
- The more intelligent and adaptive the AI, the less able people are to understand the system;
- The more capable the AI, the poorer people's self-adaptive behaviours for compensating for shortcomings;

- The more intelligent the AI, the more obscure it is, and the less able people are to determine its limitations and biases and when to use the AI;
- The more natural the AI communications, the less able people are to understand the trustworthiness of the AI.

As a way to try to overcome these ironies of AI, Endsley provides various future directions, including the importance of **human-centred AI**, the need to **explicitly identify AI** as bots, emphasis on **explainability** and **transparency**, clearly **exposing biases and limitations** of AI to users, provide humans with **meaningful control** over AI, ensure that people develop and **retain cognitive skills** for performance, and **evaluate safety and reliability of AI** systems in conjunction with people using them.

The Human Factors and Ergonomics Society (HFES) developed a **Human Readiness Level** (**HRL**) scale to provide a mechanism to evaluate, track, and communicate the readiness of a technology or system for safe and effective human use that can be applied in the context of existing systems engineering and human systems integration (HSI) processes (HFES, 2021) for advanced automation. Inspired by the Technology Readiness Level scale, the HRLs comprise nine levels, as shown in Table 1.

It is applicable to any type of technology under development in the military, government, federal agencies, industry, and academia. FAA has made steps for integrating HRLs in their research, acquisition and system development process (*Austrian et al., 2023*).

HRL	Description
1	Basic principles for human characteristics, performance, and behaviour observed and reported
2	Human-centred concepts, applications, and guidelines defined
3	Human-centred requirements to support human performance and human-technology interactions established
4	Modelling, part-task testing, and trade studies of human systems design concepts and applications completed
5	Human-centred evaluation of prototypes in mission-relevant part-task simulations completed to inform design
6	Human systems design fully matured as influenced by human performance analyses, metrics, prototyping, and high-fidelity simulations
7	Human systems design fully tested and verified in operational environment with system hardware and software and representative users
8	Total human-system performance fully tested, validated, and approved in mission operations, using completed system hardware and software and representative users
9	System successfully used in operations across the operational envelope with systematic monitoring of human-system performance

Table 1. Human Readiness Levels, as defined in (HFES, 2021)

2.8 EUROCAE and SAE working groups on AI certification

EUROCAE Working Group 114 "*Artificial Intelligence in Aeronautical Systems*" and SAE G-34 "*Artificial Intelligence in Aviation*" are jointly working on a standard for development and certification/approval of aeronautical safety-related products implementing machine learning. As a basis for the development of this standard these groups produced a report with a statement of concerns, which includes a gap analysis for limitations of applying existing standards for certification of AI (EUROCAE, 2021). The scope of the current work is on offline learning applications, where ML models are trained and implemented in a fixed production system without additional learning.

The following topics are addressed in ER-022 (EUROCAE, 2021). A wide range of AI-related definitions and a classification of AI techniques are provided. A gap analysis of existing development assurance standards is presented, discussing at a high-level limitations for AI-based systems of ARP4754A/ED-79A, ARP4761, ED-12C/DO-178C, ED-218/DO-331, ED-215/DO-330, ED-216/DO-333, ED-217/DO332, ED-80/DO-254, ED-109A/DO-278A, ED-153, etc. The identified gaps highlight that the data-driven paradigm of ML may not be adequately addressed by the existing standards. Some specific considerations are provided for the ML aspects, like data selection and validation, model selection, training, and testing, inference implementation, and system integration and verification. Aspects of a potential approach for certification/approval of AI-based systems are presented, such as learning assurance, formal methods, testing, explanation, and licensing. Finally, some use cases of applying AI in aircraft systems and ATMs are sketched.

The development of a new standard for the certification of data-driven AI systems in airborne and ATM systems considers design assurance levels (DALs) for technical systems. Ethical considerations, human factors and information security are not in the current focus. It is considered that a particular product (aircraft, ATM system) can be decomposed into a number of subsystems, which may encompass conventional software and hardware components or ML constituents. At a high level, the same domain-agnostic and application-agnostic processes are applied for conventional as well as ML-based systems:

- **System Development:** A structured process (usually top-down) to define and implement the desired system's intended function in the context of its intended usages, including the arrangement (architecture) of its subsystems and items, from initial concept to its entry into service.
- **Safety Assessment:** A process which identifies and classifies hazard effects or failure conditions associated with the system functions, generates safety objectives/criteria, and determines the minimum level of rigour to be applied to the associated development assurance activities according to the severity classification of those hazard effects or failure conditions. The process includes safety requirement(s) identification and confirmation that the implementation satisfies the safety requirements.
- **Development Assurance:** All planned and systematic actions used to substantiate, at an adequate level of confidence, that development errors have been identified and corrected such that the system satisfies the applicable safety objectives/criteria. The expected level of confidence is indicated by an assurance level (AL) for the conventional or ML constituent.

In these processes there exist specific focus points that need to be taken into account for the features of machine learning, such as:

- To account for ML specifics at the interfaces of systems, e.g. input and output validity metrics.
- To consider data collection considerations and ML robustness properties in the system architecture.
- The safety assessment process and the need to constrain MLCs to their allocated operating environment may drive the system architecture to include protections depending on the nature of the intended function and the severity of its failure conditions/hazard effects.

2.9 Test and evaluation approach for AI-enabled systems at the US Air Force

In a study for the US Air Force (NASEM, 2023), test and evaluation challenges in AI-enabled systems were addressed towards the central question: how to achieve sufficient confidence in AI-enabled systems? It discusses how human-AI interfaces present new challenges as responsibilities shift between humans and intelligent machines and new concepts of operations emerge. More prominence should be given to Human Readiness Levels (HRL) and UI/UX for AI-enabled systems. Measures of performance and effectiveness, including assessments of user trust and justified confidence, must be formulated during system design and development, and assessed throughout test and evaluation and after system fielding.

The effective use of AI implies transitioning from a waterfall to an agile development methodology. AI implementations are developed cyclically, often referred to as either AIOps or MLOps, and require continuous training, evaluation, and retraining as operational conditions change. Key components include:

- **Trained labellers:** Labellers are trained on tooling and the data they are labelling.
- **Continuous monitoring, retraining, and redeployment of AI models:** Model performance is constantly monitored. Models are regularly retrained and redeployed.
- **Instrumented deployment platforms to capture ML-ready data:** Both the deployed models and the data streams they consume must be instrumented to capture the behaviour deviation and the observations that manifested the performance shift.
- **Synthetic data engines and supporting digital twins:** Enable faster incorporation of emergent threats, observed domain shifts, or previously unknown edge cases. These components must be built for the appropriate domains and modalities.

In (NASEM, 2023) a useful overview is provided of the testing & evaluation (T&E) framework of Nvidia in the development of autonomous systems. Within this framework, the development of the AI-enabled system begins with defining the product specification (*e.g.*, what does the system need to do?). The product specification drives the risk model creation that, in turn, generates the functional requirements to achieve the goals of the system. The product specifications and the risk model are continuously updated through cyclical review. To safely and effectively integrate AI capabilities into safety-critical system processes, continual test and refinement approaches must be implemented to manage against accepted and residual risks. Validation and verification are accomplished in both complex simulation environments and real-world test fleet deployments. Both capabilities feed refinements back into product specifications. Various types of tests are employed, including replay of collected data, replay of augmented data, simulation, and track and road testing.

The concept of *"justified confidence"* is introduced as a progressive measure of trustworthiness: developers, testers, and users should gain justified confidence in AI-enabled systems over time as they

become increasingly familiar with system performance limits and behaviours. *AI assurance* is the process of evaluating, monitoring and ensuring the reliability, effectiveness, robustness, and safety of AI systems. AI assurance comprises a set of practices and methodologies for assessing the quality of AI models and systems, including verifying their accuracy and performance, detecting and mitigating potential biases, and evaluating their ethical and societal implications. The goal of AI assurance is to provide confidence in the decision-making processes of AI systems and to promote the responsible and trustworthy deployment of AI technologies.

The formulation of T&E requirements across the AI life cycle is linked inextricably to the concept of risk management. It is advised in (NASEM, 2023) that an AI Risk Management Framework (RMF) is adopted, such as (NIST, 2023). An AI RMF includes assessing and understanding the potential risks of fielding AI-enabled systems based on different levels of dedicated T&E, communicating risks to decision-makers and end-users, and determining responsibility and accountability for system failure or unanticipated performance problems. The key pillars of the AI RMF of (NIST, 2023) are shown in Figure 2-6, encompassing "*Govern: A culture of risk management is cultivated and present; Map: context is recognized and risks related to context are identified; Measure: Identified risks are assessed, analysed, or tracked; Manage: Risks are prioritised and acted upon based on a projected impact*." Potential risks must be considered at every stage of the AI life cycle. Such kinds of pillars can also be recognised in safety management frameworks (ICAO, 2018; Stroeve et al., 2022).

Key recommendations in (NASEM, 2023) for testing and evaluation of AI-based systems include

- The adoption of an agile and cyclic AIOps/MLOps approach for the development and testing of AI-based systems.
- To use red teams. These teams must be capable of emulating current and future peer competitor capabilities and performance and should be integrated into the entire AI life cycle. They should be integral to AI T&E, *although independent*, and focused on *operational performance* and *mission resilience* in the face of known and unknown—but expected—adversarial attacks, beginning with the presumption of attack at every stage of the AI life cycle, including cyberattacks, data manipulation, and data corruption and poisoning.
- To provide AI education, training and possibly certification to personnel.

Figure 7. AI Risk Management Framework of (NIST, 2023)

2.10 ISO/IEC Cross-industry standards for Software and AI

For the purposes of this document, the approach developed within the ISO/IEC framework is particularly pertinent due to a series of standards that establish criteria for evaluating the quality of software products, data, services, quality in use, the performance of machine learning models, the construction and maintenance of AI systems, explainability in AI systems, and safety-related systems under functional security. Specifically, we have considered the contributions offered by the following:

- The **ISO/IEC 25000** series of standards provides a comprehensive framework for evaluating the quality of software products, data, services, and quality in use. It aims to improve the quality of ICT systems by focusing on the early identification of **quality requirements**, the use of **quality models**, and the measurement of **product quality**. The standards are applicable to a wide range of application domains, including traditional and innovative applications. The key benefits of using the **ISO/IEC 25000** standards include increased **quality**, improved **digital skills** of personnel, and better **design** and **evaluation** of IT services. The standards are relevant to a variety of stakeholders, including innovative start-ups, industries, companies, banks, insurance companies, public administrations, and organisations for the supervision and monitoring of the quality of digital products.

- **ISO/IEC TS 4213:2022**. This standard emphasises the importance of consistent **methods for evaluating the performance of machine learning models** in **classification** tasks. It highlights that advancements are often measured against existing models or baselines, but the choice of metrics depends on the specific application and limitations. Moreover, it offers examples of performance improvements, including higher accuracy, lower training data requirements, and

faster inference speeds. To ensure the validity of these claims, **ISO/IEC** outlines factors to consider, such as **model implementation**, **dataset characteristics**, and **calculation methods**. It emphasises the relevance of these approaches for various stakeholders in the field of Artificial Intelligence, as defined by other standards of the **ISO/IEC** family. Continuing, particular attention is paid to methodological controls to achieve fair and representative results. Examples include setting up computational environments, preparing datasets carefully, and avoiding data leakage that can lead to misleading outcomes. It highlights the limitations of relying solely on accuracy, especially for unbalanced datasets, and suggests alternative metrics like macro-averaged metrics for a more nuanced evaluation. Additionally, it acknowledges that different types of classification tasks (*binary, multi-class, multi-label*) require specific performance measures. Beyond core metrics, factors like **computational cost**, **latency**, and **efficiency** might also be relevant. Finally, there is an acknowledgement of potential issues arising from data distribution and suggestions regarding the statistical significance of tests to determine if performance improvements are meaningful, emphasising the use of specific training, validation, and testing methodologies to address various scenarios during model development.

- **ISO/IEC 5338:2023**. This standard proposes leveraging existing software and system **life cycle** processes (**ISO/IEC/IEEE 12207 and 15288**) as a foundation for building and maintaining **AI systems**, **defining a taxonomy based on machine learning and heuristic systems.** It acknowledges the need for modifications and additions to accommodate the unique characteristics of AI systems, such as machine learning models that require retraining with new data. Particular emphasis is given to the benefits of integrating AI system life cycle processes with existing practices. This integrated approach improves efficiency, promotes wider adoption of AI, and fosters better understanding among stakeholders as defined in **ISO/IEC 22989**. Moreover, there is recognition that AI systems often combine traditional elements (source code, databases) with AI-specific components. Continuing, there is an aim to define processes and related concepts for the life cycle of AI systems based on machine learning and heuristic approaches, with references to existing standards (**ISO/IEC/IEEE 12207**, **15288**, **22989**, and **23053**) and provisions on the processes necessary to support various stages of the AI system life cycle, including **definition**, **control**, **management**, **execution**, and **improvement**. These processes can be applied within organisations or projects to develop or acquire AI systems. Additionally, the document clarifies that established software and system life cycle processes remain applicable for traditional elements within an AI system.

- **ISO/IEC CD TS 6254.** The standard highlights the importance of **explainability in AI systems**, particularly when they are used **in decision-making processes that impact the lives and rights of citizens**. The complexity of achieving useful explanations for AI system behaviour and the efforts from industry and academia to develop new explainability methods are both acknowledged as properties and occupy a principal role. In particular, the standard clarifies that the overarching goal of explainability is to enhance the trustworthiness of AI systems. However, different stakeholders prioritise specific objectives throughout the AI life cycle (as defined by **ISO/IEC 22989**). For instance, developers seek improved safety, reliability, and robustness by identifying and fixing bugs more easily. Users aim to understand the level of trust they can place in an AI system by uncovering potential biases or unfairness. Service providers need explainability to demonstrate compliance with regulations. Policymakers require this understanding to develop effective policy frameworks that balance societal needs

with innovation. Continuing, there is a clear focus on describing approaches and methods to achieve explainability objectives for various stakeholders regarding machine learning models and AI system behaviour, outputs, and results, offering guidance on applying these approaches and methods throughout the AI system's life cycle. On a final note, a key takeaway concerns the role played by explainability, which appears crucial for building trust in AI systems, with the standard providing a framework for understanding and achieving explainability objectives according to the different needs users may have.

- **ISO/IEC TR 5469:2024.** Moving to the final standard of this analysis, it principally addresses the challenges of applying Artificial Intelligence (**AI**) technologies in **safety-related systems under a functional security lens**. It highlights the increasing use of AI, particularly Machine Learning (**ML**), and acknowledges the difficulty of guaranteeing its performance and explaining its behaviour. The standard continues by emphasising the importance of models in demonstrating the compatibility of AI with safety requirements, including those relative to environmental safety and protection (see standard **ISO/IEC CD TR 20226** - Environmental sustainability aspects of AI systems). Transparent and complete models, based on well-understood scientific relationships, are considered ideal. However, AI is often used in complex scenarios where such models are not available. Additionally, emphasis is placed on the susceptibility of ML models to bias, amongst many technical limitations, which is tied to the undermining of safety verification and validation efforts. The overall purpose of the document is to equip developers with the knowledge to safely integrate AI technologies into safety functions, providing information on functional safety and its relation to AI, different AI technology classes and their potential compliance with existing functional safety standards, relative functional safety risks associated with different AI technology and usage level combinations, a **three-stage realisation principle** for using AI in safety-related systems, where existing safety standards cannot be directly applied and finally potential solutions from verification/validation, control measures, processes, and methodologies. Concluding, annexes to the standard provide additional details and examples, including:

    - How to apply **IEC 61508-3** (functional safety of electrical/electronic/programmable electronic safety-related systems) to AI technology elements
    - Examples of applying the **three-stage realisation principles** and defining properties
    - Detailed processes related to risk mitigation
    - Mapping between the safety life cycle in **IEC 61508-3** and the AI system life cycle in **ISO/IEC 5338**

Certification Approach Considerations. The overall **scope** of standards analysed appears to be ensuring the **safe** application of Artificial Intelligence (AI), particularly Machine Learning (ML), in aviation systems, which would be the scope of the ISO\IEC certification approach. In particular, the **focus appears to be on product quality**, which is a point of critical focus and is directly tied to the robustness, reliability and safety of AI and automation-based systems. The standards outlined above target several **key goals** to achieve these objectives, such as laying down stringent functional safety requirements (**ISO/IEC TR 5469:2024**), enhancing the explainability of AI systems (**ISO/IEC CD TS 6254**), and finally improving the overall quality of AI systems for aviation (**ISO/IEC 25000**). This encompasses not just functional safety and explainability but also aspects such as reliability, maintainability, and security. By adhering to these standards, the civil aviation domain sees the deployment of safe and trustworthy automation as a principal goal, while further reaching goals of explainable, quality models.

All of the above is supported by the already tackled standards, which both leverage existing industry best practices in automation, and address AI-specific challenges. On the second end, the focus appears to be on performance evaluation and functional safety, as technical standards are closely tied and in line with the scope of the approach. The **actors** identified as involved or responsible in carrying out the approach are developers, users (intended as those interacting directly with the AI), and aviation organisations, both from a management and technical perspective. Policy-making authorities, producers, and third-party organisations are identified as some of the principal actors involved in both ends of the approach, that is, **innovation** and **enforcement**. It is important to underline how effective communication and collaboration between these agents are essential for the successful implementation of the approach in discussion and how provisions of document preparation, testing and validation, as well as product monitoring, support such an interaction. Above all considered, a last consideration on **harmonisation** and **record-keeping** follows. Given how the approach partly relies on existing best practices and standards, it could be considered adaptable and capable of supporting innovation with an appreciable degree of flexibility. Paired with attention to record-keeping, considerations of transparency and data governance also appear in line with an approach that takes documentation and harmonisation into account, as well as their correlation.

2.11 IEEE Cross-industry standards for Software and AI

The Institute of Electrical and Electronics Engineers Standards Association (**IEEE SA**) is an operating unit within **IEEE** that develops global standards in a broad range of industries, including artificial intelligence systems, learning technology, robotics, automotive and transportation. **IEEE SA** has developed standards for over a century, through a program that offers balance, openness, fair procedures, and consensus. Technical experts from all over the world participate in the development of IEEE standards.

In the past few years, IEEE has published some standards in the field of AI and autonomous systems:

- **IEEE 7000-2021 – Standard for Model Process for Addressing Ethical Concerns During System Design**: It establishes processes for organisations to integrate ethical values into all stages of concept exploration and development. It supports transparent communication between management, engineering teams, and selected stakeholders to elicit and prioritise ethical values. The standard facilitates the traceability of ethical values throughout the operational concept, value propositions, and system design. It describes processes for tracing ethical values in the concept of operations, ethical requirements, and ethical risk-based design. This standard is applicable to organisations of all sizes and types, regardless of their life cycle models.
- **IEEE 7001-2021 – Standards for Transparency of Autonomous Systems**: It outlines measurable and testable levels of transparency, enabling objective assessment of autonomous systems and determination of compliance levels.
- **IEEE 7002-2022 – Standard for Data Privacy Process**: It defines requirements for a systems/software engineering process focusing on privacy considerations in products, services, and systems that utilise personal data of employees, customers, or other external users. It targets organisations and projects involved in the development and deployment of such products and systems. The standard offers specific procedures, diagrams, and checklists to facilitate conformity assessments of privacy practices. Additionally, it outlines privacy impact assessments (PIAs) as a tool for identifying necessary privacy controls and confirming their implementation.

- **IEEE 7003-2023 – Standard for Algorithmic Bias Considerations**: It outlines processes and methodologies to mitigate bias in algorithm creation. It includes criteria for selecting validation data sets to control bias quality, guidelines for establishing and communicating application boundaries to prevent unintended consequences, and suggestions for managing user expectations to mitigate bias stemming from misinterpretation of system outputs.

In the field of AI and advanced automation, the IEEE SA has launched the **IEEE's CertifAIEd program** for assessing the ethics of **Autonomous Intelligent Systems (AIS)**. The resulting certificate and mark is supposed to demonstrate the organisation's effort to deliver a solution with a more trustworthy AIS experience to their users.

From a substantive standpoint, **IEEE's *CertifAIEd*** program covers four key areas:

- **Transparency** criteria relate to values embedded in a system design, and the openness and disclosure of choices made for development and operation.
- **Accountability** criteria recognise that the system/service autonomy and learning capacities are the results of algorithms and computational processes designed by humans and organisations that remain responsible for their outcomes.
- **Algorithmic bias** criteria relate to the prevention of systematic errors and repeatable undesirable behaviours that create unfair outcomes.
- **Privacy** criteria are aimed at respecting the private sphere of life and public identity of an individual, group, or community, upholding dignity.

In relation to each area, the **IEEE CertifAIEd** offers some "*Ontological Specifications*", as a first level of insight into the criteria published under a Creative Commons BY-NC-ND 4.0 licence. These are extracted from the comprehensive IEEE licensed material that includes the details on the several hundred criteria.

Interestingly, the above-mentioned criteria are presented as being compatible with upcoming regulations such as the EU AI Act.[9]

Procedurally speaking, the ***IEEE CertifAIEd*** Ecosystem includes:

- **Trainers**, offering in-depth knowledge about the criteria and methodology and the ability to address a skills gap in AI Ethics certification.
- **Assessors**, which must attest to the expertise and credentials for delivering AI Ethics assessment, that flexibly integrates into the assessors' own services.
- **Certifiers**, which independently corroborate an organisation's assessment of its AIS under the given *CertifAIEd* criteria using a detailed process that can complement their own offering.

2.12 Safety assurance objectives for autonomous systems

A Safety of Autonomous Systems Working Group (SASWG) of the Safety Critical Systems Club (SCSC) identified a series of safety assurance objectives for autonomous systems, which include AI-based

[9] https://standards.ieee.org/products-programs/icap/ieee-certifaied/

systems (SASWG, 2024). These objectives are structured along three hierarchical levels, from high to low levels:

- At the **platform level, the behaviour of the autonomous system is specified,** as well as the relations with interacting items, people, and the environment. Examples of objectives at this level are: acceptably safe behaviour for the platform is defined; the specified behaviour is safe in the presence of faults and failures, as well as foreseeable misuse and abuse; operational monitoring is sufficient to identify and support the mitigation of new hazards, including emerging cyber security threats; unavailability or unreliability of interacting items does not make the platform unsafe; safety-related demands on people interacting with the platform are reasonable; suitable interfaces are provided for people that may interact with the platform; the platform is appropriately protected against harm from adversarial actors; elements of the environment relevant to the safe operation of the platform are identified and understood; situational awareness of the platform's environment is maintained.

- At the **autonomy architecture level, it addresses how computations are integrated into** a system, describing the faults and failures that it must tolerate, the information that it must maintain and provide, and the changes that it must allow during its operational life. Examples of objectives at this level are: operational inputs inconsistent with training inputs are tolerated; adversarial attempts to disrupt the computation are tolerated; incorrect computation outputs are tolerated; relevant information is presented to interacting parties; relevant information is preserved to support post-incident analysis; computation behaviour is appropriate before, during and after an adaptation.

- At the **computation level**, the implementation in software and hardware is addressed, including algorithms and data used for machine learning. Examples of objectives at this level are: data is acquired and controlled appropriately; training data pre-processing methods do not introduce errors; data captures the required algorithm behaviour; performance boundaries are established and complied with; the test environment is appropriate; the algorithm's behaviour is explainable; post-incident analysis is supported; the software is developed and maintained using appropriate standards; hardware misbehaviour does not result in incorrect outputs from the algorithm.

2.13 Process/property/risk-based approaches for Certification of AI

Current certification methods can be considered "process-based certification", where the manufacturer, in consultation with the competent authorities, defines the steps to be taken in the certification of new aircraft and systems. In the White Paper *Machine Learning in Certified Systems* (Malamet et al, 2021), this is described as follows: "the completion of a predefined development assurance process during the development of a product is the assurance that this product complies with the requirements laid down in the certification basis. Most of the standards currently used for certification are "process-based".

Demonstrating the safety and correct functioning of systems and elements is through specification, verification, safety assessment, etc.

In dealing with AI, certification can be split into:

- Check input data;
- Check the software (inference engine);
- Monitor the results.

2.13.1 Check input data

AI-systems work with a separation of input data (knowledge) and the inference engine that processes the data. In e.g. rule-based systems, the knowledge is presented to the system on the forehand, while in machine-learning applications, the knowledge is constantly updated with newly learned information.

The White Paper *Machine Learning in Certified Systems* suggests the certification of the dataset. For this, a Dataset Requirements Plan & Dataset Verification Plan need to be set up and agreed with the competent authorities.

Verification of the dataset must eliminate errors and biases. Verification and validation techniques for AI may include various approaches, such as testing the AI model against representative datasets, conducting simulations or experiments to assess its performance, analysing the model's decision-making process, and ensuring that it operates within acceptable bounds (Paraskevopoulou, 2023).

Bhattacharyya et al (2015) mentions requirements capture as one of the most difficult challenges in adaptive AI-systems. In particular, the proof of completeness is extremely challenging as the resulting system keeps changing. Model-checking techniques do offer the possibility to address non-linear behaviour and developments in the technique for verification purposes are ongoing.

2.13.2 Check inference engine

Property-based certification. The white paper *Machine Learning in Certified Systems* (Malamet et al, 2021) considers that a number of HIgh Level Properties (HLPs) need to be demonstrated as being safe. The white paper identifies seven HLPs, which are classified as probabilistic assessment, resilience, specifiability, data quality and representativeness, explainability, robustness and verifiability.

Property-based certification concerns the demonstration by the applicant that a predefined set of properties is met by a product is the assurance that this product complies with the requirements laid down in the certification basis. In this approach, the HLPs may be some (or all) of the properties to be demonstrated. Even though there is a growing interest for this approach, one of the difficulties is to prove that the selected set of properties completely covers the desired certification objectives. A clear consensus on an acceptable set of properties has not been reached yet, but “property-based” certification is in line with the Overarching Properties initiative, and it remains an option considered for the future.

Risk-based or failure-oriented approach. Another approach is the risk-based or failure-oriented approach, where the risk of hazardous operational situations is qualitatively assessed and safety measures are defined to avoid or control systematic failures and to detect or control random hardware failures, or mitigate their effects. An example is ISO 26262, which is mostly applied for road traffic. Not surprisingly, the EASA AI Roadmap (EASA, 2023) focuses on mitigating the safety risks.

Bhattacharyya et al (2015) mentions that certification approaches based on the development of a safety case for the aircraft (including its adaptive components) would in principle provide more flexibility to use advanced algorithms, demonstrating the safety of the adaptive algorithm by using the most appropriate evidence, while not sacrificing safety.

Risk acceptance principles must be applied (e.g. in accordance with EC Regulation 402/2013) for the safety assessment, according to a risk acceptance principle such as:

- Compliance with a recognised and context-specific code of practice to deal with conventional risks;
- Comparison with a "similar reference system";
- Explicit risk analysis to demonstrate quantitatively or qualitatively that the risk is kept low enough to be acceptable.

The risk-based approach is also applied in the SORA (Specific Operation Risk Assessment) for small unmanned aircraft, where the approach taken depends on the operational risk, divided into ground risk (the chance of hurting people, animals or damaging property on the ground and the air risk towards other airspace users. It must be mentioned that the SORA is a standard for a risk assessment and not for certification.

2.13.3 Monitor results

Bhattacharyya et al. (2015) indicate that online monitoring tools can be used to check the safety bounds and ensure a stable response. One way to implement this is through architectural mitigations.

Simplex architecture. A simplex architecture relies on three smaller, high-assurance functions: a system status monitor, a simpler backup for the adaptive function, and a switching function. During normal operation, outputs from the adaptive function are used by the rest of the system. If the monitor detects that the adaptive function is not behaving correctly (e.g., it has not converged or computed new output before its deadline) or the system as a whole is approaching a state in which the correct behaviour of the adaptive function has not been verified, then the system will switch to using outputs from the simpler backup function. The inherent advantage in this approach is that, due to the architecture design, the safety of the vehicle never depends solely upon the adaptive function. The adaptive function is used during "normal" operating conditions and switched off during "abnormal" conditions when it might not be dependable.

An alternative approach uses a complex adaptive function to recover the vehicle in the case of a catastrophic failure or upset condition. In this case there is a conventional system that is used during normal flight operation, and a high-assurance monitor and switch that only invokes the adaptive system when the vehicle would otherwise be destroyed. The function of the monitor is to guarantee that the adaptive function is never used during normal operations. The adaptive function is switched off during "normal" operating conditions and only switched on during "abnormal" conditions (when the vehicle would be lost anyway).

2.13.4 Alternative approaches

One approach to certify software or a system is to provide a licence for correct functioning. Similar to training pilots and air traffic controllers, adaptive systems could be trained and approved in a series of

hundreds of hours and then tested extensively. The focus of the method becomes more on proven performance than on the development process and evidence of compliance. Certification would be eventually attained through extensive, though not exhaustive, demonstration of knowledge and skill by the advanced software systems.

In the use of AI for autonomous vehicles, this approach is considered as described in Dia et al (2021). The article introduces several skill levels of the software and would allow it to be used under different, well-defined conditions initially, while after some proven skills, the licence can be extended to something like a Graduated Licence, Level A.

EUROCONTROL (Eurocontrol, 2020) also mentions the approach of certification through licensing. A system encapsulation AI-based software might, for example, be required to go for hundreds of thousands of simulated hours, and thousands of real hours, encountering thousands or millions of faults and contingencies, demonstrating competency far beyond what any human could possibly show in a lifetime.

One key problem with a licensing approach is that any test-based evidence of acceptable behaviour may be completely invalidated by a change to the system.

2.14 Emerging certification approaches for road transportation

2.14.1 Introduction

In 2021, the Committee on Transport and Tourism of the European Parliament provided an opinion (EPCTT, 2021) on the proposal for the AI Act. In its opinion, the Committee recognizes the impact of AI on the transport sector, which requires the “highest level of safety.” Specifically, road, rail, and maritime sectors are referred to next to aviation as the representing transport sector as the Union law regulates the compliance of the mentioned sectors with safety requirements. Therefore, a comparison of those sectors may be helpful.

Nevertheless, not all sectors share the same level of advancement in certification or homologation as casually referred to in these sectors. For road transport, thanks to the rather widely-used autonomous driving system on the road, the certification (homologation) is more researched and advanced. The other two sectors, rail and maritime transport, are somewhat less studied areas in terms of the certification of AI. Therefore, certification methods for AI in road transport are briefly visited in this section, as they may be comparable to aviation.

2.14.2 Main features

In road transport, AI is used foremost for “autonomous means of transport” to reduce the risk of road accidents (EC, 2019) and monitor real-time traffic for better traffic management. The in-depth analysis of the European Parliamentary Research Service (EPRS) depicts that autonomous driving is supported by infrastructure (EPRS, 2017).

According to EPRS, autonomous driving requires the following AI applications:

- driver assistance
- partial automation

- conditional automation
- high automation
- complete automation

As to the traffic management, the following may apply:

- Intelligent Transport System (ITS)
- AI for journey planning and optimisation

The certification methods for autonomous transport receive more attention in the EU than those for traffic management (EP, 2017). Based on these features, the following sections present existing frameworks, if any, relevant to the certification of AI applications in the road transport.

2.14.3 Autonomous means of transport

According to a report of the Coordination of Automated Road Transport Deployment for Europe (CARTRE, 2018), 20 out of 27 EU Member States certify the autonomous means of road transport through testing. Each country has specific regulatory frameworks and requirements for testing vehicles. Typically, they include detailed safety assessments, insurance coverage, and compliance with local road safety laws. The main focus remains that the testing does not compromise public safety and that the vehicles meet technical and operational standards required by each country's road transport authorities. Among others, the following countries have approaches which are distinctive:

- Austria: an applicant must provide detailed vehicle and driver information, proof of information, and specifications on testing period and road sections. A code of conduct must also be acknowledged. Currently, test use cases are developed and provided by the government.
- Belgium: an extensive procedure including a risk analysis, training plans for testing drivers, and a communication with relevant authorities is required. There is no fixed use case or conditions for testing.
- Czech Republic: the safety validation is the responsibility of the public authority. The application to the testing should contain the system for verifying the behaviour of the tested vehicles as a whole. No standards for such are provided by the government.
- Denmark: the safety validation relies on the independent assessor and technical service by means of expert reports. No standards for such are provided by the government.
- Estonia: the safety validation relies on the public authority. No standards for such are provided by the government.
- France: the application should be accompanied with comprehensive documentation describing objectives and safety measures of the prototype being tested, and receive authorization from three different ministries.
- Germany: testing requires approval from each function of the vehicle that is not yet permitted by law, involving an exemption approval from the federal motor transport authorities.
- Italy: two distinctive procedures exist. One focuses on both individual vehicles with temporary licensing plates and comprehensive systems involving vehicles, control systems, and infrastructure.
- The Netherlands: the application process is rather structured. It includes safety assessments and coordination with the Netherlands Vehicle Authorities.

- Norway: application must include detailed risk assessments, vehicle documentation, and data handling strategies.
- Spain: applications need to register as vehicle developers, report tests for Level 2 of the Society of Automotive Engineers and ensure that all testing is accompanied by continuous data recording.
- Hungary: in the testing process, stakeholders are closely involved starting from the early stage of the development (Joldy et al, 2020). The applicant is supposed to validate the safety of the vehicle by means of self-assessment. The approach directs towards self-certification where the highest-ranking officers of the relevant automotive corporation become liable, which is supported by the regulatory framework for certification.

As demonstrated above, an interesting feature is the responsibility to validate safety before or during testing. The safety validation relies on the applicant, public authority or independent assessors. This is another layer before certifying a vehicle or the system which is not yet harmonised within Europe.

The White Rose University Consortium suggests a framework applicable to the highly automated vehicle in the specific context of the UK (WRC, 2019). It focuses on five elements for a certification framework to cover to ensure safety as follows:

- Types of defects: in such a system, there may be requirements defects, design defects, implementation defects, verification plan defects and safety or reliability defects. A certification process should cover those aspects.
- Road testing for defects discovery: a road testing is essential to test system requirements specific for the operations. The purpose of testing is to identify road hazards, sensor failures, and unusual traffic behaviours which are critical for refining system specifications.
- Hazard analysis: methods like Systems Theoretic Process Analysis (STPA) and Functional Hazard Analysis (FHA) are recommended for understanding potential hazards and ensuring that all possible faults leading to hazardous conditions are identified and mitigated.
- Scenario requirements and verification: the development and use of scenarios to validate and verify the behaviour of highly automated vehicles are particularly emphasised. Scenarios help in testing whether the vehicles can perform safely and as expected under various conditions. The scenario should encode specific traffic situations and driving behaviours to ensure compliance with traffic laws and safety standards.
- Simulation as a verification tool: Simulation plays a central role in verifying the system by reproducing complex driving scenarios and testing the vehicle's responses. The use of digital twins and scenario replay within simulated environments may help identifying and mitigating potential failures.
- Full system testing: the need for testing the full system stack, including sensor models, vehicle dynamics, and human-machine interaction, are highlighted. This is a rather comprehensive approach to ensure all components of an autonomous vehicle operate in a harmonised way.
- Continuous improvement and adaptation: the validation and verification processes should be designed to be interactive and adaptive, with continuous updates based on new discoveries and technical advancements to ensure that the tested system remains safe and effective.

Meanwhile, Zoldy et al (2019) also proposes solutions for difficulties rising in the use of innovative technologies such as autonomous vehicles. One of its focuses is the security level of the vehicle rather than the safety level. This aspect is generally connected to the challenges that AI functioning as a safety

critical system entails. By the law of the EU, road traffic must have a human driver in the vehicle at any moment and should be able to take control of the vehicle whenever necessary. In other words, the lessons are applicable only up to Level 3 Automation but not above.

3 Evaluation of Certification Approaches

The objective of this chapter is to review the different approaches collected in Chapter 2 against a set of evaluation criteria that measure their applicability for advanced automation. This chapter is structured as follows:

- Section 3.1 develops the evaluation criteria to be used in the review.
- Sections 3.2 to 3.12 evaluate the certification approaches collected in Chapter 2 using the criteria set out in Section 3.1.

3.1 Evaluation criteria

In determining the most appropriate evaluation criteria, HUCAN drew on the standard KPAs for SESAR projects (SHS), supplemented by the objectives outlined in various frameworks, such as the S3JU Multiannual Work Framework Programme 2022-2031, the European ATM Master Plan 2020, the EASA AI Roadmap 2.0. Beyond the objectives specifically outlined for the aviation domain, HUCAN also considered the Digital Decade Policy Programme 2030 (Decision (EU) 2022/2481) and the EC Strategy 'Artificial Intelligence for Europe' (COM/2018/237 final). In addition, input from the preliminary objectives and guidelines of the S3JU, as outlined in the pre-read material of the European ATM MP Stakeholder Consultation Workshop of 8 April 2024, was taken into account. In light of this synthesis, the following criteria are used to analyse the approaches under consideration.

This section develops evaluation criteria for reviewing the different certification approaches collected in Chapter 2 related to their applicability for advanced automation. The baseline for the development is the list of constituting elements of a Certification Approach as defined at the beginning of Chapter 2. For each element, a number of evaluation criteria has been developed. Some criteria may overlap in terms of requirements against which a certification approach can be assessed. For example, systems' explainability may be relevant in *uncertainty* (lack of explainability may increase uncertainty), in *human factors* (explainability may facilitate human oversight and human-AI teaming), and in *technical complexity* (the level of explainability may require various levels of technical expertise to understand the system). Definitions of the criteria follow below:

- **Uncertainty.** A robust certification approach should account for the inherent uncertainties in various key aspects, including the technology itself, the data used, operational scenarios, environmental factors, and unforeseen behaviour in the context of autonomy and automation. This evaluation goes beyond assessing if the approach considers basic uncertainties and component failures and is particularly critical when considering the highest levels of automation and the relationship of all of the above with accountability. It also assesses how the certification approach facilitates the development of contingency plans for unforeseen events, major failures, or security breaches.
- **Safety.** Evaluate the effectiveness of the certification approach in supporting comprehensive risk control strategies. Posing the focus on safety management should facilitate robust feedback mechanisms to learn from operational occurrences involving advanced automation, as well as tackle technological safety tools in the strict sense. This includes identifying suitable indicators that effectively capture potential risks and dangerous autonomous or automated behaviour. The certification approach should support integrated risk management practices,

encompassing not only safety but also security-related interfaces for key performance areas like environmental, service-oriented and organisational security. This evaluation should consider the level of detail provided by the safety risk assessment, including the types of qualitative or quantitative results generated and the means of compliance included.

- **Accountability.** This evaluation criterion examines the effectiveness of the certification approach's accountability framework. A robust approach should clearly define a framework that assigns clear responsibilities and obligations to stakeholders throughout the civil aviation value chain. This framework should be designed to incentivize the spontaneous adoption of certification measures and ensure ongoing compliance with established safety and security standards. The evaluation should assess the level of discretion granted to stakeholders in implementing the framework. It's crucial to strike a balance between flexibility and ensuring a consistent level of safety across the industry. Furthermore, the evaluation should identify the primary entities held accountable for adherence to the framework and explore how accountability is distributed across the value chain. A well-defined approach will explicitly delineate accountability for different stakeholders involved in the design, development, operation, and maintenance of aviation systems.
- **Environmental Protection.** Assesses the certification approach's capacity to support the reduction of air travel's environmental footprint. An effective approach should address key environmental concerns associated with air travel, including mitigating climate change through CO2 emission reduction strategies, minimising aircraft noise pollution, and safeguarding local air quality around airports. International organisations establish environmental standards that member states translate into national regulations. This evaluation focuses on how effectively the certification approach fosters the adoption, consideration, or implementation of these established environmental standards.
- **Public Oversight.** Measures the extent of democratic control over the organisations, procedures, and enforcement mechanisms associated with the certification approach. It acknowledges the inherent tension between delegating certification activities and duties to private entities or non-traditional public bodies (across member states) and the need for effective public oversight. The evaluation considers concepts like "thirdness" (independence from industry or government) and potential biases within the oversight structure. Furthermore, it assesses the level of public participation in the certification process and transparency surrounding the certified products (technologies, systems etc.). A well-designed approach should ensure that public interest is served through robust oversight mechanisms and opportunities for public engagement.
- **Efficiency.** This criterion evaluates the overall efficiency of the certification process facilitated by the approach. This includes assessing the expected total completion time for technology certification. A well-designed approach should strike a balance between fostering innovation and establishing clear regulatory frameworks. It should ensure a level of rigour necessary to maintain safety without unnecessarily hindering the pace of technological advancement and production.
- **Technical Complexity.** Evaluates the level of knowledge and experience necessary to understand the certification approach, utilise it correctly, and interpret its results. This includes the explainability of the approach, ensuring transparency and clarity in its application. Additionally, the evaluation considers the complexity of tools required to utilise the approach. An ideal approach would be accessible to a reasonable range of experts within the field, utilising tools that are efficient and do not necessitate excessive computational resources.

- **Human Factors.** This criterion evaluates how effectively the certification approach considers human factors in interaction with advanced automation. A well-designed approach should account for the various ways humans will interact with the technology, encompassing considerations like human oversight and human-AI teaming strategies. The evaluation should assess how well the approach facilitates the development of comprehensive training programs for personnel. These programs should equip personnel with the necessary skills to effectively collaborate with advanced automation, while fostering a strong safety culture. This includes promoting practices that discourage overreliance on the system, encourage the reporting of issues, and emphasise situational awareness.
- **Data Governance.** Assesses the certification approach's capacity to establish robust data governance practices. Effective data governance ensures the accuracy, safety, usability, and accessibility of data used within advanced automation systems for civil aviation. This encompasses defining clear protocols for data access control, specifying who can access what data under specific conditions. The approach should also address data storage and usage practices, ensuring data integrity and adherence to relevant regulations.

In the following sections, we review the innovative certification approaches collected in Chapter 2 against each of the above-mentioned criteria. If a criterion is not applicable to an approach, or if there is insufficient information about the approach to evaluate the criterion, this is indicated.

3.2 Ethics guidelines on Trustworthy AI

The **Ethics Guidelines on Trustworthy AI** implement a series of fundamental requirements, principles and methodologies to achieve the safe, robust, fair and ultimately trustworthy research and development of AI solutions. All of the above is contextualised within key priorities, which circle around the notions of **ethical**, **lawful** and **robust** AI.

Human Factors. Amongst the various elements present in the approach, human factors seem to be prioritised, with ethical considerations such as respect for people's autonomy and agency, as well as social wellness as a whole, being critical and repeatedly present. Moreover, explainability occupies a principal role in the approach, a feature of AI directly tied to human awareness and trust. Continuing, trust itself is additionally fostered by minimising uncertainty and supporting safety. In particular, robust systems capable of being lawful, fair and non-discriminatory are one of the main goals of the approach, which explicitly tackles human factors.

Safety and Uncertainty. The ethics guidelines draw a series of principles from the domain of human factors susceptible to influence safety and uncertainty considerations. In particular, there is a clear focus on the creation of systems which mitigate or prevent human harm by design. Additionally, attention is posed on transparency as a critical value and guiding principle, an aspect tied closely to matters of safety and the reduction of unforeseen damages, coordinating with the notion of lawful and technically robust AI, with the latter an enabler of secure automation.

Accountability and Public Oversight. All of the above is framed within considerations of accountability that focus on the implementation of a by-design methodology, as well as architectures which take human oversight and agency into account. Crucially, accountability in and of itself is one of the requirements for trustworthy AI, according to the approach. Still, a critical point to underline in the evaluation of the approach is that it is a soft law act focusing on guidelines and principles. As a matter of fact, although stating requirements for AI research and development, from an enforcement, public

oversight and administrative governance perspective the ethics guidelines for trustworthy AI appear to be distant from the effectiveness shown in other approaches, though nevertheless authoritative.

Data Governance. Although capable of influencing data governance practices, the approach does not include direct references nor points of analysis.

Environmental Protection. A critical lack of attention to environmental protection can be highlighted in the ethics guidelines, particularly considering the importance of sustainable AI in the automation discussion.

Efficiency and Technical Complexity. As detailed above, the guidelines lack a focus on efficiency of implementation and technical complexity.

Conclusion. The approach, therefore, is to be taken into account as an authoritative source of priorities and principles in the domain of AI when tackling its certification within the context of advanced automation, with many guidelines being susceptible to finding application. Nevertheless, the approach in and of itself cannot be the sole basis of a certification framework due to its nature, rooted in soft law and focused on principles, architectures and methodologies.

3.3 The AI Act

As said in Section 2.3, the AI Act's approach to certification largely builds on the **New Legislative Framework** actors and procedures. In particular, the AI Act:

- **introduces** a new certification procedure if a high-risk AI system is a biometric system and/or the provider has not applied, or has applied in part, harmonised standard pursuant to Article 40 of the AI Act. The new certification procedure is largely modelled after the product safety approach, thus the New Legislative Framework ecosystem provides a reference benchmark.
- **integrates** the certification procedures contained in the New Legislative Framework's regulations, if the AI system is a component of a product or it is itself a product regulated under Union harmonisation legislations listed in Section A of Annex I. In particular, the AI Act requires that 1) the conformity assessment on essential requirements of high-risk AI systems follow the one established under those legal acts; 2) that the conformity assessment body competent under those legal acts takes into consideration technical documentation produced by the AI provider; 3) that the same conformity assessment body complies with some requirements on notified body established by the AI Act.[10]

For standalone systems, i.e. those covered in Annex III, except for biometric systems, the AI Act instead foresees an internal control procedure, as long as the provider has applied harmonised standards under Article 40. The latter provides for a presumption of conformity and states that the involvement of a third-party body is not necessary. Thus, in this case, a proper certification procedure for the AI system is not required. On a policy-making level, one might argue whether this option is adequate, given the impact of systems included in Annex III on safety, health and/or fundamental rights. However, this is beyond the scope of this analysis. In the remainder of the section, we shall not consider standalone systems.

[10] Article 43(3) AI Act.

Before delving into the analysis, it is important to notice that the AI Act's certification approach is under construction. A large role will be played by European Standardisation Organisations (ESOs) in drafting harmonised standards. On 22 May 2023, the European Commission issued an implementing decision on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence. The request covers standardisation deliverables concerning various aspects of the AI Act, including the risk management system, data governance and dataset quality, record-keeping, transparency, human oversight, accuracy, robustness, cybersecurity, and quality management. Still, it must be kept into account that the technical standards are yet to be implemented, a factor which will undoubtedly impact the approach.

Data Governance. As said in Section 2.3.2., among the essential requirements of high-risk AI systems, the AI Act requires the systems to rely on appropriate data governance and management practices, which ensure the data is high-quality (**Article 10**). This requirement is reflected in the obligation of providers to put in place an appropriate data management plan, which includes systems and procedures for data acquisition, data collection, data analysis, data labelling, data storage, data filtration, data mining, data aggregation, data retention and any other operation regarding the data that is performed before and for the purpose of the placing on the market or the putting into service of high-risk AI systems[11].

Technical Complexity. The AI Act contains a series of technical definitions which are necessary to appreciate the scope and effect of the regulation, which undoubtedly contribute to heightening the technical complexity of the corresponding approach, especially when considering that the interpretation of such definitions influences the application of the regulation itself. In particular, the AI Act highlights a series of AI applications to be considered as high-risk, with a list containing a broad and varied set of notions of a highly technical nature. Moreover, specific technical values and definitions are set for the governance and regulation of general-purpose AI (GPAI), with the act underlining the dangers posed by similar systems when the computing power used for their training is above 10^25 floating point operations (FLOPs). Finally, as mentioned above, the technical complexity of the act will change once the standards proper are adopted. Hence, considerations in this regard are preliminary and refer to the high-level approach.

Human Factors. The two most critical areas in which the AI Act takes into account human factors appear to be in the context of transparent and explainable AI, deployed under human control and oversight, both aspects which appear crucial for the standardisation following the implementation of the regulation as well as the certification of AI models. Transparency, referring to article 13, focuses on ensuring that deployers and to a certain extent affected persons appreciate how high-risk AI systems function. This includes providing clear instructions regarding the purpose of the system, as well as its capabilities and limitations, with all of the above closely tied to notions of traceability and explainability in AI. Moreover, the regulation ensures deployers are informed about the expected accuracy, potential biases, and any foreseeable factors that could impact the system performance.

Human oversight is required per Article 14, which mandates that high-risk AI systems have human supervision to mitigate their inherent and potential risks. The level of oversight is tailored to the specific system and its context of use and is closely tied to the accountability architecture of the

[11] Article 17(1)(f).

regulation. In particular, the AI Act outlines that those overseeing the AI system should be able to understand its outputs, identify potential biases, and intervene if necessary, being closely tied to the obligations of transparency mentioned above. A critical provision is that which mentions the ability to stop the operation of the system in critical situations. Finally, it must be mentioned how in particularly high-risk scenarios, such as those involving remote biometric identification for law enforcement or, more generally, those susceptible to impact the fundamental rights of affected persons, the Act requires additional safeguards and a heightened degree of human control and oversight.

Environmental Protection. Environmental protection stands as one of the overarching objectives of the AI Act[12], as mentioned in Article 1, relating to the subject matter of the regulation. However, environmental protection is not per se addressed in the certification framework. The risk assessment systems involve only considerations related to safety, health and fundamental rights, where stakeholders are encouraged to tackle and consider sustainability issues. Moreover, environmental incidents and damages are included in the regulation as a case of serious incident, influencing the risk assessment of stakeholders when dealing with systems capable of disrupting the environment, further supporting sustainability goals. An environmental impact assessment was added to the FRIA by the European Parliament amendments. However, this proposal did not make it into the final text of the Regulation.

Accountability. The AI Act approaches accountability as a complex framework, adopting a multidirectional, multilevel, and multiagent architecture in an effort to ensure that the different stakeholders involved in the research, development and deployment of AI are held responsible for implementing their obligations. The core principle of the accountability framework lies in the assignment of different duties to the above-mentioned stakeholders operating within the AI value chain, with the act distinguishing between them by categorising them into developers, manufacturers, importers, affected persons, with each having distinct roles and responsibilities according to their relationship with the AI model, and role within its value chain. In principle, the provider is the one bearing the highest degree of accountability, as it is the one marketing the high-risk AI system under its name or trademark. However, accountability obligations may change along the value chain whenever an actor (e.g., the deployer) modifies the intended purpose of the system.

Continuing, closely related to the multiagent accountability nature of the framework is the concept of human factors. As a matter of fact, the act emphasises the importance of human oversight, control, and personal agency, ensuring humans remain ultimately responsible for AI decision-making, potentially supporting the grounding of liability claims for AI harm and damages.

To achieve this multi-layered, multiagent and human-centric accountability approach, the AI Act sets a series of measures and frameworks into place, pursuing a multidirectional framework, which includes a system of fines for enforcement, the obligation to establish a **quality management system** (Article 17) and follow the lifecycle of the AI model with post-market monitoring measures. On the first end, the act empowers authorities to impose significant fines on actors who fail to comply with its provisions, serving as a strong deterrent against negligent or malicious practices in the domain. Continuing, the AI Act grounds the quality management system as a requirement of providers of high-risk AI systems, which outlines a structured approach to ensuring compliance with regulatory requirements. Finally, the AI Act mandates that providers and deployers of high-risk AI systems

[12] Article 1.

implement robust post-market monitoring mechanisms, intended as ongoing monitoring allowing for continuous evaluation and mitigation of potential risks associated with the AI system, following deployment.

Efficiency. As the AI Act is yet to be effective, and its standards yet to be adopted, no critical considerations regarding its efficiency can be drawn at the time of this report.

Public Oversight. Public oversight in the AI Act is present, yet indirect, as the principal governance levels in which oversight frameworks are implemented are the administrative and the standardisation ones. In particular, control and enforcement of AI Act norms, duties and obligations is primarily addressed towards public organisations, which guarantees an adequate level of public scrutiny over the enforcement and upholding of the regulation. Continuing, stakeholder engagement is considered a priority in the process of standardisation which would follow the adoption and implementation of the regulation, with Article 40(3) stating the goal of providing a balanced representation of interests and the effective participation of all relevant stakeholders. Finally, the possibility of claiming judicial remedy for breach of the regulation, as well as for AI harm and damages, adds an additional layer of public *ex post* scrutiny to the approach. Still, we would underline how the possibility of directly overseeing the process of application of the regulation, as well as that of standardisation, by members of the community and associations representative of social bodies is not immediately considered within the AI Act approach to certification.

Conclusion. The AI Act provides an overarching framework for the certification of high-risk AI systems in the European Union. As said above, the Regulation does not directly apply to AI developed and used in aviation. However, it will serve as a benchmark to amend current sectorial regulations to address issues stemming from the increasing uptake of automation in aviation. The framework of the AI Act largely builds on the Ethics guidelines on Trustworthy AI and attempts to translate ethical requirements into more stringent legal requirements. In terms of matters considered, the AI Act is quite comprehensive since it deals with data governance, transparency both at the deployer and user level, human oversight and human factors, and accountability. It builds governance and enforcement mechanisms which should, in principle, allow for public oversight. At the same time, the AI Act is rather “high-level” and generally framed due to its horizontal and cross-sectorial approach to regulation, meaning that many of the requirements and obligations will need to be specified both through technical standards and authoritative guidelines.

3.4 EASA AI Roadmap 2.0 and guidance for ML applications

The introduction to the EASA approach outlines the AI Roadmap 2.0, which provides a preliminary yet comprehensive overview of AI applications in aviation. A significant aspect of the EASA approach is its principle of proportionality, aimed at tailoring objectives for individual AI applications based on their AI Levels and the criticality of the systems housing the ML models. To implement this principle, the Agency has proposed a classification system comprising three levels of automation, delineating varying degrees of human oversight over AI applications.

Uncertainty. Given the current limitations of the available documentation on EASA's approach to advanced automation, in particular for AI systems at level 2B and potentially higher levels of automation, EASA recognises and addresses two main types of uncertainty: epistemic uncertainty (arising from deficiencies in knowledge or information) and aleatory uncertainty (arising from inherent randomness in the data).

The EASA guidance outlines preliminary procedural elements to iteratively address these possible uncertainties (Anticipated MOC-SA-01-4; Anticipated MOC-SA-01-5; Anticipated MOC-SA-01-6). During the analysis design phase, tasks include failure mode and effect analysis of the AI/ML item, defining performance metrics for AI/ML components, analysing and mitigating exposure effects of AI-based subsystems or components to input data beyond their operational domain, identifying and classifying uncertainties, assigning assurance levels like Software Assurance Levels (SWAL), and defining safety support requirements. Verification ensures implementation meets safety support requirements, ensuring reliability and safety of AI systems. These considerations, though currently limited, apply to preliminary and technical safety assessments, especially where uncertainties exist in learning assurance for transfer learning and surrogate models (Objective SU-02).

Safety. EASA's roadmap identifies safety as the cornerstone of its approach to AI. As previously highlighted [§ 2.4], safety assessment under the proposed new model is integrated into both the overall trustworthiness analysis and technical aspects such as learning safety, data management and information security. Although the objectives and expected means of compliance are currently limited to AI levels 1 and 2, they provide a detailed framework that integrates specific requirements of these new technologies with existing regulatory and certification frameworks. In addition, EASA's approach includes a final AI safety risk mitigation process that recognises the potential limitations of the 'AI black box' and addresses residual risks associated with the inherent uncertainty of AI. Overall, EASA's approach emphasises integrated risk management and safety-related interfaces for key aviation performance areas such as safety and human factors. However, it is less clear how EASA's guidance addresses other issues, such as the economic and environmental impact of AI solutions. While there's some adjustment of compliance burdens through proportionality and system classification, the practical implications of this principle are currently only better understood in relation to AI Levels 1 and 2.

Accountability. In terms of accountability, EASA's approach places significant responsibility on applicants, particularly developers and operators. They are generally invited to comply not only with EASA's guidance but also with the highest available safety standards. Given EASA's role in the current European certification governance system and the importance of its guidance in shaping new certification requirements for AI, accountability is a strong theme aimed at ensuring the spontaneous adoption of certification measures and compliance as a norm of behaviour. At this stage, developers bear the primary responsibility for certification, given their technical and organisational control over design and implementation aspects. In terms of the discretion left to stakeholders, it is currently moderate for AI Levels 1 and 2, with important integrations made in the second edition of the concept paper. However, for level 3 applications, the level of discretion is much higher, with indications currently provided but expected to be further integrated and improved. Generally, accountability duties are intended across the whole product lifecycle.

Environmental Protection. EASA's roadmap approach highlights, in particular, the potential benefits of AI in minimising the environmental impact of aviation operations. One of the main areas addressed in the guidance is environmental protection, which includes specific objectives and new anticipated MOC. From a certification perspective, the protection of the environment is also reflected in the ethics assessment (Gear 6). Currently, the guidance states that the applicant should carry out an environmental impact assessment to identify and assess the potential negative environmental and human health impacts of the AI-based system throughout its life cycle (development, deployment, use, end-of-life). Measures to reduce or mitigate these impacts should be identified. However, it's worth

noting that the competent authorities responsible for verifying compliance and managing environmental risks are identified as national environmental authorities for EMAS, not EASA. This reduces the impact of the results of the ethical assessment from a certification perspective.

Public Oversight. EASA has stressed the importance of a collaborative effort in shaping new certification standards for AI-based applications. The agency aims not only to support its stakeholders by establishing long-term partnerships with industry and collaborating on AI developments through Innovation partnership contracts (IPCs) and memoranda of cooperation/understanding (MoC/MoU) on innovation are the tools that EASA is using to collaborate on innovation with the industry. EASA also aims to involve its employees by gaining practical experience through industry projects and activities. Currently, the section dedicated to use cases aims to assess the practical scope of the guidelines and facilitate participatory rulemaking. Given the rapid evolution of these technologies, such dynamics are likely to become more established in the future.

Efficiency. Currently, the assessment of EASA's guidance for certifying Level 1 and Level 2 AI-based applications presents challenges. The Agency acknowledges that the guidelines have been partially implemented in selected use cases, serving as demonstrators to validate predefined objectives. Specifically concerning assurance levels, varying levels are assigned to AI-based subsystems based on safety criticality and aviation domain. However, operational experience with the guidance remains limited, particularly in supervised learning. Additionally, some anticipated Means of Compliance (MOCs) for challenging objectives, especially for the highest criticality levels, are not yet available. Similarly, MOCs for several challenging objectives in unsupervised learning are less developed. Furthermore, the reduction of assurance levels within AI/ML constituents is presently prohibited, though this restriction may be revisited as more experience with AI/ML techniques is gained. Additionally, requirements for Level 3 applications, which constitute a significant part of the advanced automation solutions investigated by HUCAN, remain uncertain.

Technical Complexity. EASA's approach requires a certain level of technical expertise to be used effectively. The AI Roadmap 2.0 serves as a bridge between the current certification system and the evolving framework for AI-based applications. As such, it requires a solid basic technical knowledge and understanding of AI solutions, with a particular focus on the implications of their implementation in the aviation sector. Several research projects have identified that both the aviation and AI sectors need to take significant steps to align the skills required for the safe and effective use of these new solutions in aviation. In this regard, it is important to emphasise that EASA's work has paid particular attention to the issues of transparency and explainability, carefully considering the different understanding needs of stakeholders involved in the implementation, management and use of AI-based systems. Currently, for Level 1 and 2 systems, it has been determined that the concept of explainability in two views: one related to end-users (operational explainability) and one related to other stakeholders involved with the AI-based system during development or in the post-operational phase (development explainability). Future developments of the concept may be necessary for level 3 applications, especially with respect to development explainability.

Human Factors. EASA dedicates an entire building block of its trustworthiness analysis to the Human Factors (HF) perspective, firmly believing that it is necessary to introduce the necessary guidance to address the specific Human Factors needs associated with the introduction of AI. In this context, a new concept is introduced and explained: human-AI teaming (HAT), which aims to ensure adequate cooperation or collaboration between end-users and AI-based systems to achieve certain goals. From this perspective, the Level 1 and Level 2 AI guidelines currently cover the main aspects related to AI

operational explainability, human-AI teaming, interaction modality and interface style, error management, workload management, failure management and alerting system, and human-AI interface customisation. Taking into account the specificities of level 3, these issues could be revisited and extended, in particular for protected advanced automation, where operators may be allowed to override the authority of the system and intervene ex post with their own initiatives. The adjustments required for the unsupervised actions of full automation should be less significant.

Data Governance. The theme of data quality and data governance emerges as a cross-cutting issue in the AI Roadmap and its complementary concept papers. In terms of learning assurance, it is observed that a correct and complete definition of the Operational Design Domain (ODD) is crucial for ensuring the quality of datasets involved in the learning assurance process (further detailed in Anticipated MOC DA-03 and Anticipated MOC DA-04). Additionally, specific objectives and requirements are explicitly dedicated to data management, covering activities ranging from data collection to data labelling in supervised learning, data preparation, data allocation, data labelling in unsupervised learning, and data validation and verification. The scope of these requirements is further informed by the importance attributed to data governance and management requirements within the ethics assessment (Gear 3), which includes privacy compliance profiles traditionally situated within the realm of security.

Conclusion. An analysis of EASA's approach reveals the benefits of a comprehensive, phased approach with development objectives that is gradually shaping the regulatory landscape, albeit at a deliberate pace. Of particular importance are the Level 2 applications, which have significant safety and human factors considerations in terms of task allocation and responsibility. The guidance provided is instrumental in effectively aligning development processes with initial objectives, supported by anticipated means of compliance. However, significant gaps remain within the HUCAN research scope. Reliable evidence for level 3 applications remains elusive. In addition, a critical aspect warrants scrutiny: the static nature of AI systems and subsystems classification according to the EASA taxonomy. This demarcation shows where the 'AI-level' classification differs from an 'automation' paradigm. Unlike the latter, where levels can adapt dynamically across operational phases or degraded modes, the 'AI level' remains fixed. It reflects the peak capability of the AI-based system, particularly in terms of user interaction or autonomy (especially at AI level 3B). The classification serves as a universal reference across all aviation domains, reinforcing the modulation of AI trustworthiness objectives beyond system criticality. This approach runs the risk of inadvertently downplaying automation-related hazards, potentially requiring design overhauls to meet less stringent certification categories than those for advanced automation.

3.5 Research roadmaps for increasingly autonomous operations

In Section 2.5 insights are presented from an early research agenda for autonomy research in civil aviation (National Research Council, 2014) as well as a more recent roadmap for autonomy verification and validation (Brat et al., 2023). Both reports provide relevant and detailed advice for types of developments that are needed in support of development and verification & validation of (AI-based) increasingly autonomous systems. Being research roadmaps, they do not provide the methods as such.

Uncertainty. Various types of uncertainty are considered in the research roadmaps, including those leading to safety risks, cybersecurity vulnerabilities, and lack of continuous human oversight. It is advised to develop methods to better describe adaptive and non-deterministic systems, since there

may emerge new types of behaviour that can affect stability and robustness of operations as a result of combination of various sources of uncertainty.

Safety. Development of methods for verification and validation to ensure the safety of increasingly autonomous systems is stressed, including high-fidelity test environments, modelling and simulation to assess safety risks and cybersecurity robustness, compositional verification techniques, runtime assurance, contingency planning, and dynamic assurance.

Accountability, Environmental Protection, Public Oversight, and **Efficiency** are Not considered.

Technical Complexity. There is a strong impetus in the research roadmaps for the development of expertise, methods and tools for development and analysis of increasingly autonomous systems, including behaviour of adaptive/non-deterministic systems, modelling & simulation, compositional verification techniques, and runtime assurance approaches.

Human Factors. More research on the roles and interfaces of personnel and systems is stressed, ensuring effective communication, situational awareness, trust, and intuitive interaction, especially in high-stress, dynamic situations.

Data Governance. Not relevant.

Conclusion. Key insights from the research roadmaps are that there is a need for new methods and tools on behaviour of adaptive/non-deterministic systems, modelling & simulation, compositional verification techniques, and runtime assurance approaches.

3.6 FAA roadmap and methods for AI Safety Assurance

As explained in Section 2.6, the FAA and stakeholders are developing a roadmap for AI safety assurance and conducted supporting workshops. As part of this endeavour a number of methods have been discussed, but it is not a complete approach for certification yet. Key insights along the evaluation criteria are discussed next.

Uncertainty. The main focus of the AI safety assurance approaches, which have been considered, regard failures to achieve the intended behaviour of an AI system (e.g. automatic recognition of a runway), incorporating uncertainty in foreseeable operating conditions, such as sensor errors, environmental conditions (e.g. lighting conditions), and component failures. These are the types of uncertainty that are most in line with existing safety assessment approaches for certification. Other types of uncertainty are considered to a smaller extent. This regards dealing with crises or major failures: how can this be achieved effectively if human operators are largely or completely out of the loop? Interestingly, as described in Section 2.6, the HASS developed a framework that includes AI contingency management. It is known that uncertainty exists due to security attacks, but systematic approaches to deal with them have not yet been developed. Uncertainty due to the possible large variability in the interaction of an AI system with humans is also not considered in detail.

Safety. The FAA AI safety assurance roadmap and methods development is mainly focused on safety risk assessment and safety assurance approaches for the design of AI-based systems. The methods build upon existing approaches for design assurance levels. It is recognized that validation and verification may require enhanced modelling and simulation approaches. The safety assurance for design is complemented with AI operation time assurance (OTA) and AI contingency management

(CM) in a framework developed by HASS. OTA monitors AI components and their enabled systems to detect any violations of operation time requirements and adjusts their functioning within safe limits. CM addresses situations that cannot be addressed by the AI and OTA, enabling last resort measures, e.g. a contingency landing. There has been little focus on methods for safety risk control (e.g. learning from occurrences), and the impact of security attacks on the safety of operations. Also the role of human factors on the safety risk in operations with AI-based systems has been considered to a small extent only.

Accountability. Existing certification approaches will be followed as much as possible. Possible changes in the accountability framework, for instance towards higher accountability for system developers, have not been considered in detail.

Environmental Protection. Means for environmental protection have not been specifically considered in the AI safety assurance development.

Public Oversight. Public oversight has not been addressed specifically in the developments. It is considered that the ethical use of AI is outside of the scope of safety assurance. Existing certification approaches will be followed as much as possible. This indicates that current approaches for public oversight will be maintained, including the involvement of governmental organisations (FAA), standardisation organisations (like RTCA), industry and universities. There is a risk that the lack of dedicated attention for public oversight may lead to a decrease in appropriate public oversight, especially if sophisticated AI methods are employed for which sufficient expertise is lacking at the certifying entities.

Efficiency. The development of new methods in support of certification is at a too early stage to evaluate their efficiency.

Technical Complexity. A variety of approaches in support of certification are being considered. These can be expected to require detailed specific expertise and dedicated tools. Management approaches trying to reduce technical complexity have not been considered extensively at this stage.

Human Factors. The emphasis in the developing approach lies on the evaluation of the (AI) software of technical systems, rather than on human factors in their application.

Data Governance. Data governance rules and practices for accuracy, suitability, sufficiency, and accessibility of data have not been considered.

Conclusion. Key insights from the FAA AI assurance roadmap are that validation and verification is likely to require enhanced modelling and simulation approaches for safety assurance during design, and that this should be complemented by AI operation time assurance (OTA) and AI contingency management (CM).

3.7 NASEM and HFES certification approaches

In Section 2.7 a number of studies are described, which focus on human-AI teaming and which effectively include human factors in the development of advanced technical innovations. These studies mainly describe human factors issues in working with advanced AI-based systems and they propose research objectives for future studies. Furthermore, the HFES developed a HRL scale which can be used

in conjunction with the TRL scale for assuring proper consideration of human factors in the development of AI-supported operations.

Uncertainty. As these studies focus on human-AI teaming, they consider uncertainty and variability in interactions between humans and AI-based systems. This type of uncertainty may not be sufficiently considered in more technological-focused safety assurance approaches.

Safety. The research objectives and the HRLs in the studies are intended to support the safety of the operations involving human-AI teaming. They mostly do not provide a basis for safety risk assessment of such operations, however.

Accountability, Environmental Protection, and **Public Oversight** are Not considered.

Efficiency. The application of the HRL scale is intended to support the efficient introduction of advanced systems by proper consideration of associated human factors. Specific methods for human-AI teaming need to be developed, their efficiency cannot be judged now.

Technical Complexity. Consideration of human factors in certification of human-AI teaming operations and technology requires appropriate human factors expertise. Other needed expertise and tools can be diverse and depend on the type of methods applied in the human-AI teaming assessment, e.g. human-in-the-loop simulation facilities, human measurement equipment (EEG, eye tracking, etc.), human behaviour models.

Human Factors. The key focus of the studies is on human factors in operations with advanced (AI-based) automation, addressing human-AI models and processes, maintaining suitable situation awareness, achieving appropriate transparency and explainability for humans working with AI-based systems, achieving effective interaction in human-AI teams, obtaining suitable trust in AI systems, avoiding biases in decision making, achieving proper training for human-AI teams, and assuring human-system integration in coordination with multidisciplinary development teams. All this concerns a wide spectrum, which will require a multitude of methods, expertise and supporting tools. The HRL scale developed by HFES may support the management of the needed human factors studies.

Data Governance. Data governance for human measurement data should comply with ethical rules for data collection, including informed consent, data privacy, data security, and data minimisation.

Conclusion. These studies provide very useful overviews of human factors to be considered in operations with AI-based systems. The HRL scale may be a suitable means for assuring proper consideration of human factors in AI-supported operations.

3.8 EUROCAE and SAE working groups on AI certification

In Section 2.8 an overview is provided for the approaches of EUROCAE and SAE working groups on AI certification focusing on a gap analysis for limitations of the existing standards for the certification of AI (EUROCAE, 2021) and the development of a new standard for certification of aeronautical safety-related products with AI. Currently this new standard has not been finalised yet, implying that the topics below may be addressed differently in the future standard.

Uncertainty. It is recognized that the data-driven approach of ML may lead to new types of uncertainty, which are not sufficiently addressed by current standards. The development of a new standard is focused on managing such data and learning process related uncertainty.

Safety. The EUROCAE and SAE working groups build on current approaches for safety assessment and certification of aeronautical products. In this context they are developing approaches to support development and assurance of ML components in airborne and ATM systems.

Accountability. The standard under development is intended to provide an acceptable means of compliance for approval in line with sectorial and EU regulations in the future.

Environmental Protection. Not considered.

Public Oversight. Standards by EUROCAE and SAE are public and they are developed by their cooperating members, including industry, service providers, regulators and research institutes.

Efficiency. The future standard builds on existing standards. The efficiency of the new approaches for approval of ML-based components cannot be judged at this stage.

Technical Complexity. As for the application of existing standards, considerable expertise is required on the technical systems, including ML-based components, and the operational context of their application. A diversity of tools is used in the development assurance activities.

Human Factors. Human factors are not addressed in specific detail.

Data Governance. Specific emphasis is placed on data management processes in support of machine learning for the development as well as validation and verification of the ML-based applications.

Conclusion. The standard for certification of aeronautical safety-related products with AI by the EUROCAE and SAE working groups is still under development. This development is focused on the assurance of ML-based systems, which is also considered in the learning assurance approach of (EASA, 2024).

3.9 Test and evaluation approach for AI-enabled systems at the US Air Force

In Section 2.9 an overview is provided of test and evaluation challenges in AI-enabled systems at the US Air Force (NASEM, 2023), aimed at the central question: how to achieve sufficient confidence in AI-enabled systems? Key insights along the evaluation criteria are discussed next.

Uncertainty. It is indicated that effective use of AI implies transitioning from a waterfall to an agile development methodology with frequent evaluation and retraining, so as to be able to adapt to changes in operational conditions and emergent threats. An AI RIsk Management Framework (RMF) is used, which includes identification and management of unanticipated performance problems. Red teams are used to try to identify adversarial attacks and weak spots.

Safety. An agile and cyclic development approach is advocated where developers, testers, and users should gain justified confidence in AI-enabled systems over time as they become increasingly familiar with system performance limits and behaviours. The reliability, effectiveness, robustness, and safety of AI systems are evaluated, monitored, and ensured using AI assurance methods to verify their

accuracy and performance, detect and mitigate potential biases, and evaluate their ethical and societal implications. In line with such a cyclic development approach, an AI RMF is adopted for continuous monitoring, assessment, and management of risks in operations with AI-based systems.

Accountability. The study emphasises testing and evaluation towards improving the systems and obtaining justified confidence, rather than certification measures, compliance, and accountability.

Environmental Protection. Not considered.

Public Oversight. Not considered explicitly, but the US Air Force is ultimately controlled by the US government.

Efficiency. An agile and cyclic approach for development and testing of AI-based systems is adopted, where AI assurance processes are used for evaluating, monitoring and ensuring the reliability, effectiveness, robustness, and safety of AI systems. As such the performance for various key performance areas may be managed in line with the strategic needs of the US Air Force.

Technical Complexity. An agile and cyclic development approach requires detailed expertise and tools, including trained labellers; continuous monitoring, retraining, and redeployment of AI models; instrumented deployment platforms to capture ML-ready data; and synthetic data engines and supporting digital twins. Validation and verification may be accomplished in both complex simulation environments and real-world tests. In such an agile approach, product specifications and the risk models are continuously updated through cyclical review. So, the technical complexity is large, since it encompasses both (cyclic) development and evaluation.

Human Factors. It is indicated that human readiness levels (HRL) and UI/UX for AI-enabled systems must be prominent. Measures of performance and effectiveness, including assessments of user trust and justified confidence, must be formulated during system design and development, and assessed throughout test and evaluation and after system fielding. The agile and cyclic approach for development and evaluation provides a basis for users to become increasingly familiar with system performance limits and behaviours, and to provide feedback on the AI systems. This should build upon a suitable culture for risk management.

Data Governance. In an agile and cyclic approach for development and testing, there exists an emphasis on updating data for adaptation of AI models, and on acquiring continuous data on the effectiveness and risks of the use of the AI-based system.

Conclusion. A key insight of this study of the US Air Force is that effective use of AI implies transitioning from a waterfall to an agile development methodology with frequent evaluation and retraining, so as to be able to adapt to changes in operational conditions and emergent threats. In line with such a cyclic development approach, an AI RMF is adopted for continuous monitoring, assessment, and management of risks in operations with AI-based systems.

3.10 ISO/IEC Cross-industry standards for Software and AI

The ISO/IEC cross-industry set of standards focuses principally on ensuring the **quality** of software and AI applications from both a technical and organisational perspective. This approach, centred around AI as a product or service, differs from considerations tied to the role of AI within advanced automation, which sees its implementation as an element of support and acceleration of automation

processes. This primary consideration must be taken into account, as the ISO/IEC approach emerges from a set of standards directly related to AI.

Uncertainty and Safety. Matters related to uncertainty, safety and the technical robustness are of principal importance within the ISO\IEC approach. As a matter of fact, these features of AI and automated systems are tackled both explicitly and implicitly by focusing on the achievement of a life cycle capable of ensuring a high standard of quality and efficiency, following the process of control, management, execution and improvement of the model.

Technical Complexity. The certification approach of ISO\IEC is fundamentally gathered from a set of harmonised standards and principles, which physiologically carry a significant technical complexity, intended both in the strict sense, as per the management of AI life cycles, machine learning assessment performance and XAI methodologies, and the broad sense, with reference to the managerial and organisational implementation of the standards.

Efficiency and Data Governance. The focus on both organisational and technical harmony within the approach, as well as the consensus surrounding the standards themselves and their adoption process, enables an efficient application of the ISO\IEC framework. On a data governance end, all of the above, including matters related to safety and technical robustness, is grounded on organisational training, record-keeping and documentation**,** further supporting the overall efficiency and stability of the approach in data practices, with particular regard to the management of data related to the AI life cycle. Continuing, an additional consideration relevant for both efficiency and data governance refers to the adoption of best practices and industry guidelines as key elements for the development of the approach, which therefore appears flexible and in line with existing private sector practices, and thus successful both formally and substantially in creating a framework capable of harmonisation and effective information management.

Human Factors. The ISO\IEC approach in discussion appears to be extremely technical in nature and primarily AI-focused, critically overlooking fundamental aspects in advanced automation such as the human factor, which is only incidentally tackled through provisions of explainability and a focus on personnel training. The same consideration explains the lack of elements for principal priorities such as enforcement, governance and accountability, as well as general attention to human-centric AI development and research.

Accountability and Public Oversight. These aspects were not considered in the standards and were analysed for the purposes of this document.

Environmental Protection. Sustainability and environmental protection are aspects considered within the approach, with particular regard to the mitigation and management of environmental damages as a security and organisational priority. Nevertheless, no singular standards directly focusing on the environmental sustainability of AI and automation are present in the analysis, except ISO/IEC CD TR 20226 - Environmental sustainability aspects of AI systems, which appears to be at the committee draft level and not currently in force.

Data Governance. These aspects were not considered in the standards analysed for the purposes of this document.

Conclusion. Overall, the approach appears technical in nature and heavily based on standards, operating on a lower level, meaning more technical and less akin to a certification high-level approach,

when compared with EASA or the AI Act. This, however, is perfectly admissible, and in line with the approach as being a derivative of common themes identified within this document from the ISO/IEC standards themselves, which are widely recognised as authoritative and effective within the aviation and AI industries.

3.11 IEEE Cross-industry standards for Software and AI

The IEEE cross-industry standards for software and AI cover a variable scope that generally includes **software capable of achieving high levels of automation and autonomy, including AI**. To ensure proper application of these standards, the IEEE now offers a **certification approach for the ethical aspects of AI through the IEEE CertifAIEd program**. The approach favoured by these standards is value-based design, and therefore they have a primarily engineering-oriented approach.

Uncertainty. The IEEE approach is characterised by open and voluntary adherence to the proposed standards and the pursuit of certification. The risks associated with the development of technologies with high levels of autonomy and automation are a fundamental assumption of this approach. In some cases, such as the standards on transparency and, in the future, on algorithmic bias, certain areas of uncertainty are addressed more specifically and in detail, with recommendations for practical mitigation measures.

Safety. This issue is implicitly addressed. Some references emerge by explicitly addressing the operational, social and legal-administrative functions of transparency. However, as clearly indicated in certain sections of the documents considered, there are specific IEEE standards on safety and security, which are also referenced in terms of ethical implications.

Accountability. Given the nature of the standards, which are inspired by a principle of open and voluntary adherence, the IEEE generally identifies the broad category of 'designers' as the primary recipients of its standards. This includes, developers, builders, maintainers, operators, and decision-makers and procurers in organisations using and deploying systems. However, it is specified that the modes of adherence and the related responsibilities, especially concerning obligations and associated risks, rest entirely with the designers. Given the approach of the IEEE's *CertifAIEd* program, it is assumed that the same applies to the ethical certification, which remains voluntary and does not preclude any concrete determination of accountability.

Environmental Protection. In general, it is explicitly stated that just as safety and security profiles are not within the declared scope of the standards, environmental protection is also not included. However, the guidelines on transparency can help convey the benefits and savings in terms of efficiency and consumption, but these are considered side effects rather than their primary objective. Similarly, IEEE's approach is fundamentally agnostic towards environmental protection, as it assists in identifying the values and ethical requirements to be implemented but does not provide a predefined value framework.

Public oversight. While the method through which standards are adopted is inspired by democratic principles of transparency, consensus, and open collaboration.

Efficiency. The voluntary nature of adherence means there are no explicit guarantees of efficiency and impartiality in their application.

Technical Complexity. The provided material presupposes a certain **level of professional competence**, not only for comprehending the standards but also for their practical application. This aligns with the **engineering orientation that consistently informs IEEE documents**, including those related to value-based design and ethics by design.

Data governance. The topic is explicitly addressed in terms of transparency and privacy by design.

Human factor. There is no explicit reference to the human factor component**.**

Conclusion. The IEEE certification approach for AI and automation in civil aviation seems to emphasise voluntary adherence to value-based design and ethical standards, focusing on key aspects such as managing uncertainty, ensuring safety, assigning accountability to designers, and maintaining transparency. On the other hand, environmental protection, public oversight and human factors are not explicitly guaranteed nor are they a point of focus in the approach. Overall, the IEEE approach requires professional competence to be applied and it is centred around the practical implementation of standards, following a traditional stance towards certification.

3.12 Safety assurance objectives for autonomous systems

In Section 2.12 an overview is provided of safety assurance objectives for autonomous systems as developed by the Safety of Autonomous Systems Working Group (SASWG) of the Safety Critical Systems Club (SCSC) in (SASWG, 2024).

Uncertainty. Objectives are provided for dealing with various types of uncertainty and hazards, including system failures and unavailability, foreseeable misuse, cyber security threats and adversarial actors.

Safety. The report provides a structured set of objectives for the assurance of the safety and security of autonomous systems. It does not define the methods to achieve these objectives.

Accountability, Environmental Protection, and **Public Oversight** are Not considered.

Efficiency. This study concerns a structured set of safety assurance objectives rather than an overall certification process, such that its efficiency cannot be judged.

Technical Complexity. This study concerns a structured set of safety assurance objectives, but specific means and their technical complexity to achieve these objectives are not provided in detail.

Human Factors. Some high-level objectives for interactions with humans are provided.

Data Governance. Objectives for data management in support of machine learning are provided.

Conclusion. This study provides a structured, hierarchical set of safety assurance objectives, which can be useful for the evaluation of the scope of a safety study for autonomous systems.

4 Conclusion

This Chapter concludes by discussing the results of Chapter 3, and analyses which of the approaches (or elements of approaches) are suitable for the aviation domain and which gaps emerge from their implementation. The result will feed into HUCAN WP4, which will use this as input/inspiration to develop a new approach.

Based on the analysis conducted in Chapter 3, several key **gaps and challenges** in the certification of AI and advanced automation for civil aviation have been identified. These primarily revolve around human factors and associated elements such as trust and accountability, which should be taken into consideration in particular, given their impactful role in socio-technical systems. In particular, the following considerations can be made in relation to the criteria analysed in the previous chapter:

- **Uncertainty.** The majority of the approaches give proper weight to uncertainty and a lack of foreseeability in advanced automation. The focus, however, is often on safety considerations in a strict sense and does not include their interconnections with human agency and oversight. The focus tends to be on managing the uncertainty inherent in high automation rather than on a broader uncertainty, covering many aspects of the relationship between humans and technologies, including how humans interact with the technology to make their decisions.
- **Safety.** All approaches prioritise safety and robust automation, adhering to the ethos of traditional certification frameworks, which share the principal goal of ensuring safety and security first. However, as for uncertainty, safety considerations are generally linked to the technical functioning (or malfunctioning) of the system and partially overlook the overall organisational aspects, including human-technology interaction. Nevertheless, there are some exceptions that establish a relationship between ethical and explainable systems to safety, considering the impact of the human factor directly.
- **Accountability**. While accountability aspects are considered in the frameworks analysed, a greater emphasis could be put on this element. Most approaches focus on technical and organisational aspects of certification, leaving accountability frameworks assumed or to be developed by the technology developers/users. This idea follows a traditional view of certification, which relies on strict requirements for technology development. The exceptions identified provide programmatic targets, rather than simply prescriptive measures, which encourage the adoption of a given accountability architecture. When advanced automation and AI are involved in the process, ex-ante one-size-fits-all prescriptive approaches might not consider the most critical features of AI, such as its adaptive behaviour and capacity to act unpredictably. Therefore, it might be adequate to incorporate accountability into the certification approach directly, at least by specifically identifying who is responsible for decision-making and who bears the responsibility of control. More attention should be put in ensuring alignment between accountability models and technical standards.
- **Environmental Protection**. Few approaches incorporate environmental protection principles, standards and norms directly within their frameworks, often referring to external sources or legislative initiatives on the matter. Flagship examples of approaches that do include a discussion on environmental protection are the EASA AI Roadmap 2.0 and the AI Act. The latter, however, while presenting programmatic goals and aims at sustainability, lack specific requirements and enforcing norms devoted to environmental protection. Our analysis

underlines that the problem of environmental protection should be considered not simply at the programmatic level but also at the level of standards.

- **Public Oversight** is partially taken into account. Some certification frameworks have a narrower view on public oversight and provide for stakeholder engagement procedures, such as those leveraging standards formed out of consensus, such as for IEEE and ISO/IEC. Other approaches do not take into account how to include effective stakeholder participation in a more structured way. In a highly technical and technologically flexible domain, the providers of automation technologies are those who hold the most information and domain proficiency. This expertise may be used to suggest the most effective paths to implement certification goals. With regard to oversight in the certification process, the emerging frameworks tend to focus on the standards to be certified rather than on the actors and procedures involved in the certification process. The value of public oversight in the certification process should be given greater importance and scope, not only including stakeholders but also the society at large.
- **Efficiency.** Efficiency of the certification process is scarcely considered as a priority. This lack finds its source in the existing trade-off between safety and efficiency, with safer approaches often being less efficient in terms of speed of implementation and regulatory weight. This consideration requires careful evaluation and a proper balance between safety and efficiency in the certification procedures. The relationship between the human factor aspects and the efficiency of the certification process shall be emphasised as equally important.
- **Technical Complexity.** The level of technical complexity is generally high amongst all approaches, which is in line with the nature of advanced automation as a safety-critical and highly technical domain. This may create a challenge in the certification process and bear relevance in terms of expertise needed by the certification entities to accomplish their task. It also may be seen as an opportunity to foster stakeholder engagement to leverage technical expertise with a view to understanding and managing complexity.
- **Human Factors.** Some approaches consider elements of human agency, explainability, and trust, which embrace a socio-technical view and include aspects of human-technology interaction. In particular, this is the case for the Ethics Guidelines and the AI Act, which stress the need for “human-centred AI” development. Also, the emergence of specific frameworks specifically focused on human-AI teaming is noteworthy. At the same time, many of those frameworks deal with the human component at a high level of abstraction without specifically determining or giving criteria to determine the level of preparedness and fitness of humans to retain human agency and maintain control over automation.
- **Data Governance.** Although authoritative approaches, such as the EU AI Act and the EASA concept paper on machine learning applications duly consider aspects of data governance, the criterion is overall under emphasised, with few other approaches tackle it in a satisfactory way. There is a chance to explicitly address data governance within certification approaches, identifying best practices and potential pitfalls to ensure robust data management and technical resilience in the context of advanced automation. Particularly, when implementing AI solutions, which introduce complications tied to their training, retraining and data life-cycle management.

Considering the analysis above, it appears that emerging certification approaches in the domain of advanced automation for aviation would heighten their effectiveness by taking particularly into account more **comprehensive and dynamic safety frameworks**, **accountability models**, **human factors analysis**, and **state-of-the-art data governance practices**.

In terms of **safety**, more comprehensive plans to achieve a proper certification of AI software are needed considering the following critical elements of AI and advanced automation in civil aviation: (1) **inference algorithm:** this can be certified using conventional methods as it is deterministic in nature; (2) **input/knowledge Base**: this component may be built dynamically (self-learning) or statically. Regardless, it requires extensive testing and certification against rigorous requirements; (3) **system Output Monitoring**: this feature measures the outputs of the AI system, which should be monitored in real-time to ensure security and reliability.

Moreover, we underline how most of the approaches analysed in this document focus on standards and a static view of safety for approval of new systems. An alternative approach to certification could be considered from a safety management cycle, where safety is assured in design as well as during operations. Such an approach would add a **dynamic element to certification within a safety-critical domain,** where human-automation interactions are being implemented at a higher rate and encompass various applications and aircraft types, including passenger planes and drones.

In this regard, the certification ecosystem present in the self-driving car industry offers an interesting parallel. In this context, automated systems learn from thousands of inputs and their behaviour is **continuously monitored**. Additional learning and certification can occur through supervised instructor sessions, as the training processes for pilots and air traffic controllers. This method aims to ensure that the AI system is **consistently improving and adhering to safety standards**, and would support an efficient implementation of licensing drawn from a different safety-critical domain.

By adopting a safety management-based strategy, certification would act in a more **versatile** way, following an approach **closer to governance** frameworks for AI and automation, such as the one in the AI Act.

An additional consideration deriving from the analysis carried out is the potential risks associated with **overreliance** and lack of **human agency**, including **automation bias**, which must be considered as a priority in the context of the human factor analysis. Advanced automation systems, including AI-driven architectures where humans are involved, must be designed to avoid these pitfalls. It is crucial to consider the human element to maintain safety and trust in aviation automation.

Finally, the implementation of machine learning-based models for aviation must be **resilient** and capable of **failing safely** in the face of unforeseen information security threats, posing particular emphasis on the critical importance of efficient and effective **data governance practices** and **cybersecurity**. In this regard, designing robust systems to withstand and recover from attacks, as well as ensuring continued safe operation despite potential vulnerabilities, appears critical, including implementing correct cybersecurity and data processing, management, and storage practices.

Ensuring these elements are integrated is crucial for fostering trust, ethical interactions, and the overall effectiveness of AI systems in aviation, with few yet significant approaches showing there is an opportunity to do so both efficiently and effectively. It is, therefore, imperative that **future certification efforts emphasise these human-centric considerations to achieve a holistic and robust regulatory framework**.

4.1 Summary Table

In this section, we present a summary of the evaluation result of innovative certification approaches conducted in **Chapter 3**, including the concluding remarks elaborated in **Chapter 4**. To do so, we present a table including an **Evaluation Summary** for each of the criteria, a **Status** report clarifying whether the criteria is **satisfied**, **partially satisfied** or **not satisfied** according to our research, as well as **Recommendations** for future shifts and changes in innovative certification approaches, tailored to each criterion and its findings.

Ultimately, what emerges is the presence of a series of patterns and trends in evaluations results, beginning with the **necessity to focus on practical, enforceable and defined standards** instead of abstract principles in certification approaches. Moreover, research highlights the effects and presence of **asymmetry of information** between agents involved in the certification process. In doing so, it **pushes for a broader view of certification** as a process including new areas such as accountability, public oversight and data governance frameworks directly within itself.

See below the **Summary Table** of our findings relating to innovative certification approaches and their evaluation:

Criteria	Evaluation Summary	Status	Recommendations
Uncertainty	Most approaches focus on managing uncertainty in automation for safety purposes, but often neglecting human-technology interaction and decision-making.	Partially Satisfied	Include human-AI and human-automation interactions as uncertainty factors.
Safety	Prioritised as a technical requirement, with scarce connections to ethics and explainability and little focus as an organisational requirement.	Partially Satisfied	Integrate organisational aspects and human factors, broaden the notion of safety.
Accountability	Considered, but often left outside of the approach and not tackled directly. Should be incorporated into the certification process.	Not Satisfied	Include accountability directly within the certification approach.
Environmental Protection	Few approaches integrate the criteria, typically referring to external standards. Should be incorporated into the certification process.	Not Satisfied	Include environmental protection directly within the certification approach.
Public Oversight	Public oversight is inconsistently addressed. Effective stakeholder participation, oversight of certification implementation and attention to actors and procedures require greater emphasis.	Not Satisfied	Ensure structured stakeholder participation and broader actor oversight.

Efficiency	Appears to be often overlooked due to a negative trade-off with safety.	Not Satisfied	Consider rebalancing the relationship between efficiency and safety.
Technical Complexity	High technical complexity guarantees effective certification, but has a negative trade-off with stakeholder engagement and enforcement.	Partially Satisfied	Ensure complexity does not impede oversight, enforcement. Focus on clear and transparent evaluation tools.
Human Factors	Considered primarily as abstract principles, such as agency, explainability, and trust, but not substantially implemented in the certification process.	Not Satisfied	Develop substantial inclusion, include clear criteria for human readiness and control in the process.
Data Governance	Underemphasized, with approaches leaving policies to external sources. Critical data management practices are excluded from certification processes.	Not Satisfied	Include data governance policies within certification approaches.

Table 2. Summary table

5 References

Austrian, E., Sawyer, M., Kring, J., Siragusa, K., & Gibson, S. (2023, 27-30 November 2023). Systematic Approach to Assessing the Readiness of a Technology for Safe and Effective Human Use SESAR Innovation Days 2023, Seville, Spain.

Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775-779. https://doi.org/https://doi.org/10.1016/0005-1098(83)90046-8

Bakirtzis, G., Carr, S., Danks, D., & Topcu, U. (2023). Dynamic Certification for Autonomous Systems. Commun. ACM, 66(9), 64–72. https://doi.org/10.1145/3574133

Balduzzi, G., Bravo, M. F., Chernova, A., & al., e. (2021). Neural Network Based Runway Landing Guidance for General Aviation Autoland. Federal Aviation Administration, 27 November 2021, DOT/FAA/TC-21/48.

Bhattacharyya, Siddhartaha, Darren Cofer (Rockwell Collins), David J. Musliner, Joseph Mueller, and Eric Engstrom (Smart Information Flow Technologies), Certification Considerations for Adaptive Systems, NASA/CR–2015-218702, March 2015

Brat, G. P., Yu, H., Atkins, E., Sharma, P., Cofer, D., Durling, M., Meng, B., Alexander, C., Borgyos, S., Fan, C., Garg, K., Topcu, U., & Bakirtzis, G. (2023). Autonomy verification & validation roadmap and vision 2045. NASA Ames Research Center, 31 January 2023, NASA/TM-20230003734.

Coordination of Automated Road Transport Deployment for Europe (CARTRE) (2018). 'D.3.8: "Guide on National Testing Regulations - Final edition"' (Horizon Europe 2020)

Dia, Hussain and Richard Tay, Ryszard Kowalczyk, Saeed Bagloee, Eleni Vlahogianni, Andy Song (2021), Artificial Intelligence Tests for Certification of Autonomous Vehicles, Academia Letters, May 2021.

EASA (2023). Artificial Intelligence Roadmap 2.0: Human-centric approach to AI in aviation. European Union Aviation Safety Agency, May 2023.

EASA (2024). EASA Concept Paper: Guidance for Level 1&2 machine learning applications. European Aviation Safety Agency, Issue 02, March 2024.

EASA and Collins Aerospace (2023). Formal Methods used for Learning Assurance (ForMuLA). European Aviation Safety Agency, 17 April 2023.

Endsley, M. R. (2023). Ironies of artificial intelligence. Ergonomics, 1-13. https://doi.org/10.1080/00140139.2023.2243404

EUROCAE (2021). Artificial intelligence in aeronautical systems: Statement of concerns. EUROCAE, April 2021, ER-022.

Eurocontrol (2020), European Aviation Artificial Intelligence High Level Group, The FLY AI Report: Demystifying and Accelerating AI in Aviation/ATM, 5th March 2020

European Commission (EC, 2019). REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL - Implementation of Directive 2010/40/EU of the European Parliament and of the Council of 7 July 2010 on the framework for the deployment of Intelligent Transport Systems in the field of road transport and for interfaces with other modes of transport (2019)

European Parliament Committee on Transport and Tourism (EPCTT) (2021). Opinion on the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://artificialintelligenceact.eu/wp-content/uploads/2022/09/AIA-TRAN-Rule-56-Opinion-Adopted-12-July.pdf

European Parliament (EP)(2017). European Parliament resolution of 16 February 2017 with recommendations to the Commission on civil law rules on robotics (2015/2103(INL)).

European Parliamentary Research Service (EPRS) (2017). Artificial intelligence in road transport.

EU (2021), AI Act, Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts

HFES (2021). Human Readiness Level Scale in the System Development Process. Human Factors and Ergonomics Society, ANSI/HFES 400-2021.

High-level Expert Group on AI (2019). Ethics guidelines for trustworthy AI. European Commission, 8 April 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

High-level Expert Group on AI (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment. 17 JUly 2020.

Holloway, C. M. (2019). Understanding the Overarching Properties. NASA Langley Research Center, July 2019, NASA/TM–2019–220292.

ICAO (2018). Safety Management Manual. International Civil Aviation Organization, Fourth edition, Doc 9858.

ISO/IEC (2024). Artificial intelligence — Functional safety and AI systems. First edition 2024-01, ISO/IEC TR 5469.

Koopman et al (WRC) (2019). Certification of Highly Automated Vehicles for Use on UK Roads : Creating An Industry-Wide Framework for Safety. Report. Five AI Ltd.

Malamet, F. et al, DEEL Certification Work Group, White Paper Machine Learning in Certified Systems, March 2021, Ref - S079L03T00-005

MLEAP Consortium (2023). EASA Research – Machine Learning Application Approval (MLEAP) interim technical report European Aviation Safety Agency, 11 May 2023.

Moss, R. J., Kochenderfer, M. J., Gariel, M., & Dubois, A. (2023). Bayesian Safety Validation for Black-Box Systems AIAA AVIATION 2023 Forum, https://arc.aiaa.org/doi/abs/10.2514/6.2023-3596

Nagarajan, P., Kannan, S., Torens, C., Vukas, M., & Wilber, G. (2021). ASTM F3269 - An Industry Standard on Run Time Assurance for Aircraft Systems. https://doi.org/10.2514/6.2021-0525

NASEM (2022). Human-AI Teaming: State of the Art and Research Needs. National Academies of Sciences, Engineering, and Medicine.

NASEM (2023). Test and Evaluation Challenges in Artificial Intelligence-Enabled Systems for the Department of the Air Force. National Academies of Sciences, Engineering, and Medicine.

National Research Council (2007). Software for Dependable Systems: Sufficient Evidence? The National Academies Press, Washington DC, doi: 10.17226/11923.

National Research Council (2014). Autonomy Research for Civil Aviation: Toward a New Era of Flight. The National Academies Press, Washington, DC, doi:10.17226/18815

NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology, January 2023, NIST AI 100-1.

Paraskevopoulou, Sivylla (2023), The Road to AI Certification: The importance of Verification and Validation in AI » Artificial Intelligence - MATLAB & Simulink (mathworks.com), Posted by Sivylla Paraskevopoulou, July 11, 2023

Paul, S., Prince, D., Iyer, N., Durling, M., Visnevski, N., Meng, B., Meiners, M., McMillan, C., Mandal, U., Siu, K., & Varanasi, S. C. (2023). Assurance of Machine Learning-Based Aerospace Systems: Towards an Overarching Properties-Driven Approach. Federal Aviation Administration, DOT/FAA/TC-23/54.

SASWG (2024). Safety Assurance Objectives for Autonomous Systems. Safety Critical Systems Club, Version 4.0, February 2024, SCSC-153C. https://scsc.uk/SCSC-153C

SESAR AI (2024). Smart ATM - Artificial intelligence. https://www.sesarju.eu/ai

SESAR JU (2024). European ATM Master Plan Stakeholder consultation workshop pre-read material, 22-23 April 2024. https://sesarju.eu/sites/default/files/documents/events/ATM%20MP%20workshop%20pre-read%20material_2024.04.08_FINAL.pdf

Stroeve, S., Smeltink, J., & Kirwan, B. (2022). Assessing and Advancing Safety Management in Aviation. Safety, 8(2), 20. https://www.mdpi.com/2313-576X/8/2/20

Zöldy et al (2020). Challenges in homologation process of vehicles with artificial intelligence. Transport, 35(4), 447. https://doi.org/10.3846/transport.2020.12904

6 List of acronyms

Acronym	Description
AI	Artificial Intelligence
AIA	AI Act
AIOps	AI-powered Operations
AIS	Autonomous Intelligent Systems
ALTAI	Assessment List for Trustworthy AI
AMC	Acceptable Means of Compliance
ANS	Air Navigation System
ARP	Aerospace Recommended Practice
ATC	Air Traffic Control
ATM	Air Traffic Management
BR	Basic Regulation
CARTRE	Coordination of Automated Road Transport Deployment for Europe
CCTV	Closed-Circuit Television
CM	Contingency Management
ConOps	Concept of Operations
DAL	Design Assurance Level
DTA	Design Time Assurance
EASA	European Union Aviation Safety Agency
EC	European Commission
ED	EUROCAE Document
EMAS	Eco-Management and Audit Scheme
EPCTT	European Parliament Committee on Transport and Tourism
EPRS	European Parliamentary Research Service
EU	European Union
EUROCAE	European Organization for Civil Aviation Equipment
FAA	Federal Aviation Administration
FHA	Functional Hazard Assessment/Analysis
FRIA	Fundamental Rights Impact Assessment
GDPR	General Data Protection Regulation
GM	Guidance Material
HASS	Highly Automated Safety Center of Excellence
HAT	Human AI-based system Teams
HFES	Human Factors and Ergonomics Society
HRL	Human Readiness Level
HSI	Human-System Integration
HUCAN	Holistic Unified Certification Approach for Novel systems based on advanced automation
IAS	Increasingly Autonomous System
ICAO	International Civil Air Organization
IEC	International Electrotechnical Commission

IEEE SA	Institute of Electrical and Electronics Engineers Standards Association
IPC	Innovation Partnership Contracts
ISO	International Organization for Standardization
ITS	Intelligent Transport System
KPA	Key Performance Area
LOA	Level of Automation
LOAT	Levels of Automation Taxonomy
ML	Machine Learning
MLEAP	Machine learning Application Approval
MLOps	ML-powered Operations
MOC	Means Of Compliance
MoC/MoU	Memoranda of Cooperation/Understanding
NAS	National Airspace System
NASA	National Aeronautics and Space Administration
NASEM	National Academies of Sciences, Engineering, and Medicine
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
ODD	Operational Design Domain
OECD	Organisation for Economic Co-operation and Development
OTA	Operation Time Assurance
RIPS	Recorder Independent Power Supply
RMF	Risk Management Framework
RTA	Run-Time Assurance
RTCA	Radio Technical Commission for Aeronautics
SA	Situation Awareness
SAE	Society of Automotive Engineers
SASWG	Safety of Autonomous Systems Working Group
SCSC	Safety Critical Systems Club
SESAR JU	SESAR Joint Undertaking
SORA	Specific Operation Risk Assessment
STPA	Systems Theoretic Process Analysis
SWAL	Software Assurance Levels
T&E	Testing & Evaluation
TIM	Technical Interchange Meeting
TRL	Technology Readiness Level
UAS	Unmanned Aerial Systems
UI/UX	User Interface / User Experience
V&V	Verification & Validation
VV&C	Verification, Validation, and Certification
WG	Working Group
WRC	White Rose University Consortium
XAI	Explainable AI

Table 3. List of acronyms

7 Glossary

Throughout our analysis, a multitude of different notions related to AI and automation have surfaced. These definitions are gathered not only from legal frameworks but also from research projects completed within the EU institutions.

A few observations are made regarding definitions. The first observation is that terms and definitions within the EU are not completely aligned with each other. While definitions indicate similar or identical concepts, their terms may differ from one another. The second observation is that the **AI Act** (EU, 2021) and **High-Level Expert Group on AI** (2019) rather generically refer to terms and define them while **EASA** (2023) and **SESAR AI** (2024) are rather aviation-specific. This results in discrepancy among terms; for example, the AI Act defines the term 'Serious Incident' more generically than how the aviation sector would define it.

The purpose of the section is not to analyse and come up with a different list of **Glossary** applicable to the certification of AI in the ATM, but to provide an overview which actors in the certification approach may refer to.

Term	Definition	Source
Adaptive Learning	Learning capability during the operations (see also online learning)	EASA (2023)
Adaptivity (of the Learning Process)	The ability to improve performance by learning from experience	EASA (2023)
Advanced Automation	The use of a system that, under specified conditions, functions without human intervention	EASA (2023)
AI Literacy	Skills, knowledge and understanding that allows providers, users and affected persons, taking into account their respective rights and obligations in the context of this regulation, to make an informed deployment of AI systems, as well as to gain awareness about the opportunities and risks of AI and possible harm it can cause	EU (2021)
AI Practitioners	All individuals or organisations that develop (including research, design or provide data for) deploy (including implement) or use AI systems, excluding those that use AI systems in the capacity of end user or consumer	High-level Expert Group on AI (2019)
AI Regulatory Sandbox	A concrete and controlled framework set up by a competent authority which offers providers or prospective providers of AI systems the possibility to develop, train, validate and test, where appropriate in real world conditions, an innovative AI system,	EU (2021)

	pursuant to a sandbox plan for a limited time under regulatory supervision	
AI System's Life Cycle	An AI system's life cycle encompasses its development (including research, design, data provision, and limited trials), deployment (including implementation) and use phase.	High-level Expert Group on AI (2019)
AI System	A machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments	EU (2021)
AI-Based System	A system that is developed with one or more of the techniques and approaches listed in Annex i to the AIA and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with	EASA (2023)
Artificial Intelligence (AI)	Technology that can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.	EASA (2023)
Artificial Intelligence (AI)	A branch of computer science that aims to create intelligent machines. it has become an essential part of the technology industry. AI can be narrow, handling just one particular task, or strong meaning a machine with the ability to apply intelligence to any problem	SESAR (2024)
Artificial Intelligence or AI Systems	Software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes	High-level Expert Group on AI (2019)

	planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems). A separate document prepared by the AI HLEG and elaborating on the definition of AI used for the purpose of this document is published in parallel, titled "A Definition of AI: Main capabilities and scientific disciplines"	
Artificial Intelligence Office	The Commission's function of contributing to the implementation, monitoring and supervision of AI systems, general purpose AI models and AI governance. references in this regulation to the artificial intelligence office shall be understood as references to the Commission	EU (2021)
Artificial Neural Network (ANN) or Neural Network (NN)	A computational graph which consists of connected nodes ('neurons') that define the order in which operations are performed on the input. neurons are connected by edges which are parameterised by weights (and biases). neurons are organised in layers, specifically an input layer, several intermediate layers, and an output layer. Especially within the context of the EASA (2023), ANN or NN are referred to as a specific type of neural network that is particularly suited to process image data: convolutional neural networks (CNNs) which use parameterised con convolution operations to compute their outputs.	EASA (2023)
Auditability	The ability of an AI system to undergo the assessment of the system's algorithms, data and design processes. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available. ensuring traceability and logging mechanisms from the early design phase of the AI system can help enabling the system's auditability	High-level Expert Group on AI (2019)
Authorised Representative	Means any natural or legal person located or established in the union who has received and accepted a written mandate from a provider of an AI system or a general-purpose AI model to, respectively,	EU (2021)

	perform and carry out on its behalf the obligations and procedures established by this regulation	
Authority	The ability to make decisions and take actions without the need for approval from another member involved in the operations	EASA (2023)
Automation	The use of control systems and information technologies reducing the need for human input, typically for repetitive tasks	EASA (2023)
Autonomy	Characteristic of a system that is capable of modifying its intended domain of use or goal without external intervention, control or oversight	EASA (2023)
Bias	An inclination of prejudice towards or against a person, object, or position. Bias can arise in many ways in AI systems. For example, in data-driven AI systems, such as those produced through machine learning, bias in data collection and training can result in an AI system demonstrating bias. In logic-based AI, such as rule-based systems, bias can arise due to how a knowledge engineer might view the rules that apply in a particular setting. bias can also arise due to online learning and adaptation through interaction. It can also arise through personalisation whereby users are presented with recommendations or information feeds that are tailored to the user's tastes. it does not necessarily relate to human bias or human-driven data collection. it can arise, for example, through the limited contexts in which a system is used, in which case there is no opportunity to generalise it to other contexts. bias can be good or bad, intentional or unintentional. In certain cases, bias can result in discriminatory and/or unfair outcomes, indicated in this document as unfair bias.	High-level Expert Group on AI (2019)
Bias (In The Data)	[The common definition of data bias is that] the available data is not representative of the population or phenomenon of study.	EASA (2023)
Bias (In The Ml Model)	An error from erroneous assumptions in the learning [process]. high bias can cause a learning algorithm to miss the relevant relations between attributes and target outputs (= underfitting).	EASA (2023)

Big Data	A recent and fast evolving technology, which allows the analysis of a big amount of data (more than terabytes), with a high velocity (high speed of data processing), from various sources (sensors, images, texts, etc.), and which might be unstructured (not standardised format).	EASA (2023)
Biometric Categorisation System	AI system for the purpose of assigning natural persons to specific categories on the basis of their biometric data unless ancillary to another commercial service and strictly necessary for objective technical reasons	EU (2021)
Biometric Data	Personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, such as facial images or dactyloscopic data	EU (2021)
Biometric Identification	Automated recognition of physical, physiological, behavioural, and psychological human features for the purpose of establishing an individual's identity by comparing biometric data of that individual to stored biometric data of individuals in a database	EU (2021)
Biometric Verification	The automated verification of the identity of natural persons by comparing biometric data of an individual to previously provided biometric data (one-to-one verification, including authentication)	EU (2021)
Ce Marking Of Conformity	A marking by which a provider indicates that an AI system is in conformity with the requirements set out in title iii, chapter 2 of this regulation and other applicable union legislation harmonising the conditions for the marketing of products ('Union Harmonisation Legislation') providing for its affixing	EU (2021)
Common Specification	A set of technical specifications, as defined in point 4 of article 2 of Regulation (EU) no 1025/2012 providing means to comply with certain requirements established under this regulation	EU (2021)
Conformity Assessment	The process of demonstrating whether the requirements set out in Title III, Chapter 2 of this regulation relating to a high-risk AI system have been fulfilled	EU (2021)

Conformity Assessment Body	A body that performs third-party conformity assessment activities, including testing, certification and inspection	EU (2021)
Convolutional Neural Networks (CNNs)	A specific type of deep neural networks that are particularly suited to process image data, based on convolution operators.	EASA (2023)
Critical Infrastructure	An asset, a facility, equipment, a network or a system, or a part of thereof, which is necessary for the provision of an essential service within the meaning of Article 2(4) of Directive (EU) 2022/2557	EU (2021)
Data Governance	A data management concept concerning the capability of an organisation to ensure that high data quality exists throughout the complete life cycle of the data, and data controls are implemented that support business objectives. The key focus areas of data governance include data availability, usability, consistency, integrity, and sharing. It also relates to establishing processes to ensure effective data management throughout the enterprise, such as accountability for the adverse effects of poor data quality, and ensuring that the data which an enterprise has can be used by the entire organisation.	EASA (2023)
Data Set	(In ml in general) — the sample of data used for various development phases of the model, i.e. the model training, the learning process verification, and the inference model verification.	EASA (2023)
Data-Driven AI	An approach focusing on building a system that can learn a function based on having been trAIned on a large number of examples.	EASA (2023)
Decision	A conclusion or resolution reached after consideration. a choice that is made about something after thinking about several possibilities.	EASA (2023)
Decision-Making	The cognitive process resulting in the selection of a course of action among several possible alternative options. automated or automatic decision-making is the process of making a decision by automated means without any human involvement.	EASA (2023)
Deep Fake	AI generated or manipulated image, audio or video content that resembles existing persons, objects,	EU (2021)

	places or other entities or events and would falsely appear to a person to be authentic or truthful	
Deep Learning	The most advanced type of machine learning. In recent years, the availability of large amount of data ("big data") and the leap forward in computing power have paved the way towards unprecedented levels of performance, allowing for new levels of automation	SESAR
Deep Learning (DI)	A specific type of machine learning based on the use of large neural networks to learn abstract representations of the input data by composing many layers	EASA (2023)
Deployer	Any natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity	EU (2021)
Determinism	A system is deterministic if when given identical inputs, it produces identical outputs	EASA (2023)
Development Assurance	All those planned and systematic actions used to substantiate, to an adequate level of confidence, that errors in requirements, design, and implementation have been identified and corrected such that the system satisfies the applicable certification basis	EASA (2023)
Distributor	Any natural or legal person in the supply chain, other than the provider or the importer, that makes an AI system available on the union market	EU (2021)
Domain	Operational area in which a system incorporating an ML subsystem could be implemented/used. Examples of domains considered in the scope of this guideline are ATM/ANS, air operations, flight crew training, environmental protection or aerodromes.	EASA (2023)
Downstream Provider	A provider of an AI system, including a general purpose AI system, which integrates an AI model, regardless of whether the model is provided by themselves and vertically integrated or provided by another entity based on contractual relations.	EU (2021)
Emotion Recognition System	An AI system for the purpose of identifying or inferring emotions or intentions of natural persons on the basis of their biometric data	EU (2021)
End User	An end user is the person that ultimately uses or is intended to ultimately use the AI-based system. This	EASA (2023)

	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	
Ethical AI	The development, deployment and use of AI that ensures compliance with ethical norms, including fundamental rights as special moral entitlements, ethical principles and related core values. It is the second of the three core elements necessary for achieving Trustworthy AI.	High-level Expert Group on AI (2019)

Table 4. Glossary