Holistic approach to approval and certification of automated systems

Deliverable ID:	**D4.4**
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:	**31 July 2025**
Edition:	**01.00**
Status:	**Official**
Classification:	**PU**

Abstract

This report presents one of the main deliverables of the HUCAN project, referred to as SESAR solution SOL.0445: a holistic framework including a toolbox of methods to support certification-aware design of human-centred ATM sociotechnical systems with high levels of automation.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Authoring & approval

Author(s) of the document

Organisation name	Date
NLR	31.07.2025
DLR	31.07.2025
DBL	31.07.2025

Reviewed by

Organisation name	Date
DBL	18.07.2025
NLR	22.07.2025
CIRA	22.07.2025
DLR	22.07.2025
EUI	17.04.2025
D-Flight	22.07.2025

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

Organisation name	Date
DBL	29.07.2025
NLR	29.07.2025
CIRA	29.07.2025
DLR	29.07.2025
EUI	29.07.2025
D-Flight	29.07.2025

Rejected by[2]

Organisation name	Date

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

[2] Representatives of the beneficiaries involved in the project

Document history

Edition	Date	Status	Company Author	Justification
0.1	09.12.2024	Draft	NLR DLR	Concept structure, development of chapters 2, 3, 4
0.2	27.03.2025	Draft	NLR DBL	Addition of new chapter 2, further development of chapters 3, 4
0.3	08.05.2025	Draft	NLR	Addition of appendix and chapter 5, further development of chapter 4
0.4	10.06.2025	Draft	NLR	Completion of chapter 4 and 5, development of chapter 6
0.5	10.07.2025	Draft	NLR	Incorporation of Expert group comments and feedback from D4.3, addition of executive summary, moved chapter 4 to appendix
0.6	11.07.2025	Draft	NLR	Version for internal review
0.7	25.07.2025	Draft	NLR	Review incorporated, version for final approval
1.0	31.07.2025	Official	NLR DBL	Final And quality check

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

Co-funded by the European Union

List of figures

List of tables

Co-funded by the European Union

Executive summary

The HUCAN project develops a novel holistic framework for certification-aware design of human-centred ATM sociotechnical systems with high levels of automation. This summary outlines the motivation and scope for this development, the elements of the HUCAN holistic framework, and recommendations for further work.

Motivation and scope

Motivation for this development

Artificial Intelligence (AI) and its Machine Learning (ML) constituent are key drivers of innovation, enabling higher levels of automation across multiple domains, improving operational efficiency for complex tasks and supporting human operators and organisations. However, the primary concern for all stakeholders involved in the transition to higher levels of automation is to establish the necessary conditions and standards to ensure that the solutions meet the stringent certification requirements.

HUCAN aims to explore these topics from the perspective of research and development (R&D) projects. A review of the technical and regulatory state of the art (HUCAN D2.1 & D3.1, 2024) reveals that current certification is predominantly based on prescriptive regulations, which mandate strict compliance with detailed requirements. This approach has proven effective in progressively enhancing safety. However, its applicability to highly automated and AI-driven technologies is increasingly being questioned, raising concerns about the suitability of existing certification frameworks.

In light of these considerations, the HUCAN project has carefully analysed the currently proposed innovative certification approaches found in the literature (HUCAN D3.2, 2024). What emerged is that there is a need for a broad-scope, holistic certification approach, with emphasis on addressing: human factors for understanding uncertainty and safety risks in sociotechnical systems with diverse levels of automation, the impact on accountability in design and operations, assuring public oversight and collaboration with diverse stakeholders, the incorporation of sustainability criteria for societal and environmental impacts, and data governance policies as part of certification. The HUCAN holistic framework aims to address these needs.

Automation and AI

The HUCAN holistic framework aims to be applicable to human-centred ATM airborne and ground systems that have high levels of automation. Automation refers to the extent to which the need for human input has been reduced, and to which technology can act or take decisions on its own. This extent is measured in terms of a Level of Automation (LOA), which ranges from LOA-0 (low automation; human has full authority), to LOA-5 (full automation; there is no human operator and the automated system decides/acts on its own). The technology includes but is not restricted to AI/ML-based systems.

In the context of this work, a *system* is an overall sociotechnical system, meaning that it describes the functioning and interface of the automation system, the functioning and interaction with other technical systems, the roles, tasks and responsibilities of human operators, and the operational conditions for which the system is designed.

Certification-aware system design as target

The HUCAN holistic framework aims at providing validation and feedback support for the iterative improvement of a system design and development, towards a future certification. As such, its application aims at obtaining a *certification-aware system design*.

Certification addresses not only safety, but performance in other areas too. Key Performance Areas (KPAs) define the areas in which the system needs to perform well to achieve its overall goals. The HUCAN framework, being holistic, covers a wide range of KPAs: Human factors, Accountability, Responsibility, Liability, Safety, Resilience, Security, Environmental sustainability, Societal sustainability, and Efficiency.

The maturity of the design is determined in terms of Technology Readiness Levels (TRL), which describe the maturity of the technology, and Human Readiness Levels (HRL), which describe the readiness of a technology for use by the intended human users. At level 1 the system design is specified by basic principles only; at level 9 the system is successfully used in operations.

Broader than learning assurance

An important consideration in the certification of AI-based applications is *learning assurance*, which 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. Internationally, many activities are ongoing with a lot of focus on this topic. Joining those activities would create overlap and less efficient use of resources. Instead, HUCAN takes a broader perspective: a holistic framework for the certification-aware design of systems with automation, dealing with the human operator as an integrated part of the system, and addressing what this means for the operation as a whole. Learning assurance is incorporated in follow-on work.

HUCAN holistic framework for certification-aware design

At a high level, the HUCAN holistic framework follows a cyclic approach as shown in Figure 1. Input is a system design at a low level of maturity, which is guided through the HUCAN holistic cycle multiple times, towards a more mature design that increasingly fulfils KPAs and certification objectives.

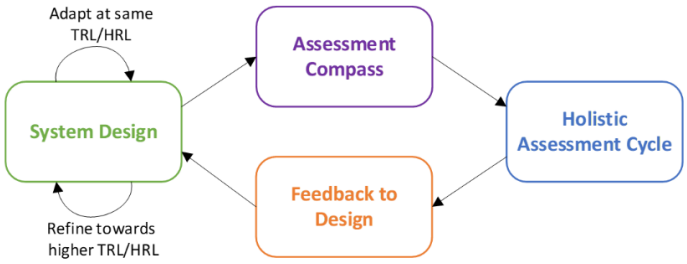


Figure 1. HUCAN holistic framework for certification-aware design.

The following main elements can be discerned:

- **System Design.** System design is the start- and endpoint of the cycle. At the start-point, it provides the input design for the assessment; at the endpoint, it provides the adapted/refined design that uses the feedback from the assessment, aiming at higher levels of maturity.

- **Assessment Compass.** This step sets the scene for the holistic assessment of the design by determining LOAs, TRLs/HRLs, KPAs, and certification objectives.
- **Holistic Assessment Cycle.** This cycle is the core of the framework by assessing multiple KPAs for critical scenarios of the sociotechnical system with (AI-based) advanced automation.
- **Feedback to Design.** Based on the combined KPA results from the holistic assessment cycle, this step identifies issues in the current design or it identifies/refines requirements or assurance levels towards a more mature design.

As is shown in Figure 2, for each of these four main elements, the HUCAN holistic framework outlines a number of steps or activities, which are supported by a HUCAN toolbox of methods.

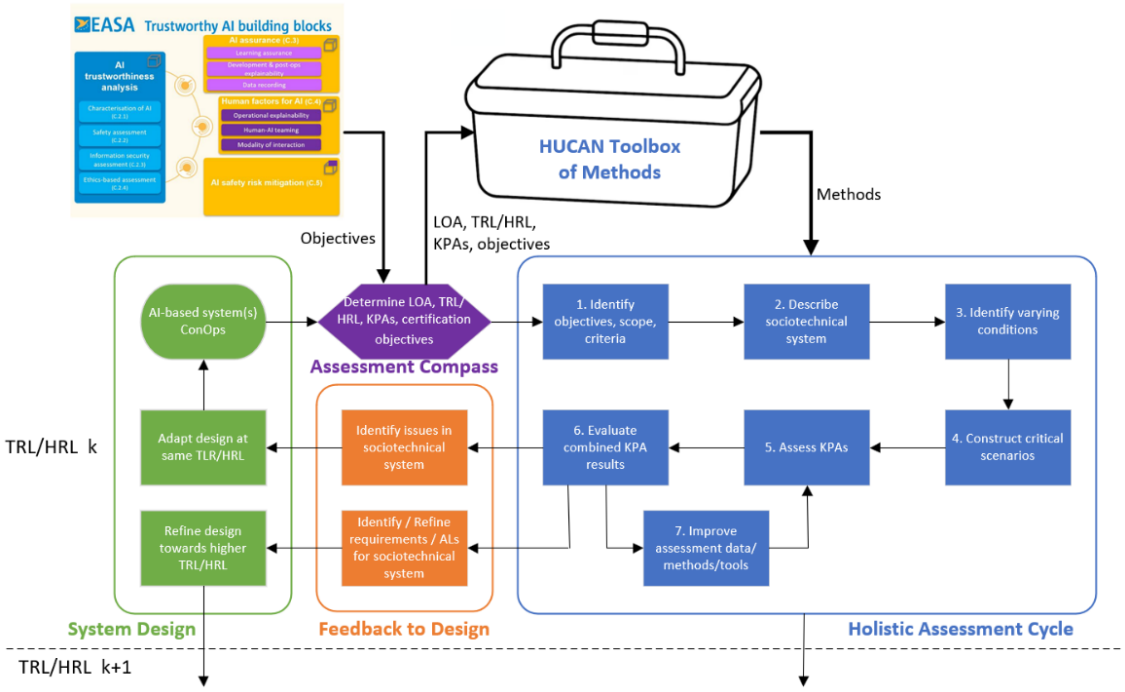
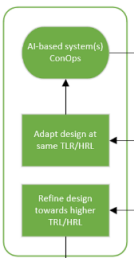


Figure 2. Steps in the HUCAN holistic framework for certification-aware design.

These steps/activities are briefly summarised below:

System design

The System Design contains three elements:

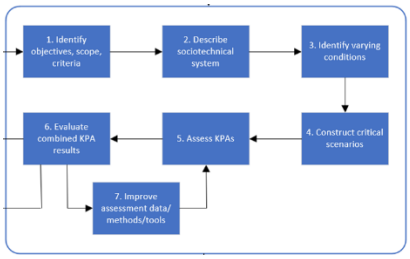


AI-based system(s) ConOps refers to the current design of the sociotechnical system and the associated advanced automation concept of operations. This is input for the Assessment Compass and the Holistic Assessment Cycle. The following two elements are applicable after completion of that cycle. Adapt design at same TRL/HRL: In case the design is considered not mature enough to proceed to the next level of TRL/HRL, the design is adapted at the same TRL/HRL. Refine design towards higher TRL/HRL: If the design is mature enough to proceed to the next level of TRL/HRL, the design is adapted towards a higher TRL/HRL.

Assessment Compass

Determine LOA, TRL/HRL, KPAs, certification objectives: The activities in the assessment compass element are to determine the levels of automation (LOA) applicable to the system, to determine current technology and human readiness levels (TRL/HRL), to determine the key performance areas (KPA) of interest, and to identify applicable certification objectives.

Holistic Assessment Cycle

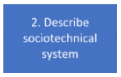


The holistic assessment cycle is the core of the framework, by assessing multiple KPAs for critical scenarios of the sociotechnical system with automation.

There are 7 activities:

1. Identify objectives, scope, criteria: In coordination with relevant stakeholders and proportional to the TRL/HRL and LOA of the sociotechnical system, step 1 defines:

- Objectives: which KPAs and which certification objectives will be considered.
- Scope: the boundaries of the operational area, and the types of functions or the types of equipment/procedures/people that are included.
- Performance criteria for the KPAs: these specify the difference between acceptable and unacceptable performance.

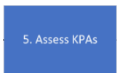


2. Describe sociotechnical system: In step 2 the sociotechnical system is described. This covers the objective of the operation, the operational context, environmental conditions, the functioning and interface of the advanced automation and of other technical systems, the roles, tasks and responsibilities of human operators and their interaction with all relevant AI-based and other technical systems. It also explicitly includes assumptions and constraints in the description of the sociotechnical system. It serves as an agreed, documented basis for the KPA assessments.

3: Identify varying conditions: Step 3 is to identify all kinds of disturbances and performance variability that can influence operations of the sociotechnical system. These can include frequently occurring conditions, like normal sensor errors, normal transmission delays, typical reaction times of human operators, differences in interpretations by humans, normal weather variability. But they also include rarer conditions, like system failures, extreme weather, particular errors by human operators.

4. Construct critical scenarios: Step 4 aims to construct scenarios that represent a critical impact on a KPA, e.g. a scenario leading to reduced safety, a scenario leading to an environmental problem, a scenario leading to a liability issue, etc. The critical scenarios are expanded by describing how agents of the sociotechnical system and related varying conditions may contribute to the KPA-critical effect.

Co-funded by the European Union



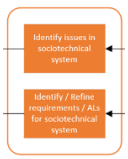
5. Assess KPAs: Step 5 aims to evaluate the constructed critical scenarios to get an assessment of the KPAs and associated criteria. For example, a quantitative assessment of safety or security risks, or a qualitative assessment of responsibility and liability issues. The step can be supported by a variety of methods and tools from the Toolbox, depending on the specific KPAs, on the types of results (qualitative/quantitative), and on the requested level of uncertainty in the results.

6. Evaluate combined KPA results: In step 6, the results of the assessment of the KPAs are combined and evaluated with respect to acceptability criteria. The result may lead to the following types of conclusions and basis for feedback to design.

1. It may be concluded with sufficient certainty that the design performance is not acceptable for one or more KPAs. Then the design cannot pass the current TRL/HRL and it would need to be adapted in an additional development cycle. Go to step "Identify issues in sociotechnical system" as part of feedback to design.
2. It may be concluded with sufficient certainty that the design's performance is acceptable for all KPAs. This implies that the design is considered suitable at its current readiness level. Go to step "Identify/Refine requirements/AIs for sociotechnical system" as part of feedback to design.
3. It may be concluded that there is insufficient certainty to evaluate the design performance for one or more KPAs, and this prevents reaching a verdict on its acceptability. In this case one of the following steps can be taken.
   a. If additional data or other methods/tools might sufficiently reduce the level of uncertainty in the assessment results, then go to step "7. Improve assessment data/methods/tools" in the holistic assessment cycle.
   b. If the level of uncertainty is high for several critical scenarios and KPAs, it may be decided to redevelop aspects of the design. Go to step "Identify issues in sociotechnical system" as part of feedback to design.

7. Improve assessment data/methods/tools: If the level of uncertainty in one or several KPA assessment results is too high to reach a conclusion on the acceptability of the design, it may be decided to improve the assessment(s). Gather additional information such as supporting data or expert opinion, or extend the models or techniques used in the assessment(s). This leads to an iteration of step "5. Assess KPAs".

Feedback to Design



Identify issues in sociotechnical system: The step at the top of the Feedback to design element is applicable in case of conclusion 1 at step "6. Evaluate combined KPA results". The design would need to be adapted in an additional development cycle. This leads to the step "Adapt design at same TRL/HRL" as part of the system design. Identify/refine requirements/ALs for sociotechnical system: The bottom step is applicable in case of conclusion 2 at step "6. Evaluate combined KPA results". The design is considered suitable at its current readiness level and can be refined at the next maturity level. This leads to the step "Refine design towards higher TRL/HRL" as part of the system design.

HUCAN Toolbox of Methods

The HUCAN holistic framework proposes a comprehensive suite of methods and tools for the holistic assessment of AI-driven and advanced automation systems. This HUCAN toolbox of methods integrates established methods as well as innovative approaches, recognizing the inherent complexity and unique challenges of integrating AI into safety-critical systems.

While acknowledging existing certification standards for non-AI systems, such as SAE ARP4761 (Aerospace Recommended Practice document 4761) and RTCA DO-178C (Software considerations in airborne systems and equipment certification) and DO-278A (Software integrity assurance considerations for CNS/ATM) systems), the HUCAN toolbox currently includes the following methods:

- ABMS (Agent-Based Modelling & Simulation)
- AI RMF (AI Risk Management Framework)
- BUSA (Bias, Uncertainty and Sensitivity Analysis)
- Environmental Assessment of AI Ecosystem
- FMEA (Failure Modes and Effects Analysis)
- FRAIA (Fundamental Rights and Algorithms Impact Assessment)
- HAZOP (Hazard and Operability study)
- Heuristic Evaluations
- HITL (Human-In-The-Loop) Simulations & Wizard of Oz
- HTA (Hierarchical Task Analysis)
- NSV-4 diagram (System Functionality and Flow model)
- Responsibility & Liability Analysis
- Safety Scanning and Security Scanning
- SecRAM (Security Risk Assessment Methodology)
- Usability Testing

For each method, this report includes a brief description, gives key benefits and limitations, and indicates which KPAs, TRLs/HRLs, LOAs, and EASA objectives are covered by the method.

Conclusions and recommendations for further work

This HUCAN deliverable presents a solution to the emerging need for a broad-scope, holistic certification approach for ATM sociotechnical systems with high levels of automation. This solution is referred to as SESAR solution SOL.0445.

From an operational standpoint, the description of the HUCAN holistic framework reveals that there is no one-size-fits-all solution. The broad spectrum of methods in the current toolbox reflects the multidimensional nature of the challenge, while the emphasis on method selection and awareness of their limitations highlights the need for a thoughtful, well-informed validation strategy.

The current version of the HUCAN holistic framework is set at TRL2 (i.e. it is a concept framework). In terms of recommendations for further development, it is important to stress that significant work is still required to systematically mature the holistic validation approach to align both with current certification standards and practices and with EASA's AI guidelines. This path also requires focus on

learning assurance, on the balance between KPAs, and on proportionality. Feedback from application to use cases will be an important contribution.

Another recommendation is regarding the anticipated update of the EASA AI guidelines. The current version, Issue 02 (EASA, 2024), is mostly focused on supervised learning, and it covers offline learning processes where the model is 'frozen' at the time of approval. The anticipated update might address other types of learning such as reinforcement learning, online learning processes, and Levels of Automation 3A and 3B (i.e. advanced automation). This update may come with additional objectives and additional challenges, to be addressed by a future update of the HUCAN Holistic Framework.

Co-funded by the European Union

1 Introduction

1.1 Objective

The HUCAN project (Holistic Unified Certification Approach for Novel systems based on advanced automation) aims to pioneer certification methods for new Air Traffic Management (ATM) systems with a focus on human centred advanced automation and Artificial Intelligence (AI)-based technologies. The project proposes a novel and holistic framework for certification-aware design of such systems. The aimed readiness level of the HUCAN holistic framework is TRL2, i.e. it is a concept framework.

The purpose of this report D4.4 is to present the development of the HUCAN holistic framework, including a toolbox of supporting methods. It builds on earlier deliverables, notably D3.2 “Innovative approaches to approval and certification”, D4.1 “Case studies introduction: Level of automation analysis and certification issues”, and D4.2 “Performance based requirements for advanced automation”. Furthermore, it incorporates the findings from D4.3 “Validation report”, which aimed at validating the D4.4 output. As such, the results presented in this report have profited from the feedback of the Expert group in the HUCAN project.

1.2 Organisation

This document is organised as follows:

- Chapter 2 gives the context for this work, by explaining the need for a holistic approach to certification of advanced automation, explaining why certification issues need to be addressed in research and development (R&D), and introducing at a very high level the HUCAN holistic approach supporting R&D towards certification.
- Chapter 3 describes the HUCAN holistic framework for certification-aware design. After an introduction of the main elements, the chapter presents each of these elements in more detail, and explains the processes in the HUCAN holistic framework. These processes form a holistic assessment cycle that gives feedback for the design, and supports a transition to a higher technology/human readiness level of the system and operation.
- Chapter 4 explains how the HUCAN holistic framework can be applied to support certification-aware design in line with the objectives of the EASA AI guidelines (EASA, 2024).
- Chapter 5 provides conclusions and recommendations.
- Chapter 6 provides references to material used.
- Chapter 7 gives a list of acronyms.
- Appendix A presents a toolbox of methods (i.e. techniques, standards, methodologies, databases, models) that can be used in support of evaluation of a KPA in an operation including advanced automation and AI-based systems.
- Appendix B provides an overview of the objectives of the EASA AI guidelines.

2 Towards a holistic approach to support the certification of advanced automation and AI-based systems

2.1 Why a holistic approach to certification of advanced automation?

Artificial Intelligence (AI) is poised to become a key driver of innovation, enabling advanced automation across multiple domains, improving operational efficiency for complex tasks and supporting human operators and organisations. As in many other industries, the aviation sector is actively working to leverage the benefits of this technological revolution. EASA's 'Roadmap 2.0 for a Human-Centric Approach to AI in Aviation' highlights the wide-ranging impact of AI across multiple domains, spanning both operational and non-operational contexts (EASA, 2023). However, the primary concern for all stakeholders involved in this transition is to establish the necessary conditions and standards to ensure that AI-driven solutions support efficient operations while meeting stringent safety requirements.

The HUCAN project is part of this ongoing debate and aims to explore these topics from the perspective of research and development projects, particularly within the framework of the S3JU (SESAR 3 Joint Undertaking) program. A review of the current state of the art, both technical and regulatory (HUCAN D2.1 & D3.1, 2024), reveals that currently, avionics and aircraft certification is predominantly based on prescriptive regulations, which mandate strict compliance with detailed implementation requirements. This approach, grounded in collective knowledge from past experiences, has proven effective in progressively enhancing safety. However, its applicability to highly automated and AI-driven technologies is increasingly being questioned. AI-based systems, characterised by varying levels of autonomy and possibly non-deterministic behaviour, often diverge from traditional software development paradigms, raising concerns about the suitability of existing regulatory frameworks.

In light of these considerations, the HUCAN project has carefully analysed the currently proposed innovative certification approaches found in the literature and standardization efforts for certifying these solutions (HUCAN D3.2, 2024). The objective was to understand, in particular, the goals, methodologies, Key Performance Areas (KPAs), and, where available, Key Performance Indicators (KPIs) that are emerging to address these technical challenges. Specifically, the analysis examined whether and how the objectives currently guiding research and development strategies—both in general and in aviation—are effectively reflected in these innovative approaches. This assessment aimed to verify whether the emerging certification frameworks are truly aligned with the technical and societal expectations of safety, efficiency, and sustainability associated with the adoption of these solutions.

In this reflection, the orientations of the EU AI Strategy, the ethical guidelines for trustworthy AI set by the European Commission, as well as the guidance provided by EASA and SESAR for the development and certification of these solutions in aviation, have played a particularly important role. What has emerged is that the reliability and technical robustness of these solutions must necessarily take into account the impacts that these innovations may have on the people who use them, the organisations that adopt them, and those who are directly or indirectly affected by their proper deployment (Lanzi et al., 2024).

In particular, the analysis in (HUCAN D3.2, 2024) was conducted in light of the standard KPAs for SESAR projects, supplemented by the objectives outlined in various frameworks such as the S3JU Multiannual

Work Framework Programme 2022-2031, the European ATM Master Plan 2020, and 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 European Commission's strategy "Artificial Intelligence for Europe" (COM/2018/237 final). Additionally, with the publication of the new ATM Master Plan 2025 by SESAR, specific criteria were identified to assess whether the innovative approaches under consideration were truly holistic. This evaluation led to the key findings listed in Table 1.

Overall, the evaluation addresses the need for a broad-scope, holistic certification approach, which emphasises addressing human factors for understanding uncertainty and safety risks in sociotechnical systems with diverse levels of automation, the impact on accountability in design and operations, assuring public oversight and collaboration with diverse stakeholders, the incorporation of sustainability criteria for societal and environmental impacts, and data governance policies as part of certification.

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	Partially Satisfied	Ensure complexity does not impede oversight, enforcement. Focus on

	trade-off with stakeholder engagement and enforcement.		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	Underemphasised, 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 1. Summary of evaluation of innovative approaches supporting certification, source (HUCAN D3.2, 2024).

2.2 Why considering certification issues in research and development?

In response, and particularly following the adoption of the EU AI Act (Reg. (EU) 2024/1689), EASA is actively engaged in a comprehensive review of the current aviation regulatory ecosystem, seeking to adapt existing provisions to the evolving technological landscape. With the establishment of the Rule-Making Task RMT.0742 – Artificial Intelligence Trustworthiness, the Agency is laying the groundwork for regulatory evolution in the aviation sector. The objective is to establish regulatory frameworks capable of supporting the safe development and deployment of AI-driven solutions in aviation while maintaining the highest safety standards. Specifically, in preparation for a certification basis, EASA is developing guidance that provides a broad perspective on key performance areas that should be addressed by objectives and anticipated means of compliance for advanced automation and AI-based systems (EASA, 2024).

An overview of the building blocks and topics with the numbers of associated objectives and provisions is given in Table 2; see also a more extended summary of the approach in (HUCAN D3.2, 2024), lists of all EASA AI objectives in Appendix B of the current document, and associated key performance indicators in (HUCAN D4.2, 2024). The way that AI objectives should be satisfied depends on assurance levels for system functions as determined from risk results in a functional hazard assessment, or on security assurance levels for the three security-related objectives, as well as on levels of AI. The scope of (EASA, 2024) is limited to Level 1 and Level 2 AI applications, as well as to machine learning techniques (particularly supervised learning). The EASA AI guidance material is expected to be extended to AI Level 3 and to a broader set of AI technology, including reinforcement learning, logic- and knowledge-based approaches, and hybrid AI.

Building Block	Themes	Explanation
Trustworthiness Analysis (21)	Characterization (7)	Identifying end users, AI interaction, concept of operation, system functions, and level of automation.
	Safety assessment (3)	Initial safety assessment, including functional hazard assessment, allocation of assurance levels,

		mitigation needs, verification of safety objectives. Continuous safety assessment to ensure continued safe operations.
	Information security (3)	Identifying security risks, mitigations, and fail-safe under security threats.
	Ethics-based assessment (8)	Ethics-based trustworthiness assessment, e.g. (High-level Expert Group on AI, 2020).
AI Assurance (65)	Learning assurance (56)	Learning assurance process following a W-shaped cycle, including management of data and learning processes, model training and implementation, and verification of learning, inference models, data management.
	Development and post-ops AI explainability (9)	Explainability of AI applications to engineers, certification authorities, and safety investigators.
Human Factors for AI (46)	AI operational explainability (10)	Explainability of AI applications to end users.
	Human-AI teaming (11)	Objectives for cooperation or collaboration between humans and AI-based systems.
	Modality of interaction and style of interface (16)	Design guidance objectives for new modes of human-machine interaction through voice, gesture, or other.
	Error management (5)	Objectives towards reducing risks of human errors.
	Failure management (4)	Objectives to support management of failure conditions.
AI safety risk mitigation (2)	AI safety risk mitigation (2)	Safety risk mitigation measures to mitigate residual risks, e.g. by real-time monitoring and passivation of AI systems.
	Theme/Provisions	
	Organisation (8)	High-level provisions to guide organisations for the introduction of AI-based systems.

Table 2. Building blocks and topics with numbers of objectives in the EASA AI guidelines (EASA, 2024).

As this summary highlights, the objectives identified by EASA move toward a more holistic approach to certification. This approach not only addresses AI assurance, safety, and risk mitigation but also calls for a rethinking and redesign of human factors paradigms, particularly in relation to redefining authority in human-AI collaboration. A preliminary AI trustworthiness analysis is introduced as a crucial step in assessing certification requirements, allowing for an early and comprehensive evaluation of both technical and operational risks across varying levels of automation.

As outlined by EASA, these AI guidelines—along with their key building blocks and associated objectives—serve as an initial reference point, offering a preliminary set of actionable objectives. The primary goal is to provide applicants with a foundational framework to guide decision-making in the development strategy of ML solutions. However, this initial set of objectives does not yet represent a definitive or detailed Means of Compliance (MOC). In alignment with the EU AI Act, while full compliance with legal and regulatory requirements is only expected at higher maturity levels, when solutions are ready for real-world testing (Reg. EU 2024/1689, Article 2(8)), this guidance implicitly promotes a progressive alignment starting from the early design phases.

In light of these considerations, the research activities carried out within HUCAN, including those based on the Use Cases (UCs) covered by the project (HUCAN D4.1 & D4.2, 2024), have highlighted the following:

- **The centrality of an iterative approach, not only from a technical perspective but also from a systemic perspective, is crucial for ensuring continuous refinement and improving the system over time** – In system life cycle processes, iteration and recursion are important for the progressive refinement of processes, system elements, and systems (ISO/IEC/IEEE, 2023). Especially, interactions between successive verification, validation, and integration processes can incrementally build confidence in the conformance of the product or service. Such iterative, cyclic development and associated verification and validation is especially important for high levels of automation concepts with key roles and responsibilities of AI-based systems. In particular, the effective use of AI requires an agile and cyclic development methodology (NASEM, 2023), where the concept of “justified confidence” can be used as a progressive measure of trustworthiness: developers, testers, and users should gain justified confidence in AI-based systems over time as they become increasingly familiar with system performance limits and behaviours. Also in (EASA, 2024) the iterative nature of learning assurance processes is stressed, as well as for the trustworthiness analysis and the assessments required in the others building blocks.

- **Aligning concepts with compliance objectives from the early stages of design is vital, as it helps address the challenges of contextualizing these objectives within the development pipeline and ensures the system meets regulatory requirements throughout its lifecycle** – While approval by certifying authorities especially concerns advanced automation and AI-based systems at high maturity levels, considering certification objectives at lower maturity levels can provide effective feedback for their development, and the documentation of the evaluation studies can provide an effective basis for the certification. However, it is not immediately clear when and how the different objectives can and should be considered throughout the development of a solution, as the available information about the concepts, the technical aspects of the solutions, and the impact on operators and organisations may

change during the process (HUCAN D4.1 and D4.2, 2024). For this reason, even though stable references – such as Technology Readiness Levels (TRLs), which provide structure to the development of new technologies like AI, and Human Readiness Levels (HRLs), which ensure that a new system can be effectively used by humans (HFES, 2021) – exist, it is still necessary to contextualise the objectives that can be considered at each stage and the methods that can contribute to achieving them.

- **The process is inherently interdisciplinary and involves multiple stakeholders, requiring careful coordination to ensure successful certification and the smooth integration of various expertise and responsibilities** – Another key consideration for the development and approval of advanced automation and AI-based systems is that they require collaboration with a diverse set of stakeholders during the life cycle stages. Stakeholders are individuals or organisations having a right, share, claim, or interest in a system or in its possession of characteristics that meet their needs and expectations (ISO/IEC/IEEE, 2023). In general, stakeholders in aviation include end users (pilots, air traffic controllers, etc.), end user organisations (airlines, ANSPs), developers and producers (aircraft manufacturers, surveillance system producers), trainers, maintainers, authorities and regulatory bodies (EASA, FAA, NSAs), and people influenced by the system (passengers, municipalities). Here, authorities have the legally responsible role to award certification and licenses that allow stakeholders to produce or utilise a particular system. For the development of AI-based systems, in particular, there are important innovative roles for the AI technology providers and for data providers. However, if the research and development process and alignment with certification require new competencies, it is necessary to coordinate different areas of expertise at various stages of validation.

2.3 Holistic approach supporting R&D towards certification

In summary, the EU AI Act, the HUCAN review and the developing EASA guidelines all stress the need of a holistic approach for evaluation and approval of AI-based systems, which should extend beyond typical safety and reliability considerations. Moreover, this concept of holism not only covers the nature of the objectives that certification should aim for, but also the way in which compliance should be addressed over time and the type of competencies required to effectively achieve the objectives.

An overview of the key elements of a holistic approach in support of development and approval of advanced automation and AI-based systems is provided in Figure 3. It shows that the approach involves coordinating with **multiple stakeholders** and addresses **multiple KPAs** in **multiple cycles** for design, development and evaluation of advanced automation and AI-based systems for a range of **levels of automation**. In these cycles, the maturity of the advanced automation concepts and supporting technology are increasing and their readiness levels are evaluated for a holistic scope of KPAs. In coordination with stakeholders, requirements addressing the various KPAs can be updated as the designers, developers, evaluators and other stakeholders achieve better understanding of the performance of the overall system and the impact on the KPAs. As such, it supports the development and certification of trustworthy advanced automation and supporting AI technology.

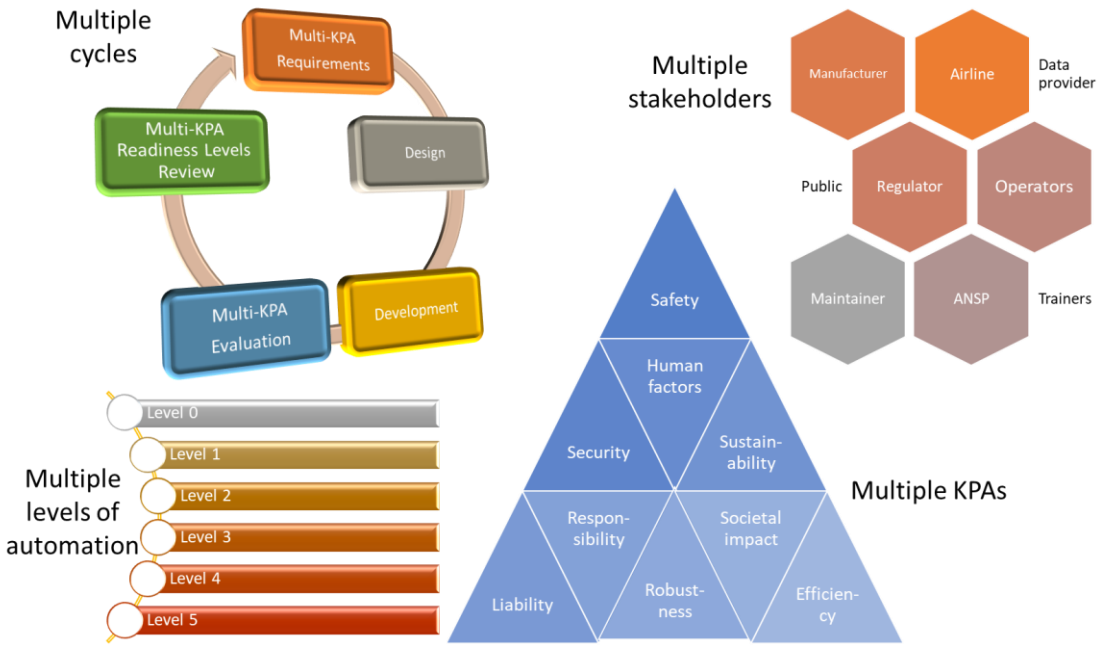


Figure 3. Holistic approach in support of development and approval of advanced automation and AI-based systems of a range of levels of automation, encompassing the interests of multiple stakeholders and addressing multiple key performance areas in multiple cycles for design, development and evaluation with increasing readiness levels.

As argued above, certification objectives and requirements are best considered from early design phases to ensure that emerging R&D solutions and concepts are aligned with compliance requirements from the outset. In particular for R&D in the S3JU framework, gradual alignment needs to be ensured of advanced automation and AI-based solutions with the objectives set out in EU AI Act and the developing guidelines (EASA, 2024).

With the aim of contributing to this alignment effort, HUCAN has developed an approach to facilitate the harmonization of objectives between the frameworks proposed by SESAR and EASA (see Figure 4). This approach defines, for each maturity level of a given solution, the specific requirements that should be met across different KPAs. By adopting the HUCAN approach, it becomes possible to map the compliance status of a solution both *as is* and *as it shall be* at a specific stage within the research and development pipeline. Furthermore, this approach enables the formulation of targeted recommendations to enhance compliance wherever gaps are identified, ensuring a structured pathway toward regulatory and operational alignment. The steps of the HUCAN holistic approach are provided next in Chapter 3.

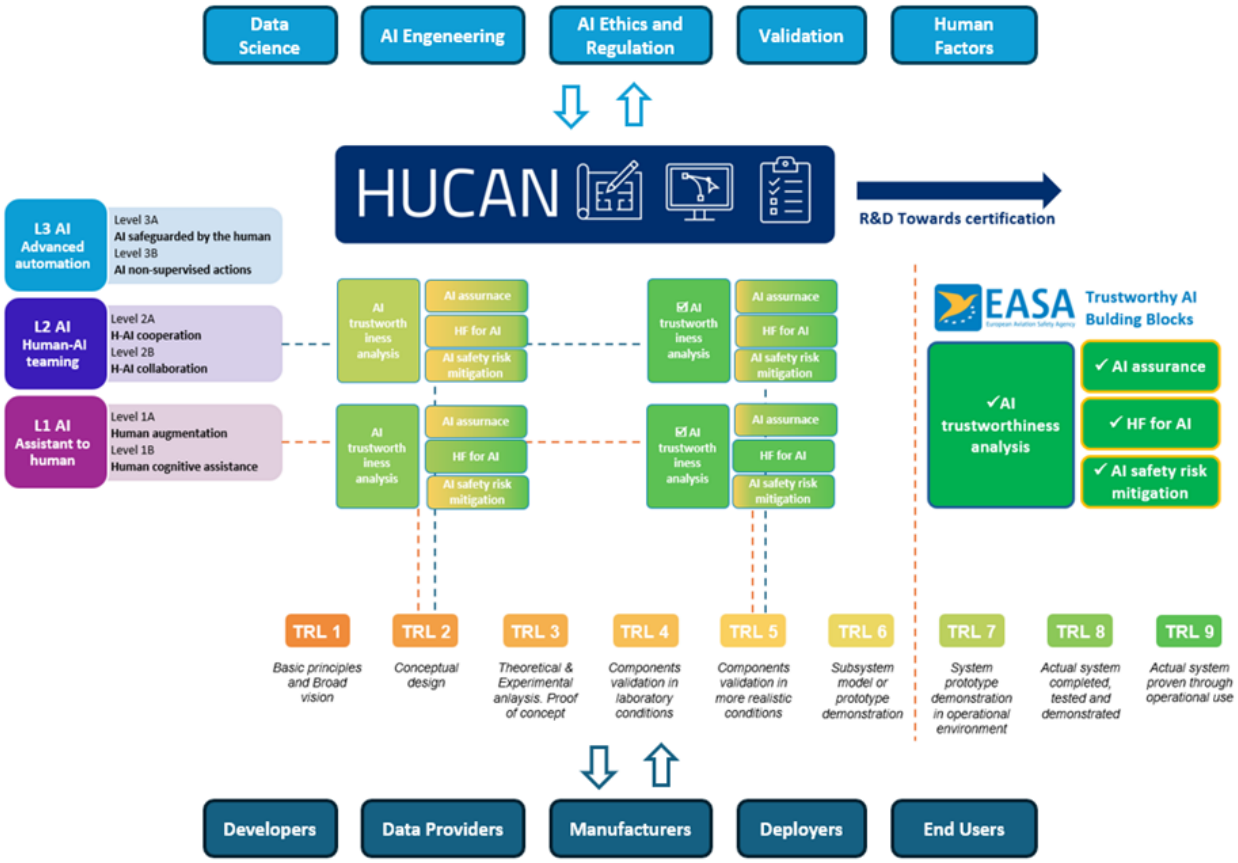


Figure 4. HUCAN holistic approach to facilitate the harmonization of certification objectives in the development along maturity levels of AI-based systems and advanced automation.

3 HUCAN holistic framework for certification-aware design

3.1 Introduction

Following the motivation of a holistic framework for certification-aware design in Chapter 2, this chapter elucidates the steps of the approach as developed by HUCAN. A high-level overview is shown in Figure 5. The following main elements can be discerned:

- **System Design.** System design is the start- and endpoint of the cycle by providing the basis for the assessment, as well as the updated design given the feedback from the assessment. In this context, the **system** is the **overall sociotechnical system**, meaning that it describes the functioning and interface of the AI-based system(s), the functioning and interaction with other technical systems, the roles, tasks and responsibilities of human operators, and the operational conditions for which the system is designed. The way that the design is changed is up to the design team and it is separated from the assessment of the design.
- **Assessment Compass.** This step sets the scene for the assessment by determining levels of automation, technology and human readiness levels (TRLs/HRLs), key performance areas, and certification objectives. These elements are explained in Section 3.2.
- **Holistic Assessment Cycle.** This cycle is the core of the framework by assessing multiple KPAs for critical scenarios of the sociotechnical system with (AI-based) advanced automation. Its steps are explained in Section 3.3.
- **Feedback to Design.** Based on the combined KPA results from the holistic assessment cycle, this step identifies issues in the current design or it identifies/refines requirements or assurance levels towards a more mature design. The types of feedback are explained in Section 3.4.

The HUCAN holistic framework is supported by a toolbox of methods/tools, which is provided in Appendix A.

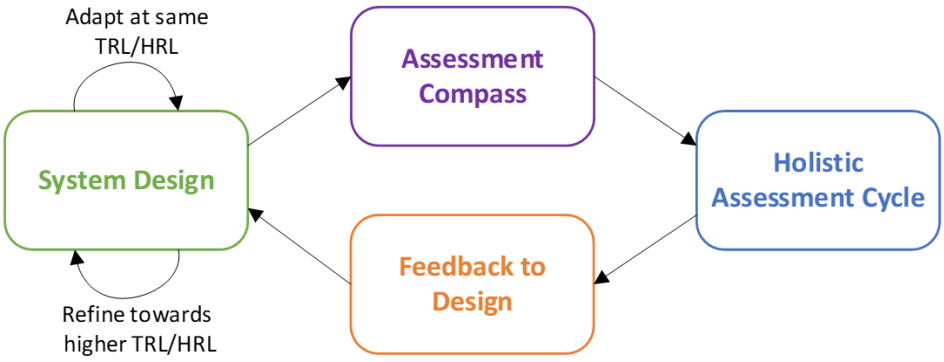


Figure 5. HUCAN holistic framework for certification-aware design.

3.2 Assessment Compass

3.2.1 Determine levels of automation/AI

Where automation may be defined as “The use of control systems and information technologies reducing the need for human input, typically for repetitive tasks” (EASA, 2024), a level of automation (LOA) then refers to the extent by which the need for human input has been reduced. Various taxonomies for LOAs exist (Vagia et al., 2016), typically ranging from manual operation without any assistance to fully autonomous operations. In (EASA, 2024) AI applications are classified along six categories AI Level 1A to AI Level 3B, which describe the level of human involvement in the applications, and Section C.2.1.4 provides guidance on choosing these AI levels. In (SESAR JU, 2024) these six AI levels are associated with six LOAs, as shown in Figure 6. At all levels the system provides full support for perception and analysis, but there are differences with regard to the degree of decision-making, execution, and the authority of the human operator.

- LOA-0 “low automation”: The human operator has full authority, takes all decisions and implements actions with or without execution support by a Level 1A “human augmentation” AI system.
- LOA-1 “decision support”: The human operator has full authority, receives decision support by a Level 1B “human assistance” AI system and implements actions with or without support by the system.
- LOA-2 “resolution support”: The human operator has full authority, receives resolution support by a Level 2A “human-AI cooperation” AI system and validates the proposed solution or comes up with a different solution. The system implements the actions under direction of the operator.
- LOA-3 “conditional automation”: The human operator has partial authority and supervises the performance of a Level 2B “human-AI collaboration” AI system. The system selects solutions and implements the actions. The operator overrides or improves solutions that are not deemed appropriate.
- LOA-4 “confined automation”: The human operator has limited authority, and only supervises and possibly intervenes, if requested by the Level 3A “safeguarded advanced automation” AI system. The system takes all decisions and implements all actions, except if it comes out of a predefined scope.
- LOA-5 “full automation”: There is no human operator. The Level 3B “non-supervised advanced automation” AI system decides and acts on its own.

Detailed guidance on how to determine the LOA for an application is provided in e.g. section C.2.1.4 of the EASA AI guidelines (EASA, 2024).

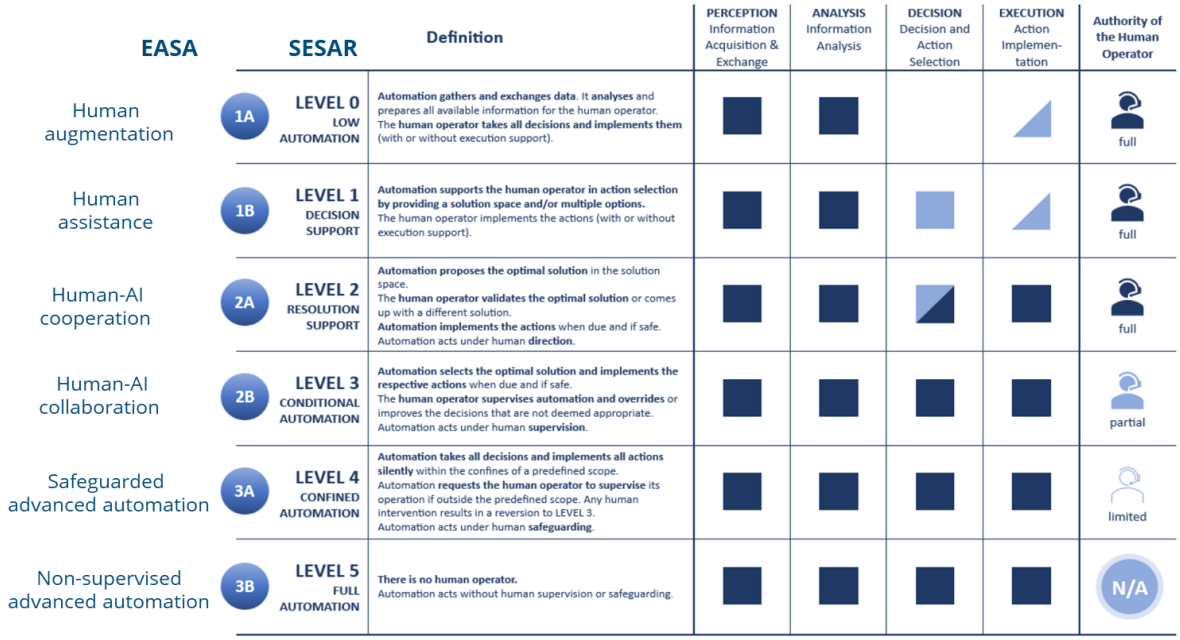


Figure 6. SESAR JU proposed new Levels of Automation Taxonomy and correspondence to EASA AI Levels, adapted from (SESAR JU, 2024).

3.2.2 Determine technology and human readiness levels

Technology and human readiness levels (TRLs/HRLs) support programme management for development of new technology and ways of working. TRLs provide a nine-level scale to describe the maturity of technology, which have been developed at NASA (Mankins, 2009) and now widely used. Recognizing that many system development programs have been deficient in applying established and scientifically-based human system integration (HSI) processes, the Human Factors & Ergonomics Society (HFES) developed a standard to define a nine-level scale of HRLs and provide guidance for their application (HFES, 2021). Here human readiness is the readiness of a technology for use by the intended human users in a specified intended operational environment. Table 3 gives an overview of HRLs and TRLs as provided in (HFES, 2021).

Levels 1 to 3 regard basic research and development, levels 4 to 6 regard human factors and technology demonstrations for increasing levels of fidelity, and levels 7 to 9 regard full-scale testing, production, and deployment. Transgression from lower to higher levels is based on (funding) decisions/approvals of R&D organisations and/or system designers/manufacturers at lower and middle levels. Approval by a certifying authority especially concerns TRL/HRL 8, but as part of the cyclic development the regulator should be involved at lower TRL/HRL to assure that the viewpoints and feedback of this key stakeholder are incorporated at a sufficiently early stage. TRL/HRL 9 applies to continuous safety management in operation, including oversight by regulatory bodies. As explained in (HFES, 2021), TRL and HRL levels should remain aligned in design and development activities, as misalignment may generate programme risks, depending on the phase of the development and the extent of the discrepancy.

The types of evaluation methods that can be used effectively in the development and that may provide objective evidence that may be used in certification depend on the TRL and HRL of the system under

consideration. Therefore, as a first step in the evaluation approach, a suitable TRL and HRL should be determined using the definitions in Table 3. Guidance on determining HRL is in (HFES, 2021), while guidance on TRL for SESAR projects is in (SESAR JU, 2025).

Level	HRL	TRL
1	Basic principles for human characteristics, performance, and behaviour observed and reported	Basic principles observed and reported
2	Human-centred concepts, applications, and guidelines defined	Technology concept and/or application formulated
3	Human-centred requirements to support human performance and human-technology interactions established	Analytical and experimental critical function and/or characteristic proof of concept
4	Modelling, part-task testing, and trade studies of human systems design concepts and applications completed	Component and/or breadboard validation in laboratory environment
5	Human-centred evaluation of prototypes in mission-relevant part-task simulations completed to inform design	Component and/or breadboard validation in relevant environment
6	Human systems design fully matured and demonstrated in a relevant high-fidelity, simulated environment or actual environment	System/subsystem model or prototype demonstration in a relevant environment
7	Human systems design fully tested and verified in operational environment with system hardware and software and representative users	System prototype demonstration in an operational environment
8	Human systems design fully tested, verified, and approved in mission operations, using completed system hardware and software and representative users	Actual system completed and qualified through test and demonstration
9	System successfully used in operations across the operational envelope with systematic monitoring of human-system performance	Actual system proven through successful mission operations

Table 3. Human and technology readiness levels; based on Table 5-1 of (HFES, 2021).

3.2.3 Determine key performance areas

The most important characteristic of the holistic validation framework is that it covers a broad scope of KPAs. As a basis the assessment team should determine in coordination with the stakeholders what

Co-funded by the European Union

KPAs should be addressed for a particular application. The following KPAs are included in the holistic framework.

- **Human Factors (HF).** Topics for validating human-system integration (HSI) in operational concepts with advanced automation include human-AI team models, processes and interaction, situation awareness in higher LOAs, AI transparency and explainability in operations, trust, decision bias, training, and overall HSI (NASEM, 2022). HF have a key impact on robustness, safety and security of AI-supported operations, variability in human behaviour contributes to flexibility and uncertainty, and accountability and responsibility of human operators in relation with other stakeholders is a key aspect for successful introduction of higher LOAs. In the EASA AI guidelines the importance of HF for AI is recognised and a range of objectives are presented in Section C.4 of (EASA, 2024) for topics covering AI operational explainability, human-AI teaming, human-AI interfaces, and error, failure and workload management. The HRLs can be used as a structure to validate the HSI in advanced automation concepts using AI-based systems. In particular, (HFES, 2021) provides for each HRL, a series of evaluation activities and the associated supporting evidence, and the exit criteria to a next HRL level.
- **Accountability.** Accountability is one of the seven requirements for trustworthy AI systems as defined in (High-level Expert Group on AI, 2019) and explained in (Díaz-Rodríguez et al., 2023). Accountability is linked to the principle of fairness and as such is closely related to risk management, so as to prevent unfair adverse effects. Here, risks must be identified and mitigated transparently, to allow verification by third parties. Independent auditing of data, algorithms and design processes are needed for this and must be supported by techniques and tools. The use of impact assessments (e.g. red teaming or forms of Algorithmic Impact Assessment) both prior to and during the development, deployment and use of AI-based systems can be helpful to minimise potential negative impact. For advanced automation cases where AI-based systems interact with humans, grading schemes can be used, which address aspects such as the predictability of an AI-based system, its reliability in performing its tasks, its competence in dealing with similar future situations, and trust by users in the overall system. Accountability also implies that tensions between requirements or the interests of stakeholders must be traded off in a rational and methodological manner. Trade-offs should be explicitly acknowledged and documented as part of the risk management, supporting the continuous review of their appropriateness. Furthermore, accountability concerns the possibility to redress an AI-based system, if it has contributed to adverse outcomes.
- **Responsibility.** Responsibility generally means that persons in charge of tasks accept the consequences of their actions/decisions to undertake the tasks, whether they result to be eventually right or wrong. When translating this concept of responsibility to AI-based systems, decisions issued by the system in question must be legally compliant, ethical, and traceable to an accountable person or organisation. A responsible AI-based system requires ensuring auditability and accountability during its design, development and use, according to specifications and the applicable regulation of the domain of practice in which the AI system is to be used (Díaz-Rodríguez et al., 2023). In operational concepts with increasing levels of automation there is a shift in the level of authority of human operators to AI-based systems. This shift implies a shift in responsibility from human operators to shared responsibility between human operator and system, and to full responsibility of the system at the highest level of automation. Responsibility in advanced automation concepts thus requires ensuring auditability and accountability of the human-system integration aspects in the operations,

including system design, development, manufacturing and maintenance, human-machine interfaces, training of human operators, explainability, etc.

- **Liability.** Liability is the state of being legally responsible for something, e.g. a manufacturer's legal responsibility to the consumer of its product. In complex organisations liability and responsibility can be tied to potential faults or accountabilities of multiple stakeholders, which is known as the problem of many hands (Thompson, 2017). Advanced automation concepts that include shifts in authority and responsibility can lead to uncertainty and concern about responsibility and liability in the case of incidents and accidents. This can have an impact on the safety culture and in particular the just culture in organisations like ANSPs and airlines (Kirwan, 2024). He argues that if just culture is to be preserved, rationales and arguments need to be developed that will stand up in courts of law. These must protect crew and workers who made an honest (i.e., a priori reasonable) judgement about whether to follow AI advice, and whether to intervene, contravening AI autonomous actions seen as potentially dangerous. AI regulatory sandboxes are test environments described by the AI Act in Article 57 (EU, 2024). They act as test beds and safe playgrounds that allow assessing the compliance of AI systems with respect to regulation, risk mitigation strategies, conformity assessments, accountability and auditing processes established by the law (Díaz-Rodríguez et al., 2023). Such sandboxes support pre-market auditability and conformity checks, as well as post-market monitoring and accountability.
- **Safety.** Safety has been and maintains the prime KPA in aviation certification, also for advanced automation and AI-based systems. As described in (HUCAN D3.1, D3.2, 2024) functional hazard assessment (FHA) approaches lead to ranges of requirements for system design and development using allocation of assurance levels, but they do not provide the means to assess in detail whether safe performance has been sufficiently attained for a specific advanced automation operational concept. In particular, methods are needed that assess the detailed functioning of the AI-based systems, while interacting with other systems and human operators in a traffic environment. The dynamics, feedback loops and typically non-linear behaviour of the interacting agents need to be accounted for in situations with normal variability, as well as in situations with non-nominal or failure conditions. The overall performance and risks that may emerge in such operational concepts need to be assessed and evaluated.
- **Resilience.** In a holistic framework the level of resilience of AI-based advanced automation operations needs to be considered. A sociotechnical system is resilient, if it can adjust its functioning prior to, during, or following changes and disturbances, and thereby sustain required operations under both expected and unexpected conditions (Hollnagel, 2014). Resilience Engineering is the discipline that focuses on developing principles and practices to support resilience of sociotechnical systems (Hollnagel et al., 2006), so as to support safety and efficiency of its operations. In a review of resilience papers by Bergström et al. (2015) it was identified that overall the prime need for resilience is considered to be the complexity of modern sociotechnical systems and their inherent risks, the prime object of resilience is the capacity to adapt, so as to keep the complex and inherently risky system within its functional limits, and the prime subject is the individual. As such, analysing and improving resilience has links to various KPAs, including human factors, safety, and efficiency.
- **Security.** Security incidents, i.e. intentional events by attackers that may lead to operational interruption or disruption, have to be avoided, as they pose a risk for safety and the continuity of operations. The use of (AI-based) advanced automation in highly connected sociotechnical

systems poses new cyber-security threats. Evaluation and control of security risks is an important component in the development of AI-based systems for advanced automation.

- **Environmental sustainability.** The validation of sustainability criteria for environmental impacts of the advanced automation and AI-based systems should be incorporated in the validation framework, as laid out in the AI Act (EU, 2024). It entails that the system's development, deployment and use process, as well as its entire supply chain, is assessed with regard to its environmental impact, such that the most environmentally friendly choices can be made. Existing guidelines for environment assessment like (SESAR JU, 2024a) focus on the environmental impact of flight operations and on changes due to adaptations in the operations. For the impact of AI-based systems such assessments may be extended by assessment of impact of the AI ecosystem.
- **Societal sustainability.** The validation of sustainability criteria for societal impacts of the advanced automation and AI-based systems should be incorporated in the validation framework, as laid out in the AI Act (EU, 2024). It includes AI ethics like passenger safety, third-party safety, fairness, privacy, transparency and human oversight.
- **Efficiency.** Efficiency relates to the costs and benefits for stakeholders. In ATM, efficiency also addresses matters like punctuality, and number of flight movements (traffic volume) processed (per hour, per year, per inbound peak, etc). Here, often a trade-off can be observed between efficiency and safety: if the number of flight movements in a given time period is increased, a higher number of conflicts between those flights can be expected, and vice versa.

3.2.4 Identify certification objectives

In this step relevant objectives are identified that have been defined by the certifying authority. For AI-based systems and advanced automation concepts the key basis for these objectives are the objectives of the EASA AI guidelines with guidance for ML applications (EASA, 2024). An overview of the topics for these objectives is listed in Table 2 in Section 2.2, while a list of all objectives and associated key performance indicators is in Appendix A of (HUCAN D4.2, 2024). The scope of (EASA, 2024) is limited to Level 1 and Level 2 AI applications, as well as to machine learning techniques (particularly supervised learning). The guidance material is expected to be extended to AI Level 3 and to a broader set of AI technology, including reinforcement learning, logic- and knowledge-based approaches, and hybrid AI. As such the set of objectives for certification is expected to grow given these extensions. Also the objectives for topics that are in the current scope may change as result of continuing research on the requirements for trustworthy AI in aviation and ATM.

Considering the current set of objectives in (EASA, 2024) it can be recognised that they cover a broad scope of KPAs and have various levels of granularity. For instance, for safety and information security risk there three objectives per KPA, addressing that assessments should be done and be supported by suitable data and metrics. For AI assurance and HF large sets of objectives are defined, which describe in detail objectives for data handling, learning processes, and interactions of humans with AI-based systems. A mapping of applicable objectives has been made depending on the level of automation, where particular objectives are only addressed for higher levels. In a systematic analysis of the applicability of the objectives to four use cases in HUCAN, it was found that in the range of 65% to 86% of them are relevant (HUCAN D4.2, 2024). Interestingly, there are also objectives that are defined out of the scope for particular levels of automation in (EASA, 2024), but that are considered relevant for use cases, such as particular objectives for ethics and human factors. Furthermore, the applicability of each relevant objective was assessed for the technology readiness level of each use case. Here it was

found that in the range of 6% to 37% of the overall sets of objectives are applicable for the TRL of the use case.

For the identification of certification objectives in the Assessment Compass, it is worthwhile to consider the relevance of all objectives of (EASA, 2024) and not to limit the set a priori based on the level of automation. Next, an evaluation needs to be made of the applicability of the objectives given the level of automation, the type of AI system (e.g. using supervised learning or not), the relevant KPAs, and the maturity of the design. Certification objectives that are not considered relevant in the current holistic assessment cycle must be listed, as they may become relevant at a later assessment cycle. Further guidance on how the HUCAN framework can support achieving objectives of the EASA AI guidelines is provided in Chapter 4.

3.3 Holistic Assessment Cycle

A detailed overview of the processes in the HUCAN holistic framework is shown in Figure 7. It shows the design of an AI-based system and the associated advanced automation ConOps, the assessment compass, the holistic assessment cycle, and the feedback processes for the design, where a transition to a higher technology/human readiness level of the system and ConOps is supported.

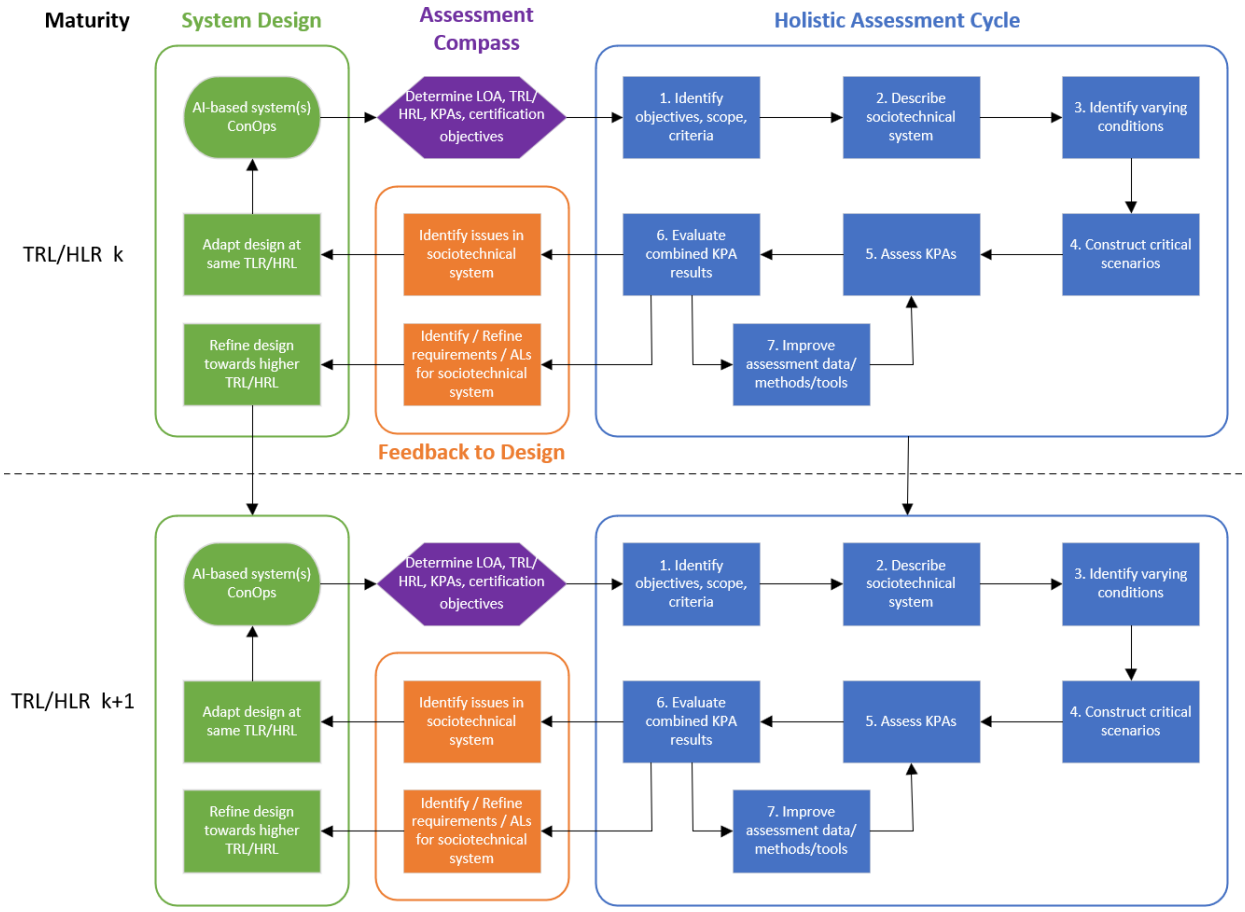


Figure 7. Overview of relations between system design and feedback to design from a holistic assessment cycle in support of transitioning to next technology/human readiness levels.

As highlighted, the holistic assessment cycle consists of the following subsequent steps:

1. Identify objectives, scope, criteria
2. Describe sociotechnical system
3. Identify varying conditions
4. Construct critical scenarios
5. Assess KPAs
6. Evaluate combined KPA results
7. Improve assessment methods/tools (as a feedback loop from step 6 to step 5).

These steps are explained in the following sections.

The framework is supported by the toolbox of methods/tools that is described in Appendix A. An overview of these methods/tools and their associated KPAs, TRL/HRL, and LOAs is provided in Table 4.

Method	KPAs	TRL/HRL	LOA	Section
ABMS (Agent-Based Modelling & Simulation)	Safety, Security, HF, Resilience	TRL 2-9 HRL 2-8	0 - 5	A.2
AI RMF (AI Risk Management Framework)	Accountability, Responsibility, HF, Safety, Security	TRL 4-9 HRL 4-9	0 - 5	A.3
BUSA (Bias, Uncertainty and Sensitivity Analysis)	All	TRL 2-9 HRL 2-9	0 - 5	A.4
Environmental Assessment of AI Ecosystem	Environmental sustainability	TRL 3-9	0 - 5	A.5
FMEA (Failure Modes and Effects Analysis)	Safety	TRL 3-6	0 – 5	A.6
FRAIA (Fundamental Rights and Algorithms Impact Assessment)	Societal sustainability	TRL 4-9 HRL 4-9	0 - 5	A.7
HAZOP (Hazard and Operability study)	HF, Safety	TRL 3-6 HRL 3-6	0 - 5	A.8
Heuristic Evaluations	HF, Safety, Efficiency	HRL 3-6	0 - 4	A.9
HITL (Human-In-The-Loop) Simulations & Wizard of Oz	HF, Safety, Efficiency	HRL 5-9	0 - 4	A.10
HTA (Hierarchical Task Analysis)	HF, Efficiency	HRL 3-6	0 - 4	A.11
NSV-4 diagram (System Functionality and Flow model)	Safety	TRL 2-6	0 - 5	A.12
Responsibility & Liability Analysis	Liability, Responsibility, Accountability	TRL 4-9 HRL 4-9	0 - 5	A.13

Safety Scanning and Security Scanning	Safety, Security	TRL 1-6 HRL 1-6	0 - 5	A.14
SecRAM (Security Risk Assessment Methodology)	Security	TRL 2-6	0 - 5	A.15
Usability Testing	HF, Safety, Efficiency	HRL 3-6	0 - 4	A.16

Table 4. List of validation methods in the toolbox and associated KPAs, TRL/HRL, LOA, and a reference to the associated section in Appendix A.

3.3.1 Identify objectives, scope, criteria

Objectives. The objectives of the holistic assessment are defined in coordination with relevant stakeholders and commensurate with (and proportional to) the technology and human readiness levels of the sociotechnical system as well as the level of automation. These objectives include the definition of the KPAs that will be included in the assessment (see Section 3.2.3) and relevant certification objectives (see Section 3.2.4). The objectives set may depend on the TRL/HRL of the system design. For instance, at low readiness levels it may be decided to exclude particular KPAs, since the details of the system design are not sufficiently known for meaningful assessment.

Scope. The operational scope of the assessment is defined, expressing the boundaries of the operational area considered, and the types of functions or the types of equipment/procedures/people that are included. The scope may depend on the TRL/HRL of the system design. For instance, at low readiness levels a broad scope, involving the global performance of various interacting agents, may be used, while at higher readiness levels a more focused scope may be used to study particular agents in detail, e.g. human-AI interaction in selected scenarios.

Criteria. In coordination with relevant stakeholders absolute or relative performance criteria for the KPAs, which define acceptable versus unacceptable performance, are adopted. Relative criteria specify that the performance of the sociotechnical system for a particular KPA should improve by a certain extent or should not degrade by more than a certain extent with respect to a reference (e.g. an existing system). Absolute criteria specify the required performance at a fixed scale. A well-known example of absolute criteria is a risk matrix, as shown in Table 5. This defines the acceptability of combinations of severity and likelihood levels of scenarios that may occur in the sociotechnical system.

As there are typically multiple KPAs that are considered in a holistic assessment cycle, there are multiple criteria that need to be defined, for instance criteria for safety, security, sustainability, resilience, and responsibility. These criteria are used in the last step of the holistic assessment cycle (Section 3.3.6), where the combined results of the KPA assessments are evaluated as a basis for the feedback to design (Section 3.4). Basically, a refinement of the design to a higher TRL/HRL is supported if the KPA results are sufficiently acceptable, while an adaptation of the design at the same TRL/HRL is needed if the KPA results are not acceptable.

Likelihood level	Severity level				
	1	2	3	4	5
A	Unacceptable	Unacceptable	Unacceptable	Unacceptable	Tolerable

B	Unacceptable	Unacceptable	Unacceptable	Tolerable	Acceptable
C	Unacceptable	Unacceptable	Tolerable	Acceptable	Acceptable
D	Unacceptable	Tolerable	Acceptable	Acceptable	Acceptable
E	Tolerable	Acceptable	Acceptable	Acceptable	Acceptable

Table 5. Example of a generic risk matrix with criteria for acceptability of likelihood-severity combinations.

3.3.2 Describe sociotechnical system

In this step the sociotechnical system, including the AI-based system and advanced automation, is described. It involves the objective of the operation, the operational context, environmental conditions, the functioning and interface of the AI-based system, the functioning and interaction of other technical systems, the roles, tasks and responsibilities of human operators and their interaction with all relevant technical systems (including the AI-based systems). It also explicitly includes assumptions and constraints in the description of the sociotechnical system. It serves as an agreed, documented basis for the KPA assessments.

The main input for the description is documentation on the system design on the ConOps and the AI-based system(s) considered. Additional information may be needed, and this can be formalised as assumptions and constraints in the assessment cycle (which may become requirements in the feedback to design). As the systems and ConOps mature at higher TRLs/HRLs, the descriptions typically become more detailed. This higher level of detail also reflects the requirements and assurance levels that have been set in the feedback to design at a lower TRL/HRL.

3.3.3 Identify varying conditions

The purpose of this step is to identify all kinds of disturbances and performance variability that can influence operations of the sociotechnical system. These can include frequently occurring conditions, like normal sensor errors, normal transmission delays, typical reaction times of human operators, differences in interpretations by humans, normal weather variability, but they also include rarer conditions, like system failures, extreme weather, particular errors by human operators. Varying conditions are identified with respect to all (possibly AI-based) technical systems, human operators and their interactions in the operational context. Sources for the identification of varying conditions include lists of hazards and issues, safety studies, AI and HF literature, and brainstorm sessions.

The identification of varying conditions is performed for a sociotechnical system at a particular TRL/HRL and for the KPAs that are in the scope of the assessment cycle. If the sociotechnical system gets more mature, with more detailed information on its (AI-based) systems and the relations between humans, equipment and procedures, then more specific types of varying conditions can be identified, whereas more abstract varying conditions are identified in early readiness levels.

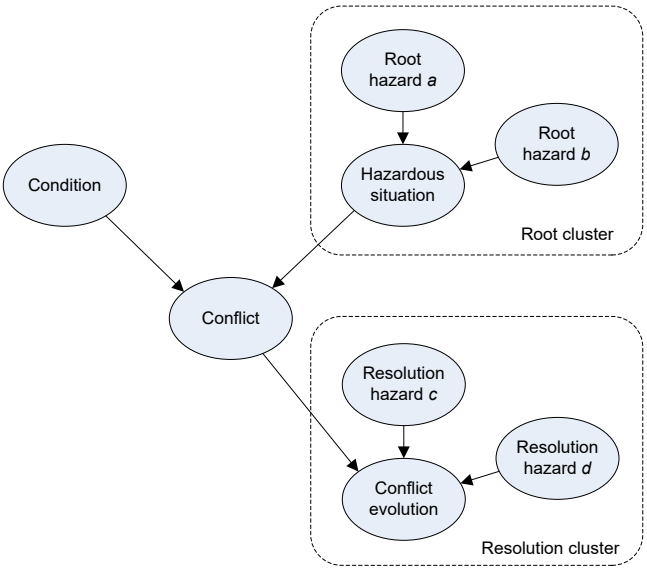
3.3.4 Construct critical scenarios

The purpose of this step is to construct scenarios that represent a critical impact on a KPA, e.g. a scenario leading to small distance between a pair of aircraft (safety), a scenario leading to an environmental problem, a scenario leading to a liability issue, etc. The critical scenarios are expanded by describing how agents of the sociotechnical system and related varying conditions may contribute

to the KPA-critical effect. The aim of these critical scenarios is to bring into account all relevant ways by which varying conditions for the AI-based system and other agents in the sociotechnical system may have an impact on a KPA.

For each operational condition (e.g. certain flight phases and associated geographical locations) that fall within the scope of the assessment, relevant scenarios are developed that may result from varying conditions, e.g. 'conflict between two aircraft converging on one route'. Such scenarios are then used as focal points for attaching associated varying conditions and effects in KPAs. To cope with the complexity in these scenarios, clusters of similar varying conditions are identified. Such clusters may play a role in multiple scenarios. This way of constructing scenarios on the basis of varying conditions sets a basis for a systematic assessment of KPAs.

An example of the construction of a critical scenario for the KPA safety is shown in Figure 8. The diagram shows a combination of conditions and root hazards that may lead to a conflict, and resolution hazards that may complicate an effective resolution of the conflict, such that it may evolve into some incident or an accident.



Example for a runway incursion scenario:

- Root hazard a: Pilots react on clearance for another aircraft and start crossing;
- Root hazard b: Pilots cross without clearance;
- Hazardous situation: Aircraft crossing runway while it should not;
- Condition: Other aircraft has initiated take-off;
- Conflict: Aircraft taking off while another aircraft is crossing the runway and it should not;
- Resolution hazard c: Pilots of crossing aircraft do not frequently look for conflicting traffic;
- Resolution hazard d: Pilots of crossing aircraft are not tuned to frequency of runway controller communication system;
- Conflict evolution: Possible ways of evolution of the runway incursion conflict, e.g. leading to some incident or an accident.

Figure 8. Example of a scenario diagram for the KPA safety, addressing risks associated to a runway incursion (Stroeve et al., 2008).

3.3.5 Assess KPAs

In this step assessments are made of the constructed critical scenarios (from Section 3.3.4) for the KPAs and associated criteria in the scope of the study (see Section 3.3.1). So, for instance, this may consider a quantitative assessment of safety or security risks, or it may consider a qualitative assessment of responsibility and liability issues. The assessment of KPAs typically takes most effort of the steps in the holistic assessment cycle. It can be supported by a variety of methods and tools, depending on the specific KPAs, on the types of results (qualitative/quantitative), and on the level of uncertainty in the KPA results that is acceptable. There can be feedback from the step "Improve assessment methods/tools" if the uncertainty in the KPA results is considered excessive (see Section 3.3.7).

The assessment of KPAs is supported by the methods in the toolbox as presented in Appendix A. This toolbox is not exclusive, and also other methods may be used for the assessment of particular KPAs. The choice of suitable methods is supported by the overview listed in Table 4, which provides associated KPAs, readiness levels and levels of automation. In addition to this overview the advantages and disadvantages of the methods as documented in Chapter Appendix A should be taken into consideration when choosing appropriate methods for the objectives and scope of the study.

The result of the KPA assessment is an overview of the assessed KPAs, as well as an overview of the uncertainties in the assessment results and/or of the assumptions or limitations of the assessment processes.

3.3.6 Evaluate combined KPA results

The results of the assessment of the KPAs are combined and they are evaluated with respect to acceptability criteria. The explicit evaluation of uncertainty in the results and/or argumentation on the assumptions or limitations of the assessment processes form a key basis for this evaluation. The result of the evaluation may lead to the following types of conclusions and basis for feedback to design.

1. It may be concluded with sufficient certainty that the performance of the AI-supported sociotechnical system is not acceptable for one or more KPAs. This means that the AI-based system design cannot pass the current TRL/HRL and that the AI-based system and/or aspects of the encompassing sociotechnical system would need to be adapted in an additional development cycle. This leads to the step “Identify issues in sociotechnical system” as part of the feedback to design.

2. It may be concluded with sufficient certainty that the performance of the AI-supported sociotechnical system is acceptable for all KPAs. This implies that the AI-supported sociotechnical system is considered suitable at the current readiness level of the design. This leads to the step “Identify/Refine requirements/AIs for sociotechnical system” as part of the feedback to design.

3. It may be concluded that there is insufficient certainty to evaluate the performance of the AI-supported sociotechnical system for one or more KPAs, and this prevents reaching a verdict on the acceptability of the sociotechnical system. In this case one of the following next steps can be taken.

    a. It may be decided that additional data or other methods/tools may sufficiently reduce the level of uncertainty in the assessment results. This leads to the step “Improve assessment data/methods/tools” in the holistic assessment cycle.

    b. If the level of uncertainty is high for several critical scenarios and KPAs, it may be decided to redevelop the AI-based system and/or aspects of the encompassing sociotechnical system. This leads to the step “Identify issues in sociotechnical system” as part of the feedback to design.

3.3.7 Improve assessment data/methods/tools

If the level of uncertainty in one or several KPA assessment results is too high to reach a conclusion on the acceptability of the advanced automation (AI-based) sociotechnical system, it may have been decided to improve the assessment(s). This can be achieved in various ways.

1. Additional information may be gathered to improve the assessment(s), such as supporting data, expert opinion, or related research results from the literature. This new information may reduce the uncertainty in the assessment(s).

2. Models or techniques used in the assessment(s) may be extended, such that the uncertainty in the assessment results may be reduced.

Other assessment techniques or tools (from the list in Table 4 or otherwise) may be applied in an effort to reduce the level of uncertainty in the assessment results for the KPAs.

3.4 Feedback to Design

Based on the decision reached in the step “Evaluate combined KPA results”, feedback to design is achieved by either

1. Identify issues in sociotechnical system, or
2. Identify/Refine requirements/ALs for sociotechnical system.

3.4.1 Identify issues in sociotechnical system

If the performance of the AI-supported sociotechnical system is (potentially or certainly) not acceptable for one or more KPAs, the system design needs to be adapted. In support of such redevelopment the critical scenarios, varying conditions, and actions of agents that have been assessed to contribute to the poor performance for a KPA are shared with the developers. The information on these issues supports the designers in improving the AI-based system(s) and ConOps.

3.4.2 Identify/Refine requirements/ALs for sociotechnical system

If it has been assessed with sufficient certainty that the performance of the sociotechnical system is acceptable for all KPAs in the scope of the study, then the premises of the assessment provide a basis for identifying or refining requirements and assurance levels in the system design towards higher readiness levels.

Assumptions made in the assessment may be transformed into requirements. For instance, if it was assumed that operations are done within a particular speed range, then it may be included as a requirement to operate within such a range. Alternatively, such assumption may be a reason to improve the assessment (by extending the assessed speed range).

In the case that a model of the sociotechnical system was used for the assessment of KPAs, wherein the model explicitly describes the performance of the agents, including nominal and non-nominal modes, then the models form a basis for the identification of requirements for the system design. For instance, if a model for a communication link applies a particular failure rate, then, given that the KPA results are acceptable, this failure rate in the model may be used as an upper bound for a communication link failure rate as a design requirement.

It may follow from the KPA assessment(s), that the performance of particular agents (like human operators or AI-based systems) or groups of agents is critical for the acceptability of the overall performance of the sociotechnical system. This means that the overall performance is acceptable if the performance of the agents is within an expected range, but that it is sensitive for performance

variations, such that it may become unacceptable for agents' performance outside of the expected range. In line with such level of criticality, assurance levels can be set for the performance of agents, which define the level of rigour in assuring that the performance of the agents remains within ranges that are compatible with overall acceptable performance of the sociotechnical system.

The requirements and assurance levels determined at a particular TRL/HRL readiness level *k* are a basis for the refinement of the design towards a higher TRL/HRL *k+1*. Furthermore, the requirements and assurance levels from TRL/HRL *k* may provide a basis for determining the requirements and assurance levels for the more detailed design at TRL/HRL *k+1*.

4 Applying the HUCAN framework for achieving EASA objectives

The purpose of this chapter is to explain how the HUCAN holistic framework (Chapter 3) and associated methods in the toolbox (Appendix A) can be applied in association with the objectives of the EASA AI guidelines (EASA, 2024). For a list of all objectives, see Appendix B.

4.1 Coupling with EASA objectives

4.1.1 Characterisation and classification of the AI application

There are 7 objectives of the Trustworthiness Analysis of (EASA, 2024) that identify end users, AI interactions, the concept of operation, system functions, and the level of automation. These objectives can be addressed in several components of the HUCAN framework.

Obj.CO-01: *The applicant should identify the list of end users that are intended to interact with the AI-based system, together with their roles, their responsibilities (including indication of the level of teaming with the AI-based system, i.e. none, cooperation, collaboration) and expected expertise (including assumptions made on the level of training, qualification and skills).*

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

Objective CO-01 is addressed in the ConOps description in the System Design and in the initial two stages of the Holistic Assessment Cycle. Objective CO-02 is focused on end users' goals and high-level tasks, rather than on detailed tasks (EASA, 2024). It is addressed in the ConOps description in the System Design and in the initial two stages of the Holistic Assessment Cycle. Furthermore, in consideration of the goals of end users, relevant KPAs can be determined in the Assessment Compass.

The activities are also supported by methods in the HUCAN toolbox, such as

- Responsibility & Liability Analysis (Appendix A.13), which starts with identifying the list of end users, and proceeds to identify tasks, roles, and responsibilities of human operators and their interaction with all relevant technical systems (including the AI-based systems).
- Safety scanning and Security scanning (Appendix A.14) show stakeholders the loose ends that require further attention from safe/secure concept development, safety/security oversight, legislation, regulation, safety/security management, operational safety/security, and technology.
- Usability Testing (Appendix A.16), which starts with identifying the list of end users, and identifies the tasks the users will perform using the user interface.

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

This objective is addressed by the initial two steps of the Holistic Assessment Cycle, so as to define the AI system in the light of the scope of the study.

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

The operational domain (OD) considers the operating conditions under which a given AI-based system is specifically designed to function as intended, including operating in normal, non-normal, and emergency conditions. Anticipated MOC-CO-04 in (EASA, 2024) describes the expected scope and level of detail of the ConOps, including end-user-centric descriptions of operational scenarios, task allocation and interaction schemes between end-users and AI-based system, and earlier identified risks, mitigations, limitations and conditions on the AI-based system. Operational scenarios should cover nominal cases, degraded modes of AI-based systems and environmental conditions. In the HUCAN framework this broad objective is addressed by the ConOps description in the System Design and in the initial four stages of the Holistic Assessment Cycle, explicitly including the identification of varying conditions and critical scenarios.

Similar as at Obj.CO-02, HUCAN toolbox methods Responsibility & Liability Analysis (Appendix A.13) and Usability Testing (Appendix A.16) are of use here as well, as they support looking at the task allocation between user and AI-based system.

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

Anticipated MOC-CO-05 indicates that end-user representatives should be engaged in planning, design, validation, verification and certification/approval of an AI-based system (EASA, 2024). It entails that end-user representatives should be engaged in all phases of the HUCAN framework, including system design, determining relevant KPAs, the assessment and evaluation of KPAs, and the feedback to design.

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

Anticipated MOC-CO-06 indicates that this includes the allocation of subfunctions to subsystems, including AI-based constituents (EASA, 2024). This decomposition and allocation can be performed in Step 2 *Describe sociotechnical system* of the holistic assessment cycle.

Supporting methods from the HUCAN toolbox include

- AI RMF (Appendix A.3), which includes methods for AI characterization.
- HTA (Appendix A.11), which aims at functional decomposition at the level of human or automation tasks.
- NSV-4 diagram (Appendix A.12), which looks at the functional decomposition from the perspective of the technical system.

Obj.CL-01: *The applicant should classify the AI-based system, based on the levels presented [by EASA], with adequate justifications.*

Anticipated MOC-CL-01-1 and MOC-CL-01-2 of (EASA, 2024) provide guidance on classifying the AI-based system. In the HUCAN framework, the classification of the level of automation is part of the Assessment Compass.

4.1.2 Safety assessment of ML applications

There are 3 objectives for safety assessment in (EASA, 2024), describing initial and continuous safety assessment processes that are achieved in development and operational phases, respectively.

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

This objective addresses the need for a safety (support) assessment during the development phase. In Section C.2.2.3 of (EASA, 2024) various anticipated MOCs are presented regarding DAL/SWAL allocation; metrics; identification, classification, assessment, and mitigation of (data) uncertainties; failure modes of AI-based systems; verification and links of the safety assessment with performance metrics and generalisation bounds.

In the HUCAN framework such safety assessment is done using the holistic assessment cycle for the KPA safety, the feedback to design, and the supported maturing of the design. The assessment cycle includes the identification of uncertainties and failure modes, and the assessment of their impact on the safety of operations with (AI-based) automation. The feedback to design includes critical issues that need to be improved in the design, and requirements and assurance levels for further development of the design. The HUCAN toolbox includes several methods that support such safety assessment, including:

- ABMS (Appendix A.2), which can be used in safety assessment of a variety of operations and conflict scenarios, including but not limited to air operations (e.g. en-route traffic, self-separating traffic, unmanned systems), maintenance, training and aerodromes (e.g. runway incursions, taxiing traffic).
- AI RMF (Appendix A.3), which can be used in support of design, development, deployment, or use of AI systems to manage risks of AI, including safety risks.
- BUSA (Appendix A.4), which aims to get detailed insight into the effect of all uncertainties including parameter value variations and assumptions adopted hence is an important step in verification and validation of model-based safety risk assessment.
- FMEA (Appendix A.6), which can be used for analysis of failure modes of technical systems, including but not limited to aircraft systems (initial and continuing airworthiness) and ATM/ANS ground equipment, including maintenance. Variations such as SWFMEA (Software Failure Modes and Effects Analysis) or HEMECA (Human Error Mode, Effect and Criticality Analysis), are available, dedicated to the analysis of software and human errors, respectively.
- HAZOP (Appendix A.8), which can be used to discover potential hazards, operability problems and potential deviations from intended operation conditions, and to establish approximate qualitative likelihoods and consequences of events.
- HITL (Appendix A.10), which can be used to evaluate the performance of humans dealing with the AI-based systems and the impact on safety.
- NSV-4 diagram (Appendix A.12), which can be used in support of functional modelling and deriving safety requirements.
- Responsibility & liability analysis (Appendix A.13), which can be used to assess the responsibility and liability of stakeholders in safety-critical scenarios.
- Safety Scanning (Appendix A.14), which can be used to check that all aspects important for safety regulation, safety risk management, safety architecture and operational safety have been addressed.

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

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

Anticipated MOC for above objectives in (EASA, 2024) indicates that data should be collected in support of the monitoring of in-service events to detect potential issues or suboptimal performance trends that might contribute to safety margin erosion, or to service performance degradations. In addition, data should be collected in support of the guarantee that design assumptions hold. Associated metrics, thresholds and target values are needed for safety management.

A continuous safety assessment supports safety management of a fully developed and certified system at TRL/HRL 9. Although the HUCAN framework is mostly focused on supporting certification-aware design, it also includes methods for risk management during operations, such as AI RMF. In support of Obj.SA-02 and Obj.SA-03, requirements for the operational use (at TRL/HRL 9) should be formulated, which enables the identification of relevant data in operations as well as acceptable bound on derived metrics.

Supporting HUCAN Toolbox methods include ABMS, AI RMF, BUSA, and Safety Scanning as explained at Obj.SA-01.

4.1.3 Information security risks management

There are 3 objectives for security risk management in (EASA, 2024), which describe the identification of security risks, mitigations, and fail-safe under security threats.

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

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

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

Anticipated MOCs for these objectives in (EASA, 2024) refer to a report on securing ML algorithms by (ENISA, 2021). Specifically, they refer to evasion and poisoning attacks, which can modify the behaviour of an AI/ML-based system, and to oracle attacks, i.e. espionage of the functioning of AI/ML-based systems. Based on the identified threats, security controls should be applied that are specific to applications using ML, besides the security controls already in place. The verification of the effectiveness of the security controls is a part of the verification in the development cycle.

In the HUCAN framework, security risks and their impact on KPAs can be evaluated in the Holistic Assessment Cycle, providing feedback to design, and supporting maturing of the design. For instance, induced sensitivity in an AI-based system by an evasion attack can be included as a threat in steps 3 and 4, and the assessment in step 5 can determine the impact of such sensitivity on the overall sociotechnical system including the AI-based system.

Supporting methods in the toolbox include:

- ABMS (Appendix A.2), which can be used to assess the impact of security hazards on safety.
- AI RMF (Appendix A.3), which can be used in support of design, development, deployment, or use of AI systems to manage risks of AI, including information security risks.
- BUSA (Appendix A.4), which can be used to assess uncertainty and sensitivity in security risk results, and for the impact on other KPAs, e.g. safety.
- Responsibility & Liability Analysis (Appendix A.13), which can be used to assess the responsibility and liability of stakeholders in critical scenarios including information security scenarios.
- Security Scanning (Appendix A.14), which can be used to check that all aspects important for security regulation, security management, security architecture and operational security have been addressed.
- SecRAM (Appendix A.15), which can be used to identify information security risks, identify security controls, and assess the effectiveness of those controls.

4.1.4 Ethics-based assessment

The EU Commission's High-level Expert Group on AI (2019) provided ethics guidelines for trustworthy AI regarding a set of 7 gears (human agency and oversight; technical robustness and safety; privacy and data governance; transparency, diversity, non-discrimination and fairness; societal and environmental well-being; accountability). In (EASA, 2024) it is explained in detail how these gears are addressed by objectives of the EASA AI guidelines. In particular, some gears are handled by eight specific ethics-based objectives, while other gears are handled by other objectives of (EASA, 2024).

The overarching ethics-based objective is

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

The other seven ethics-based objectives are listed in Appendix B (Obj.ET-02 to Obj.ET-08). They address issues like not creating overreliance, complying with GDPR, avoiding bias in ML, environmental impact, and need for new skills as well as risk of deskilling by end-users.

In the HUCAN framework, ethics-based objectives can be supported by evaluation of their impact in the Holistic Assessment Cycle for the associated KPAs. Based on their evaluation the design of the system and its incorporation in the organisation can be adapted, or requirements can be derived to support design towards a higher TRL/HRL.

Specific methods in the toolbox that can support such assessment include:

- AI RMF (Appendix A.3), which includes methods for ethics-based assessment.
- BUSA (Appendix A.4) can be used to assess uncertainty and sensitivity in environmental impact assessment.
- Environmental Assessment of AI Ecosystem (Appendix A.5), which can be used to assess the environmental impact due to changes in flight operations.
- FRAIA (Appendix A.7), which helps to map the risks to human rights in the use of algorithms including AI-based algorithms, and to take measures to address these risks.

- Safety Scanning and Security Scanning (Appendix A.14) address confidentiality and integrity considerations such as unauthorised disclosure of or access to data.

It is noted that the use of Human-in-the-loop (HITL) Simulations (Appendix A.10) and Usability Testing (Appendix A.16) also includes ethics considerations when making use of human participants in the analysis, and HITL could be used in support of Obj.ET-08 training needs analysis. EASA Gear 7 (accountability) is not linked to an objective, but could be addressed by Responsibility & Liability Analysis (Appendix A.13).

4.1.5 Learning assurance

There are 56 objectives for learning assurance (on various topics DA, DM, LM, IMP, CM, QA, RU, SU) in Section C.3.1 of (EASA, 2024), which describe needs for learning assurance processes following a W-shaped cycle for supervised learning applications, including management of data and learning processes, model training and implementation, and verification of learning, inference models, and data management. These objectives and anticipated MOCs describe in considerable detail the new types of focus points that need to be addressed for assuring 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. In addition to these detailed objectives and means, also in associated projects like CODANN “Concepts of Design Assurance for Neural Networks” (EASA and Daedalean, 2021), ForMuLa “Formal Methods Use for Learning Assurance” (EASA and Collins Aerospace, 2023) and MLEAP ”Machine learning Application Approval” (MLEAP Consortium, 2023) considerable R&D has already been done on methods for learning assurance. Given this wealth of recently developed methods for learning assurance, the focus in the HUCAN project has not been on the identification of additional methods.

In the HUCAN framework, assessment results from the holistic assessment cycle provide the basis for the definition of assurance levels for the processes supporting the learning assurance processes for the system design. Particular methods for learning assurance can be found in (EASA, 2024) as well as in the ForMuLa, MLEAP and CODANN reports.

4.1.6 Development and post-ops AI 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 (EASA, 2023c). The target audience for development and post-ops explainability includes engineers, certification authorities and safety investigators to support the development, and learning from occurrences. There are 9 objectives (Obj.EXP-01 to Obj.EXP-09) and some anticipated MOCs for development and post-ops AI explainability in (EASA, 2024), see also Appendix B. These concern e.g. the inclusion of indications of level of confidence of AI-based systems, the capability for system monitoring to ensure that they remain within specified bounds, and the provision of means to record operation data for post-ops explanations.

In the HUCAN framework a distinction can be made between the AI explainability in the development phase and in the post-operation phase. In the development phase the objectives support obtaining sufficient confidence that overall system performance is acceptable to transfer to a higher TRL/HRL. It means tracking proper system data for the evaluation of the combined KPA results in the holistic assessment cycle, and the adaptation of the design or reduction of the uncertainty in the assessment

methods, if the KPA results are not acceptable. Various methods of the toolbox can be applied, depending on the relevant KPAs.

Post-ops AI explainability is a constituent of an AI risk management system for supporting continuous assessment of operations, including safety assurance in a safety management system. For the system design this implies developing the means that support organisations in achieving suitable system data and interpretations during operations. In the HUCAN framework it means assessing that the post-ops AI explainability objectives and the associated AI risk management can be sufficiently supported by the designed sociotechnical system. The AI RMF and BUSA in the toolbox are expected to support such assessment.

There are several methods in the HUCAN toolbox that support such assessment, including

- AI RMF (Appendix A.3), which includes methods for AI explainability.
- BUSA (Appendix A.4): BUSA-attained knowledge on uncertainty and sensitivity of KPAs, like safety, can be used to determine the need for explainability.
- Heuristic Evaluations (Appendix A.9): It may support the definition of operational data that needs to be recorded for post-ops analysis of interaction between AI-based system and end-users.
- Usability testing (Appendix A.16): It may support the definition of operational data that needs to be recorded for post-ops analysis of interaction between AI-based system and end-users.

4.1.7 Operational AI explainability

There are 10 objectives (Obj.EXP-10 to Obj.EXP-19) for operational AI explainability in (EASA, 2024), which concern the need to provide end users with understandable information on how the AI-based system came to its results, see also Appendix B. Anticipated MOCs for these objectives provide useful details on the objectives and ways to achieve them.

The holistic assessment cycle of the HUCAN framework provides the means to assess and evaluate the explainability performance of AI-based systems and its impact on relevant KPAs (e.g. safety, security, HF). Such assessment can provide requirements for further development of the explainability functions, e.g. defining when explanations are clear, when explanations should be provided, what the level of validity of an explanation must be.

There are several methods in the HUCAN toolbox that support such assessment, including

- AI RMF (Appendix A.3), which includes methods for AI explainability.
- BUSA (Appendix A.4): BUSA-attained knowledge on uncertainty and sensitivity of KPAs, like safety, can be used to determine the need for explainability.
- Heuristic Evaluations (Appendix A.9): It can support analysis and improvement of interfaces between humans and AI-based systems.
- HITL & Wizard of Oz (Appendix A.10) are a prime means to analyse and improve the interface between AI-based systems and end-users, including operational AI explainability.
- Usability Testing (Appendix A.16): It can support analysis and improvement of interfaces between humans and AI-based systems.

4.1.8 Human-AI teaming

There are 11 objectives for human-AI teaming (Obj.HF-01 to Obj.HF-09 and two corollaries) in (EASA, 2024), which concern cooperation or collaboration between humans and AI-based systems (see Appendix B). 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 to jointly achieve a common goal (or work individually on a defined goal) and to solve a problem through a co-constructive approach.

The holistic assessment cycle of the HUCAN framework provides the means to assess and evaluate the performance of human-AI teaming and its impact on relevant KPAs (e.g. safety, security, HF). In particular, assessments can concern (shared) situation awareness, the diagnosis by an AI-based system in complex situations, or the capability of an AI-based system to detect poor decision-making by end users. Such assessment can provide requirements for further development of the human-AI functionalities, e.g. defining means to further support shared situation awareness or recognition of poor decision-making.

There are several methods in the HUCAN toolbox that support such assessments, including

- ABMS (Appendix A.2): It can be used to represent situation awareness of human agents as well as AI-based agents, and to evaluate the implications of decisions and coordination schemes by these agents on KPAs like safety.
- AI RMF (Appendix A.3), which includes methods for Human-AI teaming.
- BUSA (Appendix A.4): BUSA attained knowledge on uncertainty and sensitivity of KPAs, like safety, can be used for diagnosis of complex situations in human-AI interactions.
- HAZOP (Appendix A.8): It can be used to analyse decision-making flows between agents in a human-AI team.
- Heuristic Evaluations (Appendix A.9): It can support analysis and improvement of interactions between humans and AI-based systems.
- HTA (AppendixA.11): It can support analysis of task allocation in human-AI teams.
- HITL & Wizard of Oz (Appendix A.10): It can support analysis and improvement of interactions between humans and AI-based systems.
- NSV-4 diagrams (Appendix A.12): It can support analysis of task allocation in human-AI teams.
- Responsibility & liability analysis (Appendix A.13): It can be used to assess the responsibility and liability of stakeholders in safety-critical scenarios.
- Usability Testing (Appendix A.16): It can support analysis and improvement of interactions between humans and AI-based systems.

4.1.9 Modality of interaction and style of interface

There are 16 objectives (Obj.HF-10 to Obj.HF-25) in (EASA, 2024) that provide design guidance objectives for new modes of human-machine interaction through voice, gesture, or other (see Appendix B).

The holistic assessment cycle of the HUCAN framework provides the means to assess and evaluate the performance of different interaction modalities and its impact on relevant KPAs (e.g. safety, security,

HF). For instance, assessments can concern the error rate in spoken natural language or gesture interpretation and the impact on operations. Such assessment can provide requirements for further development of the interfaces for various modality types.

There are several methods in the HUCAN toolbox that support such assessments, including

- AI RMF (Appendix A.3), which includes methods for AI interfacing.
- BUSA (Appendix A.4): BUSA-attained knowledge on uncertainty and sensitivity of KPAs, like safety, can be used for diagnosis of interaction modes and interface style.
- Heuristic Evaluations (Appendix A.9): The prime purpose of heuristic evaluations is to analyse and improve the interface between AI-based systems and end-users, including interaction modes and interface style.
- HITL simulations (Appendix A.10) can support analysis and improvement of the interface between AI-based systems and end-users, including interaction modes and interface style.
- Usability Testing (Appendix A.16): A main purpose of usability testing is to analyse and improve the interface between AI-based systems and end-users, including interaction modes and interface style.

4.1.10 Error management

There are 5 objectives (Obj.HF-26 to Obj.HF-30) in (EASA, 2024) that describe objectives towards reducing risks of human errors (see Appendix B). It is recognised that AI-based systems can contribute to human errors in several ways, such as over-reliance on the system, error in complex decision-making, unexpected failure modes that are not well handled, or errors due to lack of transparency. The objectives state that the likelihood of errors should be minimised and that it must be possible to detect and correct errors.

The holistic assessment cycle of the HUCAN framework provides the means to assess and evaluate the impact of errors on relevant KPAs. As such it can be determined what types of errors are especially critical and what requirements on error likelihood should be posed to achieve acceptable performance in the operations.

There are several methods in the HUCAN toolbox that support such assessments, including

- ABMS (Appendix A.2) can represent error modes of human agents and evaluate the impact of errors on KPAs like safety. Such knowledge provides a basis for setting requirements on the likelihood of errors in the overall design.
- AI RMF (Appendix A.3), which includes methods for error management.
- BUSA (Appendix A.4) BUSA-attained knowledge on uncertainty and sensitivity of KPAs, like safety, can be used to improve error robustness of the design
- HAZOP (Appendix A.8) can be used to analyse the impact of errors in sociotechnical systems.
- Heuristic Evaluations (Appendix A.9) can support the analysis and design of fault tolerant interfaces and suitable information provision to users in the case of errors.
- HITL simulations (Appendix A.10) can support the analysis and design of fault tolerant interfaces and suitable information provision to users in the case of errors.
- Usability Testing (Appendix A.16) can support the analysis and design of fault tolerant interfaces and suitable information provision to users in the case of errors.

4.1.11 Failure management

There are 4 objectives (Obj.HF-31 to Obj.HF-34) in (EASA, 2024) that describe objectives to support management of failure conditions, such as the provision of information to an end user for failure diagnosis, and the support of an end user to propose and implement a solution for a failure condition (see Appendix B).

The holistic assessment cycle of the HUCAN framework supports assessing and evaluating means of end users to handle failure conditions, and to determine the impact on several KPAs. As such the effectiveness of failure management in the sociotechnical system can be evaluated and feedback towards improving the failure management can be achieved.

Methods in the HUCAN toolbox that support this include:

- ABMS (Appendix A.2): ABMS can represent failure modes of AI-based systems and evaluate the impact of failures on KPAs like safety. Such knowledge provides a basis for failure management strategies.
- AI RMF (Appendix A.3), which includes methods for failure management.
- BUSA (Appendix A.4): BUSA-attained knowledge on uncertainty and sensitivity of KPAs, like safety, can be used to improve failure robustness of the design
- FMEA (Appendix A.6) analyses failure modes of systems and evaluates the impact on safety. Such knowledge provides a basis for failure management strategies.
- HAZOP (Appendix A.8) can be used to analyse the impact of failure modes.
- Heuristic evaluations (Appendix A.9) can support the analysis and design of suitable information provision to users in the case of failures.
- HITL simulations (Appendix A.10) can support the analysis and design of suitable information provision to users in the case of failures.
- Usability Testing (Appendix A.16) can support the analysis and design of suitable information provision to users in the case of failures.

4.1.12 AI safety risk mitigation

There are 2 objectives for AI safety risk mitigation in (EASA, 2024), which concern measures to mitigate residual risks, e.g. by real-time monitoring and passivation of AI systems.

Obj.SRM-01: *Once activities associated with all other building blocks are defined, the applicant should determine whether the coverage of the objectives associated with the explainability and learning assurance building blocks is sufficient or whether an additional dedicated layer of protection, called hereafter safety risk mitigation, would be necessary to mitigate the residual risks to an acceptable level.*

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

In the HUCAN framework Obj.SRM-01 is addressed by the evaluation of KPA results in step 6 of the holistic assessment cycle. If it is determined that safety risks are not acceptable, then associated issues are identified in the sociotechnical system, providing feedback to adapt the design at the same TRL/HRL. One of the options for such adaptation is the inclusion of an additional safety risk mitigation means (Obj.SRM-02) in the system design, building on the identified critical issues. As highlighted in

Figure 7 (page 32) such system redesign should be followed by another assessment cycle, so as to assess its effectiveness and to evaluate potential detrimental impact on other elements of the sociotechnical system. While the EASA objectives are focused on safety risks, the same approach can be applied for other KPAs with unacceptable performance.

Methods in the HUCAN toolbox that support the assessment of relevant KPAs can be used including

- ABMS (Appendix A.2), which provides feedback on issues contributing to remaining safety risks, and it provides a means to assess the effectiveness of mitigating measures.
- AI RMF (Appendix A.3), which includes methods for AI safety risk mitigation.
- BUSA (Appendix A.4): BUSA-attained knowledge on uncertainty and sensitivity in safety risk can be used to determine the need for safety risk mitigation, and to identify issues that contribute most to remaining risk
- FMEA (Appendix A.6) assesses the safety impact of failure modes, potentially addressing unacceptable remaining safety risks, which need to be mitigated.
- HAZOP (Appendix A.8), assesses the safety impact of hazards, potentially addressing unacceptable remaining safety risks, which need to be mitigated.
- NSV-4 diagram (Appendix A.12), which can be used in support of deriving safety requirements.

4.1.13 Organisations

There are 8 provisions (Prov.ORG-01 to Prov.ORG-08) in (EASA, 2024) that guide organisations for the introduction of AI-based systems. These provisions regard review of organisational processes, continuous assessment of information security risks, safety risks, and ethics, support of auditing of AI-based systems, and training and licensing of end-users. The details of the provisions are listed in Appendix B.

These provisions support the management in organisations of fully developed and certified systems at TRL/HRL 9. Although the HUCAN framework is mostly focused on supporting certification-aware design, it also includes methods for risk management during operations, such as AI RMF. Furthermore, in support of the organisational provisions, requirements can be identified in the HUCAN framework for the operational use of the AI-based system, which should be managed by the organisation.

Methods in the HUCAN toolbox

- ABMS tooling (Appendix A.2) can support data-driven continuous safety assessment by evaluation of safety events in operations.
- AI RMF (Appendix A.3), which has a strong focus on risk management processes in organisations.
- BUSA (Appendix A.4): BUSA-attained knowledge on uncertainty and sensitivity in safety risk can support organizations in continuous assessment of assumptions and conditions.
- FRAIA (Appendix A.7) supports organisations in continuous assessment of ethics-based aspects of applying AI-based systems.
- HILT simulations (Appendix A.10) can support the development of training processes for interacting with AI-based systems by end-users.
- Safety Scanning and Security Scanning (Appendix A.14) show stakeholders the loose ends that require further attention from safe concept development, safety oversight, legislation, regulation, safety management, operational safety, and technology.

- SecRAM (Appendix A.15) can support organisations in continuous assessment of information security risks.

4.2 Identifying new objectives for supporting certification-aware design

In above section it was explained how the HUCAN framework and its Holistic Assessment Cycle can support attaining the objectives of (EASA, 2024) during subsequent system design and assessment phases. As explained in (EASA, 2024), it provides a first set of usable objectives, but it does not constitute definitive or detailed guidance. Furthermore, it is focused on Level 1 and 2 supervised learning applications, meaning that higher levels of automation and AI applications using other AI techniques have not yet been explicitly addressed.

The HUCAN framework allows to identify new objectives beyond the set of objectives in (EASA, 2024). In particular, following the evaluation of combined KPA results in the Holistic Assessment Cycle, problematic issues and/or requirements of the sociotechnical system can be identified as feedback to design. These issues and requirements may be generalised and provide a basis for the identification of new types of objectives that would need to be considered for the approval of AI-based systems.

Co-funded by the European Union

5 Conclusions

Artificial Intelligence (AI) and its Machine Learning (ML) constituent are key drivers of innovation, enabling higher levels of automation across multiple domains, improving operational efficiency for complex tasks and supporting human operators and organisations. However, the primary concern for all stakeholders involved in this transition is to establish the necessary conditions and standards to ensure that the solutions meet the stringent certification requirements.

The HUCAN project aims to explore these topics from the perspective of research and development projects. A review of the current technical and regulatory state of the art reveals that currently, certification is predominantly based on prescriptive regulations, which mandate strict compliance with detailed requirements (HUCAN D2.1 & D3.1, 2024). This approach has proven effective in progressively enhancing safety. However, its applicability to highly automated and AI-driven technologies is increasingly being questioned, raising concerns about the suitability of existing certification frameworks.

In light of these considerations, the HUCAN project has carefully analysed the currently proposed innovative certification approaches found in the literature (HUCAN D3.2, 2024). What has emerged is that there is a need for a broad-scope, holistic certification approach, which emphasises addressing human factors for understanding uncertainty and safety risks in sociotechnical systems with diverse levels of automation, the impact on accountability in design and operations, assuring public oversight and collaboration with diverse stakeholders, the incorporation of sustainability criteria for societal and environmental impacts, and data governance policies as part of certification.

The HUCAN framework essentially seeks to bridge validation methodologies with the new objectives outlined by EASA for a given AI level, the maturity level of the concept, and the Key Performance Areas (KPAs) of interest to the stakeholders involved in the research process (Figure 9).

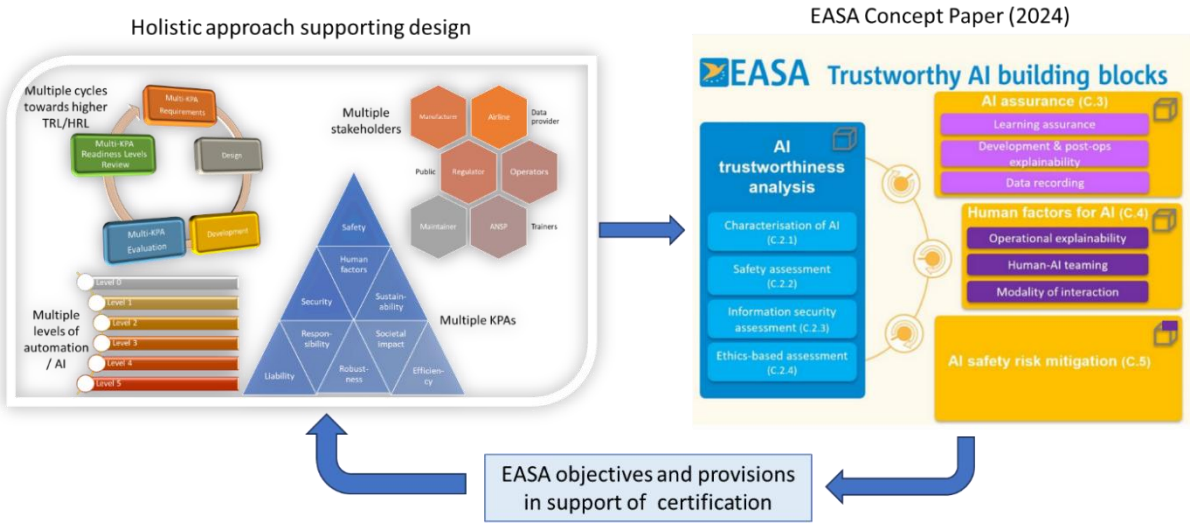


Figure 9. Elements of a holistic approach for certification-aware design.

To support the achievement of these objectives—as well as of other relevant ones identified by the stakeholders involved—the framework proposes a comprehensive suite of methods and tools for the holistic assessment of AI-driven and advanced automation systems. This HUCAN toolbox of methods integrates both established methodologies and innovative approaches, recognizing the inherent complexity and unique challenges of integrating AI into safety-critical systems.

From an operational standpoint, the definition of this approach and the analysis of the characteristics and applicability of widely used validation methodologies across different automation levels and maturity stages reveal that there is no one-size-fits-all solution. The selection of appropriate methods depends on key factors, including the specific KPAs under evaluation, the system's TRL/HRL, and its LOA. The broad spectrum of methods presented reflects the multidimensional nature of the challenge, while the emphasis on method selection and awareness of their limitations highlights the need for a thoughtful, well-informed validation strategy (Figure 10).

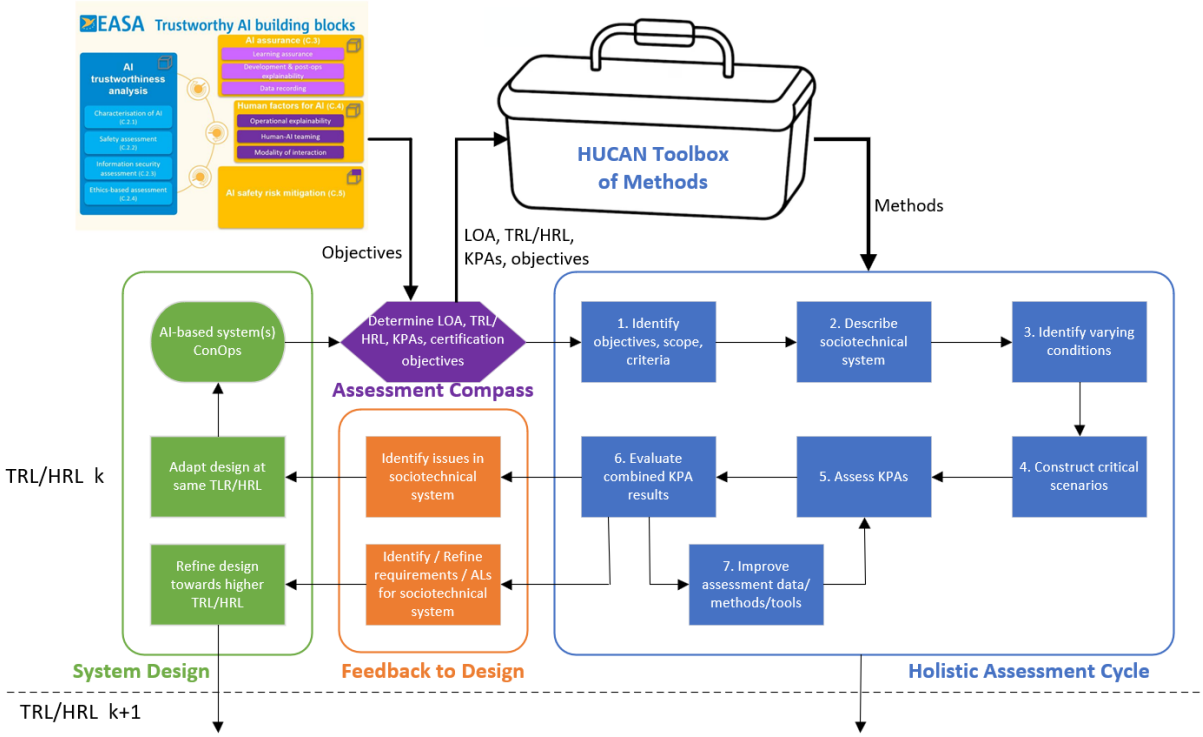


Figure 10. Steps in the HUCAN holistic framework for certification-aware design.

As demonstrated by the integrated analysis in the previous sections and listed in Table 4 in Section 3.3, the methods currently most widely used in the aviation sector generally align with different LOAs. Furthermore, the analysis reveals that although these methods can accommodate various maturity levels, their coverage is uneven, with the early design phases remaining largely under-addressed. These initial stages demand special attention and dedicated research efforts.

A summary of the coverage of the EASA objectives themes by the HUCAN holistic assessment framework and the methods in the toolbox is provided in Figure 11. It follows that the elements of the framework can largely support the objectives themes. In particular, the evaluation of KPAs in the assessment cycle provides feedback to mature the system design for the objectives and KPAs set in

the assessment compass. The methods in the toolbox provide a broad range of support for the EASA objectives themes, with some methods providing wide scope support and other methods being more focused. No methods are included for the theme Learning Assurance in this report and we refer to (EASA, 2024; EASA and Daedalean, 2021; EASA and Collins Aerospace, 2023; MLEAP Consortium, 2023) for a range of associated methods.

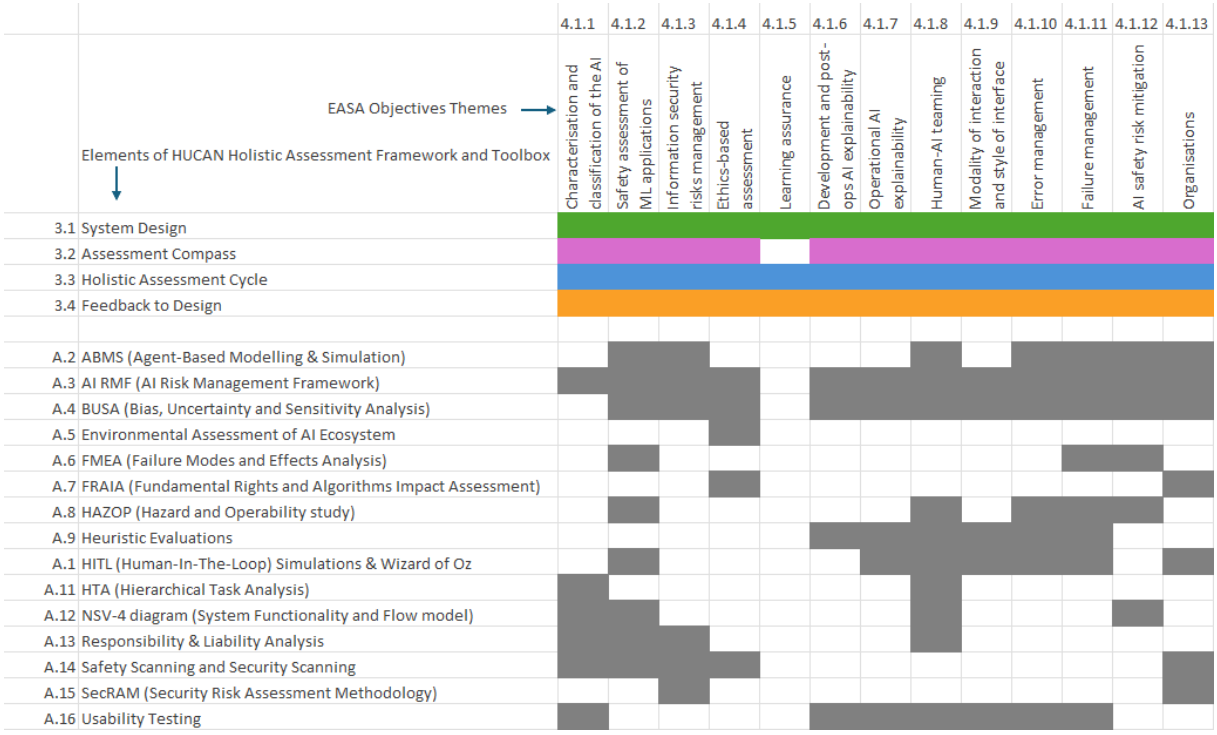


Figure 11. Coverage of EASA objectives themes by HUCAN Holistic Assessment Framework and Toolbox. The numbers at the left-hand side and top refer to sections in this report.

In terms of recommendations, it is important to stress that significant work is still required to further mature the HUCAN holistic framework to further align with EASA's certification objectives. The current version of the HUCAN holistic framework is set at TRL2. Against this background, the framework and its supporting validation toolbox can be used to identify new objectives to which certification-aware design should aspire, given a solution's level of automation and maturity. Additionally, it supports the identification of the most recommended validation methods, based on their applicability to the current state of the solution and the objectives to be pursued. Feedback from application to use cases will be an important contribution in order to further mature the toolbox of methods.

Another recommendation is regarding the anticipated update of the EASA AI guidelines. The current version (EASA, 2024) is mostly focused on supervised learning, and it covers offline learning processes where the model is 'frozen' at the time of approval. The anticipated update might address other types of learning such as reinforcement learning, online learning processes, and Levels of Automation 3A and 3B (i.e. advanced automation). This update of the EASA AI guidelines may come with additional objectives and additional challenges, to be addressed by a future update of the HUCAN Holistic Framework.

6 References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 chi conference on human factors in computing systems*, 1-13. https://doi.org/10.1145/3290605.3300233

Arnaldo Valdes, R.M., Gomez Comendador, V.F., Perez Sanz, L. (2019). Risk assessment under uncertainty, IntechOpen. DOI: 10.5772/intechopen.89445

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.

Beale, C.K. (2006). Uncertainty in the risk assessment process – The challenge of making reasonable business decisions within the framework of the precautionary principle. IChemE Symposium series No. 151. https://www.icheme.org/media/9789/xix-paper-01.pdf

Bergström, J., van Winsen, R., & Henriqson, E. (2015). On the rationale of resilience in the domain of safety: A literature review. *Reliability Engineering & System Safety*, *141*, 131-141. https://doi.org/http://dx.doi.org/10.1016/j.ress.2015.03.008

Bhattacharyya, S., Cofer, D., Musliner, D. J., Mueller, J., & Engstrom, E. (2015). *Certification Considerations for Adaptive Systems.* NASA, NASA/CR–2015-218702.

Blom, H. A. P., & Bakker, G. J. (2012, 17-19 September 2012). *Can airborne self separation safely accommodate very high en-route traffic demand?* Proceedings AIAA ATIO Conference, Indianapolis, Indiana, USA.

Charnes, A., Cooper, W.W., Rhodes, E. (1978). Measuring the Efficiency of Decision Making Units. European Journal of Operational Research. 2 (6): 429–444.

Cooke, Nancy & Demir, Mustafa & Huang, Lixiao. (2020). A Framework for Human-Autonomy Team Research. 10.1007/978-3-030-49183-3_11.

Derby, J. (2023). Designing Tomorrow's Reality: The Development and Validation of an Augmented and Mixed Reality Heuristic Checklist. https://commons.erau.edu/edt/771

Dan Diaper and Neville Stanton (2004) (Editors), The Handbook of Task Analysis for Human-Computer Interaction, Lawrence Erlbaum associates, Publishers, London, 2004

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, *99*, 101896. https://doi.org/https://doi.org/10.1016/j.inffus.2023.101896

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

EASA (2023b). Detailed Specifications and Acceptable Means of Compliance & Guidance Material for certification or declaration of design compliance of ATM/ANS ground equipment, DS-GE.CER/DEC, Issue 1, 26 October 2023

EASA (2023c). *EASA Concept Paper: First usable Guidance for Level 1&2 machine learning applications.* European Aviation Safety Agency, Issue 01, February 2023.

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

EASA (2024b). *Opinion No 03/2024 "Implementation of the regulatory needs in support of the SESAR deployment. Introduction of ACAS Xa for operations & PBN specifications for oceanic operations in the single European sky (SES)"*. European Union Aviation Safety Agency, June 2024

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

EASA and Daedalean (2021). Concepts of Design Assurance for Neural Networks (CoDANN) II, May 2021

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors, 59*(1), 5–27. https://doi.org/10.1177/0018720816681350

ENISA (2021). Securing machine learning algorithms. European Union Agency for Cybersecurity, https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms, December 2021

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

EUROCAE (2021b). *ED-12C: Software considerations in airborne systems and equipment certification.* EUROCAE, February 2021.

EUROCONTROL (2022). *European Airborne Collision Avoidance System (ACAS) Xa Change Proposal (CP)1 validation report*. Version 1.0, 16 June 2022.

European Aviation Artificial Intelligence High Level Group (2020). *The fly AI report: Demystifying and accelerating AI in aviation/ATM.* EUROCONTROL, 5 March 2020.

EU (2024). *Artificial Intelligence Act: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence.* Official Journal of the European Union, 13 June 2024. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Everdij, M.H.C., Blom, H.A.P., Stroeve, S.H. (2006). *Structured assessment of bias and uncertainty in Monte Carlo simulated accident risk*. Proceedings 8th International Conference on Probabilistic Safety Assessment and Management, New Orleans, USA, 14-18 May 2006

Everdij, M. H. C., Blom, H. A. P., Stroeve, S. H., & Kirwan, B. (2014). *Agent-based dynamic risk modelling for ATM: A white paper.* EUROCONTROL, January 2014. http://www.nlr-atsi.nl/downloads/agent-based-dynamic-risk-modelling-for-atm.pdf

Everdij, M. H. C., Blom, H. A. P. (2020). *Safety Methods Database*. Version 1.2, November 2020. Maintained by Royal Netherlands Aerospace Centre NLR, The Netherlands. Available at http://www.nlr.nl/documents/flyers/SATdb.pdf

FAA (2000). System Safety Handbook, December 2000, http://www.faa.gov/regulations_policies /handbooks_manuals/aviation/risk_management/ss_handbook/

FAA (2004). Federal Aviation Administration Safety Management System Manual, Version 1.1, May 21, 2004, http://www.atcvantage.com/docs/FAA_ATO_SMSM_v1.1.pdf

FAA (2024). *Roadmap for Artificial Intelligence Safety Assurance.* Federal Aviation Administration, Version I, 23 July 2024.

Hardy, T. (2004). Assessment of Alternate Methodologies for Establishing Equivalent Satisfaction of the Ec Criterion for Launch Licensing, Terry Hardy, FAA AST-300/Systems Engineering and Training Division, May 19, 2004, http://www.faa.gov/about/office_org/ headquarters_offices/ast/advisory_committee/meeting_news/media/2004/may/Hardy.ppt

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.

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

Hollnagel, E. (2014). *Safety-I and Safety-II: The past and future of safety management*. Ashgate.

Co-funded by the European Union

Hollnagel, E., Woods, D. D., & Leveson, N. (2006). *Resilience engineering: Concepts and precepts*. Ashgate.

HUCAN D2.1 (2024). *Advanced automation in aviation.* SESAR JU, Edition 01.00, 28 February 2024.

HUCAN D3.1 (2024). *Certification methods and automation: benefits, issues, and challenges.* SESAR JU, Edition 01.00, 28 February 2024.

HUCAN D3.2 (2024). *Innovative approaches to approval and certification.* SESAR JU, Edition 02.00, 30 August 2024.

HUCAN D4.1 (2024). *Case studies introduction: level of automation analysis and certification issues.* SESAR JU, Edition 01.00, 28 August 2024.

HUCAN D4.2 (2024). *Performance based requirements for advanced automation.* SESAR JU, Edition 01.00, 29 November 2024.

ISO/IEC/IEEE (2023). *Systems and software engineering — System life cycle processes.* Edition 2, May 2023, 15288:2023.

Kirwan, B. (2024). The Impact of Artificial Intelligence on Future Aviation Safety Culture. *Future Transportation*, *4*(2), 349-379. https://www.mdpi.com/2673-7590/4/2/18

Lanzi, P., Spiller, E., Jameel, M., Lothar, C., Gigante, G., Contissa, G., Sanchi, M., Everdij, M., and Stroeve, S. (2024) Challenges and new directions for the certification of AI and advanced automation in civil aviation. In: *SESAR Innovation Days 2024. SESAR. SESAR Innovation Days 2024*, 12-15 November 2024, Rome, Italy. doi: 10.61009/SID.2024.1.26

Lin, J.Y., Donaghey, C.E. (1993). A Monte Carlo simulation to determine minimal cut sets and system reliability, Proceedings of Reliability and Maintainability Symposium, 1993

Macal, C. M., & North, M. J. (2010). Tutorial on agent-based modelling and simulation. *Journal of Simulation*, *4*, 151-162.

Mankins, J. C. (2009). Technology readiness assessments: A retrospective. *Acta Astronautica*, *65*(9), 1216-1223. https://doi.org/https://doi.org/10.1016/j.actaastro.2009.03.058

Ministry of I&KR (2022). Dutch Ministry of the Interior and Kingdom Relations, Impact Assessment: Fundamental rights and algorithms, https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms

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

NASA (2002). Fault Tree Handbook with Aerospace Applications, August 2002, https://www.mwftr.com/CS2/Fault%20Tree%20Handbook_NASA.pdf

NASEM (2022). *Human-AI Teaming: State of the Art and Research Needs.* National Academies of Sciences, Engineering, and Medicine, (978-0-309-27017-5). https://www.nap.edu/catalog/26355/human-ai-teaming-state-of-the-art-and-research-needs

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.

Nielsen (2024, Jan 30). *10 Usability Heuristics for User Interface Design.* https://www.nngroup.com/articles/ten-usability-heuristics/

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

NIST (2024). *AI RMF Playbook.* National Institute of Standards and Technology. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

NoMagic (2025). https://docs.nomagic.com/display/UAF12P2024x/NATO+Systems+Viewpoint

Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G., Chin, M.H. (2018). Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med. 2018 Dec 4;169(12):866–872.

RTCA DO-178C (2011). Software considerations in airborne systems and equipment certification, prepared by SC-205, issued 12-13-2011.

RTCA DO-278A (2011). Software integrity assurance considerations for communication, navigation, surveillance and air traffic management (CNS/ATM) systems, prepared by SC-205, issued 12-13-2011.

SAE ARP4761 (1994). Guidelines and methods for conducting the safety assessment process on civil airborne systems and equipment, S-18 Committee, Society of Automotive Engineers, Inc. (SAE), March 1994.

SAE ARP4754A (2010). Certification considerations for highly-integrated or complex aircraft systems, Society of Automotive Engineers, Inc. (SAE), December 2010.

SCAN TF (2010). SCAN Task Force, Safety Fundamentals for Safety scanning, Edition 1.1, 11 March 2010, O. Straeter, H. Korteweg.

SCAN TF (2010a). SCAN Task Force, Safety Scanning Tool, Excel-based Tool, 11 March 2010, A. Burrage, O. Straeter, M.H.C. Everdij.

SCAN TF (2010b). SCAN Task Force, Guidance on Interpreting and Using the Safety scanning results, Edition 1.0, 11 March 2010, O. Straeter, G. Athanassiou, H. Korteweg, M.H.C. Everdij.

SESAR JU (2024a). *SESAR Environment Assessment process*. Edition 05.00, 23 July 2024.

SESAR JU (2024b). *DES security risk assessment methodology (SecRAM 2.0).* Edition 03.00.01, 16 April 2024.

SESAR JU (2025). *Maturity gate guidance.* Edition 02.00.00, 24 January 2025

Standfuss, T., Hirte, G., Schultz, M., Fricke, H. (2024). Efficiency assessment in European air traffic management—A fundamental analysis of data, models, and methods, Journal of Air Transport Management, 115 (2024) 102523.

Storey, N. (1996). Safety-Critical Computer Systems, Addison-Wesley, Edinburgh Gate, Harlow, England, 1996

Straatman, J., Muis, I., Bössenecker, G., Koole, R., Draijer, R., van Deijck, N., van Vledder, I. (2024). FRAIA in action: Lessons learned from 15 FRAIA projects at Dutch government organisations,

Stroeve, S., Bakker, B., Villanueva Cañizares, C. J., & Fota, N. (2021). *Close proximity and collision risk assessment of drones and urban air mobility* 11th SESAR Innovation Days, Virtual conference.

Stroeve, S. H., Blom, H. A. P., & Bakker, G. J. (2013). Contrasting safety assessments of a runway incursion scenario: Event sequence analysis versus multi-agent dynamic risk modelling. *Reliability Engineering & System Safety*, *109*, 133-149. https://doi.org/10.1016/j.ress.2012.07.002

Stroeve, S. H., Bosse, T., Blom, H. A. P., Sharpanskykh, A., & Everdij, M. H. C. (2013, 26-28 November 2013). *Agent-based modelling for analysis of resilience in ATM* Third SESAR Innovation Days, Stockholm, Sweden.

Stroeve, S. H., & Everdij, M. H. C. (2017). Agent-based modelling and mental simulation for resilience engineering in air transport. *Safety Science*, *93*, 29-49. https://doi.org/http://dx.doi.org/10.1016/j.ssci.2016.11.003

Stroeve, S.H., Blom, H.A.P., Hernandez Medel, C., Garcia Daroca, C., Arroyo Cebeira, C., Drozdowski, S. (2020). *Modeling and simulation of intrinsic uncertainties in validation of collision avoidance systems*. Journal of Air Transportation 28(4):173-183

Sun, R. (Ed.). (2006). *Cognition and multi-agent interaction: From cognitive modeling to social simulation*. Cambridge University Press.

Thompson, D.F. (2017). Designing Responsibility: The Problem of Many Hands in Complex Organizations. In: Van den Hoven, J., Miller, S., Pogge, T., *Designing in Ethics*, Cambridge University Press, Cambridge.

Van Dam, K. H., Nikolic, I., & Lukszo, Z. (2013). *Agent-based modelling of socio-technical systems*. Springer.

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., & Bai, C. (2022). *Sustainable AI: Environmental implications, challenges and opportunities* Proceedings of Machine Learning and Systems, Santa Clara (CA), USA.

7 List of acronyms and terms

Acronym/Term	Description
ABMS	*Agent-Based Modelling & Simulation*
AI	*Artificial Intelligence*: 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).
AI RMF	*AI Risk Management Framework*
AI system	Machine-based system that is 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, 2024).
ANSP	*Air Navigation Service Provider*
ATM	*Air Traffic Management*
Automation	The use of control systems and information technologies reducing the need for human input, typically for repetitive tasks (EASA, 2024).
BUSA	*Bias, Uncertainty and Sensitivity Analysis*
CBA	*Cost Benefit Analysis*
CNS	*Communication, Surveillance, Navigation*
CODANN	*Concepts of Design Assurance for Neural Networks*
ConOps	*Concept of Operations*
EASA	*European Union Aviation Safety Agency*
EU	*European Union*
FHA	*Functional Hazard Assessment*
FMEA	*Failure Modes and Effects Analysis*
ForMuLa	*Formal Methods Use for Learning Assurance*
FRAIA	*Fundamental Rights and Algorithms Impact Assessment*
FRIA	*Fundamental Rights Impact Assessment*
HAZOP	*Hazard and Operability study*
HF	*Human Factors*
HITL	*Human-In-The-Loop simulations*
HRL	*Human Readiness Level:* Readiness of a technology for use by the intended human users in the specified intended operational environment (HFES, 2021).
HSI	*Human System Integration*
HTA	*Hierarchical Task Analysis*

HUCAN	*Holistic Unified Certification Approach for Novel systems based on advanced automation*
KPA	*Key Performance Area*
KPI	*Key Performance Indicator*
Learning Assurance	*Learning assurance:* All of those planned and systematic actions used to substantiate, at an adequate level of confidence, that errors in a data-driven learning process have been identified and corrected such that the AI/ML constituent satisfies the applicable requirements at a specified level of performance, and provides sufficient generalisation and robustness capabilities. (EASA, 2024)
LOA	*Level of Automation*
ML	*Machine Learning*
MLEAP	*Machine learning Application Approval*
MOC	*Means of Compliance*
NIST	*National Institute of Standards and Technology*
NOV	*NATO Operational Viewpoint*
NSOV	*NATO Service-Oriented Viewpoint*
NSV	*NATO Systems Viewpoint*
R&D	*Research and Development*
SecRAM	*Security Risk Assessment Methodology*
SecST	*Security Scanning Tool*
SESAR JU	*Single European Sky ATM Research Joint Undertaking*
SME	*Subject matter Expert*
SST	*Safety Scanning Tool*
TRL	*Technology Readiness Level*
UC	*Use Case*

Table 6. List of acronyms and terms.

Appendix A Toolbox of methods for holistic validation of AI-based systems and advanced automation

This appendix provides a toolbox of methods that can be used to support the elements of the holistic assessment cycle described in Section 3.3. In this context, a ***method*** is defined as *any technique, method, standard, methodology, database, or model* that can be used in support of evaluation of a KPA in an operation including advanced automation and AI-based systems.

The Toolbox includes established methods as well as more innovative methods. When it comes to application to advanced automation, most methods will have their limitations. This appendix aims to explain these limitations, and outline for which types of advanced automation the method can be used.

Two methodologies have not been included in the toolbox since for many decades they are already considered Means of Compliance for certification of Technical Systems. They have been evaluated on their applicability for Automated Systems in (HUCAN D3.1, 2024). They are:

SAE ARP4761 (Aerospace Recommended Practice document 4761).

- ARP4761 provides guidelines and methods for conducting the safety assessment process on civil airborne systems and equipment, which are developed using the ARP4754A central standard. ARP4761 and ARP745A are applicable to the development phases of the aircraft and its systems.
- The EASA AI guidelines refer to ARP4761 in relation to objectives CO-03 and SA-01, noting that ARP4761 can be used for DAL/SWAL allocation of embedded systems, and in support of safety assessment of non-AI/ML items.

RTCA DO-178C (Software considerations in airborne systems and equipment certification) and DO-278A (Software integrity assurance considerations for communication, navigation, surveillance and air traffic management (CNS/ATM) systems).

- DO-178C and DO-278A aim to provide guidance for the production of software for airborne systems and equipment (DO-178C) and non-airborne CNS/ATM (Communication, Surveillance, Navigation/Air Traffic Management) systems (DO-278A), that performs its intended function with a level of confidence in safety that complies with airworthiness/approval requirements.
- This method is not explicitly referred to in the EASA AI guidelines but it is the industry standard for software.

A.1 Evaluation criteria

Each method in the toolbox is evaluated according to the following criteria:

- **Related KPAs**: To which key performance areas can the method contribute? This refers to the areas listed in Section 3.2.3.
- **Applicable readiness levels**: At what technology and human readiness levels (TRL and HRL) can the method be applied? This refers to the levels listed in Section 3.2.2.
- **Levels of automation**: For what levels of automation can the method be applied? This refers to the levels listed in Section 3.2.1.

- **Link with EASA objectives themes:** For which themes of objectives of (EASA, 2024) can the method be a potential Means of Compliance (MOC)? This refers to the themes listed in Table 2 in Section 2.2.
- **Uncertainty**: How does the method deal with uncertainty in inputs, outputs, assumptions, probabilistic elements, lack of data, etc.? Also are results reproducible if the methods are used by different experts?
- **Technical complexity**: What is the level of technical complexity of the method? What is the level of education/training/expertise required for using the method? What are the levels of traceability and documentability of the results?
- **Benefits**: What are the benefits of the method?
- **Limitations**: What are limitations of the method?

A.2 ABMS (Agent-Based Modelling & Simulation)

A.2.1 Concept

Agent-based modelling and simulation (ABMS) is an approach for modelling and analysing sociotechnical systems by describing the behaviour and interactions of human and technical agents. The overall system behaviour emerges as a result of the individual agent processes and their interactions. ABMS provides a highly modular and transparent way of structuring a model, thus supporting systematic analysis, both conceptually and computationally. ABMS has been used in a wide range of application fields. In aviation safety studies, ABMS has been used for assessment of the risk of collisions and close encounters for a variety of operations and conflict scenarios, including en-route self-separating traffic, runway incursions, and unmanned aircraft systems (Blom & Bakker, 2012; Everdij et al., 2014; Stroeve et al., 2021; Stroeve, Blom, et al., 2013). Typically in these kinds of applications dedicated models and simulation software are developed for performing large numbers of Monte Carlo simulation runs, but it was shown in (Stroeve & Everdij, 2017) that agent-based modelling may also be used in combination with mental simulation to qualitatively analyse the interactions and dynamics in an agent-based model for analysing and improving resilience in air transport.

In support of safety risk assessment, the agent-based models represent the dynamics and stochastic variability of operations involving complex interactions of technical systems, human operators and environmental conditions. In particular, the models represent the performance and variability of systems and humans in normal operations (including normal sensor errors, delays, traffic density variation, weather changes), as well as the impact of off-nominal/failure conditions that may affect the operations (e.g. surveillance systems not working, or having excessive errors). Dedicated techniques, like Interacting Particle System and risk decomposition, have been developed to accelerate the Monte Carlo simulation process such that close proximities or mid-air collisions can be assessed in reasonable time (see e.g. their application for drone operations in (Stroeve et al., 2021)).

This type of agent-based modelling and simulation can support the evaluation of AI-based systems in operations with interrelated technical systems, human operators, and other AI-based systems in an air traffic environment, as illustrated in Figure 12. In this diagram an AI-based system produces output based on sensor data that stem from other agents and the environment, where the sensor data includes errors based on error models. The performance of the AI-based system as well as the performance of other technical and human agents can differ in various working modes. Performance

variability and disturbances are represented in the agents' behaviour, such that the AI-based system is evaluated in a broad context of operational conditions.

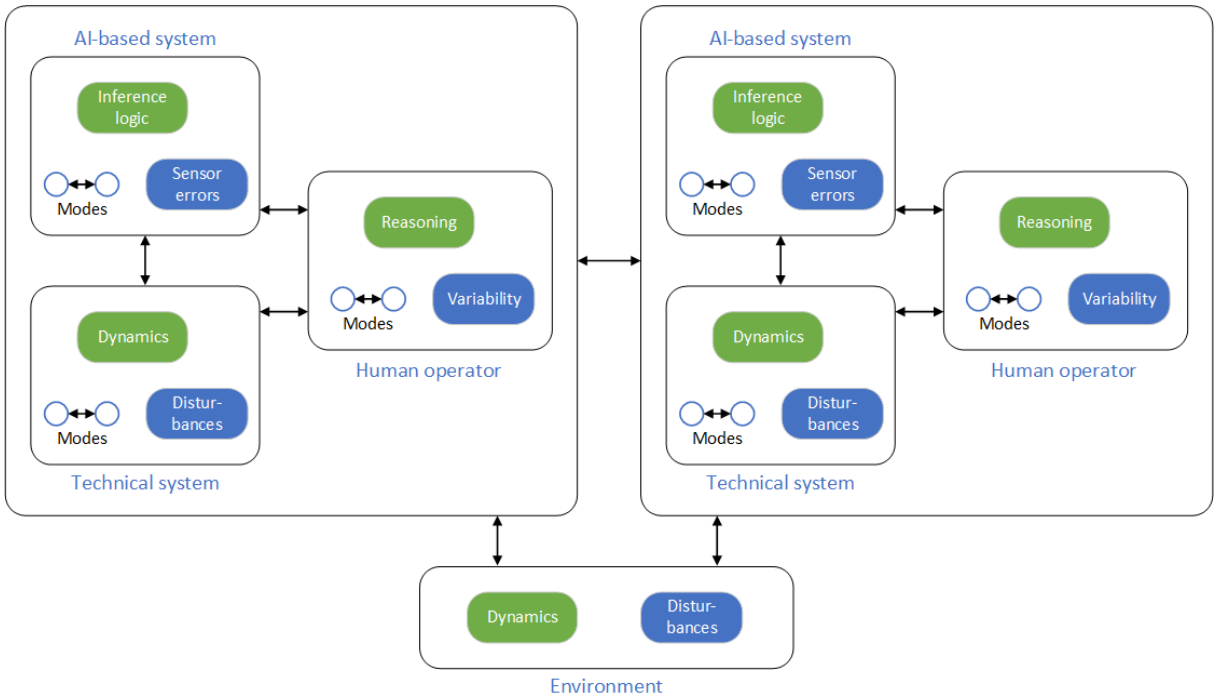


Figure 12. Agent-based modelling and simulation framework for performance assessment of AI-based systems in operations with interrelated technical systems, human operators, and other AI-based systems in an environment.

A.2.2 Method

Key elements of an agent-based approach for holistic performance assessment are the explicit consideration of agents (humans, technical systems), their behaviour and interactions, and disturbances and performance variability that influence the overall performance of the sociotechnical system. Various KPAs may be assessed including safety, security, human factors, and resilience. Depending on the KPAs specific models and simulation techniques may be needed. An agent-based analysis can be applied with various techniques at different levels of sophistication, thus supporting a range of TRL/HRL levels. Next, three agent-based approaches are described: (1) initial agent-based assessment, (2) qualitative ABMS, and (3) quantitative ABMS. These approaches are discussed next in connection with the holistic assessment cycle and feedback to design explained in Chapter 3.

A.2.2.1 Initial agent-based assessment

Steps 1 to 4 of the holistic assessment cycle are followed as described in Section 3.3.

Step 5 "Assess KPAs" now considers a qualitative assessment of the constructed critical scenarios for the KPAs and associated criteria in the scope of the study. Such qualitative assessment can be attained by structured argumentation about varying conditions and interacting agents that can lead to a critical condition in the advanced automation concept, mitigating actions of agents to avoid the negative impact and conditions that may hamper such mitigation. The assessment requires sufficient insight in

the AI-based system and the operations in the sociotechnical system, which may be supported by operational, HF, and technical experts. The arguments in the assessment should be carefully described, including supporting evidence. Uncertainty in the argumentation and its impact on uncertainty in the assessment results should be explicitly presented.

Step 6 “Evaluate combined KPA results” is performed as described in Section 3.3 using the qualitative assessment results.

Step 7 “Improve assessment data/methods/tools” is performed as described in Section 3.3. Means to reduce the level of uncertainty in the assessment results for the KPAs include the application of qualitative and quantitative ABMS approaches presented next in Appendix A.2.2.2 and A.2.2.3.

A.2.2.2 Qualitative ABMS

The qualitative ABMS can build on the initial agent-based assessment explained above in Appendix A.2.2.1. In such assessment it has been concluded that the level of uncertainty in the performance of the sociotechnical system is too high for particular KPAs and scenarios, and that qualitative ABMS may be used to reduce the level of uncertainty. Next the following steps are followed (Stroeve & Everdij, 2017):

1. *Identify objectives, scope, criteria.* The objectives and scope of the qualitative ABMS are determined. This is similar to Step 1 of the initial assessment of Appendix A.2.2.1, but the objectives and scope are typically more restricted so as to focus on the parts where the level of uncertainty needs to be reduced. This step also defines what agents of the sociotechnical system, what varying conditions, and what critical scenarios are included in the scope of the qualitative ABMS.
2. *Identification of model constructs*. A model construct is a generic model describing particular aspects of the ways that agents behave and evolve in interactions with other agents and conditions in the environment. For the identification of model constructs it needs to be understood, which key aspects of the entities in the sociotechnical system drive their behaviour and contribute to the uncertainty in the overall performance, for the conditions in the scope of the ABMS study. Such understanding has been achieved in the initial holistic performance assessment. As a basis for the identification of model constructs, the ABMS literature provides a large variety of agent models (Macal & North, 2010), including for modelling of sociotechnical systems (Van Dam et al., 2013) and social interactions (Sun, 2006). In support of resilience engineering in air traffic management (ATM), Stroeve, Bosse, et al. (2013) developed a library of model constructs for agent-based modelling. This set contains 38 model constructs, which were identified in the ABMS literature and which were evaluated for their capability to support the modelling of a broad range of conditions and events that may contribute to unsafe situations.
3. *Qualitative description of the model details.* In this step, the details of the model constructs are determined at a qualitative level for all agents in the scope. The specifically required model details depend on the model construct considered. The list below provides an overview of the types of details that may be specified in this stage. All these aspects are provided qualitatively using textual descriptions.
    - State variables, such as the position and speed of an aircraft, or the situation awareness of an air traffic controller about the position of an aircraft;

- Mode variables, describing an operating mode of a technical system (e.g. some normal working mode, or a failure mode), or of an human operator (e.g. tactical or opportunistic contextual control mode);
- Types of tasks that a human operator may perform;
- Types of behaviour that an agent may show;
- The way that a model construct is influenced by other model constructs within the same agent (intra-agent input);
- The way that a model construct influences other model constructs within the same agent (intra-agent output);
- The way that a model construct is influenced by model constructs of other agents (inter-agent input);
- The way that a model construct influences model constructs of other agents (inter-agent output).

4. *Mental simulation of the qualitative agent-based model.* The developed qualitative agent-based model provides a structured representation of the sociotechnical system for a particular operational context. Mental simulation employs reasoning on the basis of the developed qualitative agent-based model. Next two approaches are provided (Stroeve & Everdij, 2017).
    a. *Analysis of interactions.* This analysis focuses on interactions during operations that are considered to be normal, or on interactions following some varying condition of interest. It consists of the following steps.
        - As a starting point, an initial condition of agents should be formulated. This initial condition specifies the states and modes of the agents at the start of the mental simulation.
        - One or several triggering events or occurrences of varying conditions are specified, which describe conditions of interest for the critical scenario.
        - Next it is argued what the main changes are in the states and modes of the agents' models following the initial condition. This argumentation is structured by listing sequences of triggers and resulting actions in the agents. Such trigger-action pairs can be internal to an agent (e.g. an observation leading to a decision to coordinate) or it can impose an interaction between agents (e.g. a communication act leading to a change in situation awareness). As such this exercise provides instantiations of sequences of interactions that may occur in the agent-based model. As the state space of the overall model can be extensive, this argumentation is done for the states that are judged to be most relevant for the situation studied.
        - A case of multiple varying conditions leads to multiple instantiations of interaction sequences that need to be accounted for, e.g. a bad weather condition versus a bad weather condition in combination with a technical failure.
    b. *Analysis of dynamics.* This analysis focuses on dynamic relations between states and modes of agents in the sociotechnical system. It consists of the following steps.
        - As a starting point of the analysis it is decided what states or modes need to be studied, e.g. key states identified in an analysis of interactions or other relevant indicators of the overall system performance.
        - An initial condition of the agents' states and modes is specified.

- One or several triggering events or occurrences of varying conditions are specified, which extend the initial condition or occur at a later stage.
- It is qualitatively argued how the relevant states change in time due to the interactions in the agent-based model. The results of this reasoning about the agent states are described in narratives and can be illustrated by graphs as a function of time. These graphs provide qualitative indications of the variation in the selected (aggregate) state variables, which are supported by the argumentation of the elements in the agent-based model that are expected to give rise to them. It can be useful to compare the qualitative graphs for several cases, e.g. a new versus an old operation, or an operation in condition 1 versus condition 2. Also in this type of mental simulation, the dynamics of the complete state space are not described in detail, but rather it is judged during the mental simulation what the most relevant state dynamics are.
- In the case of multiple varying conditions, the above process needs to account for the triggers they induce for the state dynamics.

5. *Conclusions and feedback.* Similar to the initial agent-based assessment, the additional results of the qualitative ABMS may lead to the following types of conclusions and feedback.
    a. It may be concluded with sufficient certainty that the performance of the AI-supported sociotechnical system is not acceptable for one or more KPAs, thus implying that it cannot be certified and further development is needed.
    b. It may be concluded with sufficient certainty that the performance of the AI-supported sociotechnical system is acceptable for all KPAs.
    c. It may be concluded that there is insufficient certainty to evaluate the performance of the AI-supported sociotechnical system for one or more KPAs. There are several possible ways to handle such cases.
        i. Additional analysis may be done using an extension of the qualitative ABMS. This may reduce the uncertainty in the assessment.
        ii. Other assessment techniques may be applied in an effort to reduce the level of uncertainty in the assessment results for the KPAs, such as the quantitative ABMS approach presented next in Appendix A.2.2.3.
        iii. If the level of uncertainty is high for several critical scenarios and KPAs, it may be decided to redevelop the AI-based system and/or aspects of the encompassing sociotechnical system.

A.2.2.3 Quantitative ABMS

The quantitative ABMS builds on the initial holistic assessment of Appendix A.2.2.1 and it may also build on results of qualitative ABMS of Appendix A.2.2.2. In these assessments it has been concluded that the level of uncertainty in the performance of the sociotechnical system is too high for particular KPAs and scenarios, and that quantitative ABMS may be used to reduce the level of uncertainty. It includes the following steps:

1. *Objectives and scope.* The objectives and scope of the quantitative ABMS are determined, see also Step 1 of Appendix A.2.2.1 and A.2.2.2. This step includes what agents, varying conditions, and critical scenarios are considered in the quantitative ABMS assessment.
2. *Development of a formal quantitative agent-based model.* In this step a formal quantitative agent-based model is developed for agents, varying conditions and critical scenarios that are in the scope. It includes the identification of suitable model constructs to describe the

performance of agents in normal conditions as well as relevant off-nominal conditions, the interactions between model elements within agents and between agents, the choice of parameter values of the agent-based models, and the definition of simulation approaches to evaluate performance indicators for the KPAs of interest. Given the range of varying conditions that need to be covered, typically stochastic models are included in the agent-based model and the simulation approaches enable the evaluation of these models (e.g. Monte Carlo simulation). General approaches for quantitative ABMS are provided in (Macal & North, 2010; Van Dam et al., 2013); dedicated approaches for aviation safety risk assessment are in (Everdij et al., 2014; Stroeve et al., 2021).

3. *Implementation of the agent-based model and interfaces.* The developed quantitative agent-based model and the simulation methods are implemented in a suitable computing environment. The agent-based model of the overall sociotechnical system may include models for its AI-based systems, but it may also have interfaces with (extensive) prototypes of AI-based systems at varying levels of maturity. This is the inclusion of an AI system in a simulation environment for the licensing approach described in Appendix A.2.1.
4. *Computer simulation of the agent-based model.* Computer simulations are performed for the agent-based model of the sociotechnical system and its associated AI-based systems to arrive at performance indicators for the KPAs of interest. Given the varying conditions and associated stochastic models, the performance indicators often represent statistics of the performance of the sociotechnical system and its AI-based system, e.g. probabilities of incidents or accidents, mean duration of problematic condition, etc. In addition to such statistics, the computer simulation can provide specific realizations of scenarios that illustrate in detail the performance of the sociotechnical system and its AI-based systems. These results support the assessment and give detailed feedback to designers about the system performance.
5. *Conclusions and feedback.* Similar to the other assessments, the additional results of the quantitative ABMS may lead to the following types of conclusions and feedback.
    a. It may be concluded with sufficient certainty that the performance of the AI-supported sociotechnical system is not acceptable for one or more KPAs, thus implying that it cannot be certified and further development is needed.
    b. It may be concluded with sufficient certainty that the performance of the AI-supported sociotechnical system is acceptable for all KPAs.
    c. It may be concluded that there is insufficient certainty to evaluate the performance of the AI-supported sociotechnical system for one or more KPAs. There are several possible ways to handle such cases.
        i. Additional analysis may be done using an extension of the quantitative ABMS. This may reduce the uncertainty in the assessment.
        ii. Other assessment techniques may be applied in an effort to reduce the level of uncertainty in the assessment results for the KPAs, such as human-in-the-loop simulations.
        iii. If the level of uncertainty is high for several critical scenarios and KPAs, it may be decided to redevelop the AI-based system and/or aspects of the encompassing sociotechnical system.

A.2.3 Evaluation

Related KPAs: Safety, Security, Resilience, HF.

Applicable readiness levels: The associated readiness levels depend on the type of agent-based approach. The initial agent-based holistic performance assessment is especially useful for assessment of concepts at TRL 2-3 and HRL 2-3. The qualitative ABMS approach can support evaluation of concepts and models at TRL 2-4 and HLR 2-4. The quantitative ABMS covers a broad spectrum of TRLs where it can effectively contribute. Using simple models and scenarios, it can already provide feedback-to-design at TRL 2. Using more advanced models and interfaces with development versions of an AI-based system it can support feedback to design at TRL3-7. An advanced quantitative agent-based model & simulation tool can support testing and demonstration in support of approval at TRL 8. Such a tool may also be used to evaluate system performance as part of safety management at TRL 9. An example of such an agent-based tool is the Collision Avoidance Validation and Evaluation Tool (CAVEAT), which was developed by NLR and everis/NTT-Data for EUROCONTROL (Stroeve et al., 2020). CAVEAT was used by EUROCONTROL in a validation study of the AI-based ACAS Xa system for EASA (EUROCONTROL, 2022), which supported EASA in achieving an opinion on the introduction of ACAS Xa in Europe (EASA, 2024b). With regard to HRLs quantitative ABMS can evaluate designs using human performance models at HRL 2-5. Furthermore, it can provide a basis for identification of safety-critical scenarios that can be used in evaluation of human systems design at HRL 6-8.

Levels of automation: It can support all levels of automation LOA-0 to LOA-5.

Link with EASA objectives themes:

- *Safety assessment*: ABMS can be used to assess the likelihood of safety scenarios.
- *Information security*: ABMS can be used to assess the impact of security hazards on KPAs, e.g. safety.
- *Human-AI teaming*: ABMS can be used to represent situation awareness of human agents as well as AI-based agents, and to evaluate the implications of decisions and coordination schemes by these agents on KPAs like safety.
- *Error management*: ABMS can represent error modes of human agents and evaluate the impact of errors on KPAs like safety. Such knowledge provides a basis for setting requirements on the likelihood of errors in the overall design.
- *Failure management*: ABMS can represent failure modes of AI-based systems and evaluate the impact of failures on KPAs like safety. Such knowledge provides a basis for failure management strategies.
- *AI safety risk mitigation*: ABMS provides feedback on issues contributing to remaining safety risks, and it provides a means to assess the effectiveness of mitigating measures.
- *Organisation*. An ABMS tool can support data-driven continuous safety assessment by evaluation of safety events in operations.

Uncertainty: The handling of uncertainty depends on the type of approach. The initial agent-based assessment and the qualitative ABMS are both qualitative approaches. As is typical in qualitative approaches, these have no methods for evaluating uncertainty, except for recognition of uncertainty in the results by their users. In contrast, quantitative ABMS allows to represent and evaluate the impact of a variety of uncertainties in sociotechnical systems, such as sensor errors, delays, and operator performance variability. Importantly it supports the evaluation of the interactions between the technical and human agents in the sociotechnical system, and to understand the implications on KPIs of the nonlinear dynamics. In addition, a systematic approach exists for evaluation of potential bias and uncertainty in ABMS-based risk assessment results due to modelling assumptions and parameter values (Everdij et al., 2006).

Technical complexity: The technical complexity depends on the type of ABMS approach and the scope of the study. The technical complexity of an initial agent-based holistic performance assessment is relatively low. It is a qualitative approach, which uses structured reasoning on a sociotechnical system in scenarios. Although the approach is still qualitative, the technical complexity of the qualitative ABMS is somewhat higher, since it requires knowledge of agent-based model constructs. The technical complexity of quantitative ABMS is considerably higher, since it requires the development of a quantitative agent-based model, its implementation in software, and the use of Monte Carlo simulation to achieve results. For all approaches the technical complexity grows if the scope and the numbers of interacting agents grow. At higher TRL the numbers of agents and types of variability in these agents, which need to be accounted for, are larger, thus increasing the technical complexity.

Benefits:

- The method is recognised by European stakeholders as a potential approach for application to advanced automation, since it is able to evaluate advanced automation operational concepts by testing AI-based systems in interaction with human operators and other systems in large numbers of scenarios that represent normal variability, non-nominal or failure conditions, and contingencies.
- The agent-based perspective supports structured analysis of sociotechnical systems with many interacting technical systems, human operators, and contextual conditions.
- The three levels of agent-based modelling support a range from qualitative to quantitative analyses, where decisions for the type and scope of analysis are based on uncertainty levels in assessment results.
- By using dynamic stochastic models of technical systems and of human performance, and by integrating AI-based systems in simulation environments, quantitative ABMS can provide detailed and traceable risk results, which account for the nonlinear dynamic interactions of relevant agents. Such results cannot be attained by static models, like logic diagrams and fault trees.
- Quantitative ABMS is a means for so-called ‘licensing of AI systems’ (Bhattacharyya et al., 2015). It means that certification is attained by extensive testing in large numbers of simulated hours, including large numbers of faults, and contingencies, where the system demonstrates adequate performance. Licensing-like certification has been indicated as a potential approach in a Fly AI study (European Aviation Artificial Intelligence High Level Group, 2020) and by EASA (EASA, 2024). Advantages recognised by Bhattacharyya et al. (2015) are:
    - *Performance focus.* A licensing approach places more emphasis on the performance of a system, than on its development methodology/process. One of the criticisms of DO-178C is that it focuses more on the development process and producing evidence of compliance, than on evaluation of the resulting software. In a licensing approach there is a shift towards more extensive testing and revision of the actual safety-critical system.
    - *Reduced cost.* Once a high-fidelity simulation environment has been developed, the cost of retesting and re-licensing a new or revised system would become much lower than current certification costs.
    - *Realistic expectations and reduced liability.* A licensed system would not have an implied guarantee of perfection, it would have a proven track record of performance. Properly legislated, realistic expectation could relieve a system developer from the legal liability that may prevent the introduction of advanced technologies.

Limitations:

- The development of a quantitative ABMS environment requires considerable multi-disciplinary expertise for the development of agent-based models, for the software development of the simulation environment, and for the application in a risk assessment.
- The lead time and effort for quantitative ABMS are larger than for qualitative approaches, if the simulation environment has not yet been developed.

A.3 AI RMF (AI Risk Management Framework)

A.3.1 Method

The AI Risk Management Framework (AI RMF) of the National Institute of Standards and Technology (NIST) (NIST, 2023) supports organisations that design, develop, deploy, or use AI systems to manage risks of AI and promote trustworthy and responsible development and use of AI systems; see also Section 2.9 of HUCAN D3.2. In its core, the framework consists of four functions: Govern, Map, Measure, and Manage, and it is supported by a library of methods (AI RMF Playbook) (NIST, 2024).

Figure 13. AI Risk Management Framework of (NIST, 2023).

The *Govern* function cultivates and implements a culture of risk management within organisations that are designing, developing, deploying, evaluating, or acquiring AI systems. It includes the following main categories (see also subcategories in (NIST, 2023)):

1. Policies, processes, procedures, and practices across the organisation related to the mapping, measuring, and managing of AI risks are in place, transparent, and implemented effectively.
2. Accountability structures are in place so that the appropriate teams and individuals are empowered, responsible, and trained for mapping, measuring, and managing AI risks.

3. Workforce diversity, equity, inclusion, and accessibility processes are prioritised in the mapping, measuring, and managing of AI risks throughout the lifecycle.
4. Organisational teams are committed to a culture that considers and communicates AI risk.
5. Processes are in place for robust engagement with relevant AI actors.
6. Policies and procedures are in place to address AI risks and benefits arising from third-party software and data and other supply chain issues.

The *Map* function establishes the context to frame risks related to an AI system. It includes the following main categories (see also subcategories in (NIST, 2023)):

1. Context is established and understood.
2. Categorization of the AI system is performed.
3. AI capabilities, targeted usage, goals, and expected benefits and costs compared with appropriate benchmarks are understood.
4. Risks and benefits are mapped for all components of the AI system including third-party software and data.
5. Impacts to individuals, groups, communities, organisations, and society are characterised.

The *Measure* function employs quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyse, assess, benchmark, and monitor AI risk and related impacts. It includes the following main categories (see also subcategories in (NIST, 2023)):

1. Appropriate methods and metrics are identified and applied.
2. AI systems are evaluated for trustworthy characteristics.
3. Mechanisms for tracking identified AI risks over time are in place.
4. Feedback about efficacy of measurement is gathered and assessed.

The *Manage* function entails allocating risk resources to mapped and measured risks on a regular basis and as defined by the Govern function. Risk treatment comprises plans to respond to, recover from, and communicate about incidents or events. It includes the following main categories (see also subcategories in (NIST, 2023)):

1. AI risks based on assessments and other analytical output from the Map and Measure functions are prioritised, responded to, and managed.
2. Strategies to maximise AI benefits and minimise negative impacts are planned, prepared, implemented, documented, and informed by input from relevant AI actors.
3. AI risks and benefits from third-party entities are managed.
4. Risk treatments, including response and recovery, and communication plans for the identified and measured AI risks are documented and monitored regularly.

In developing, using, and maintaining trustworthy (AI-based) advanced automation, developers/ manufacturers and users of AI-based systems cooperate with other users and/or developers/ manufacturers in the context of general demands and interactions by society, and specific regulations and oversight actions by regulators as well as related guidance and standards. Figure 14 provides a schematic diagram of the interactions between these entities, where it is considered that each organisation that uses or develops/manufactures/maintains AI-based systems apply (to some extent) an AI Risk Management Framework with its core functions Govern, Map, Measure, and Manage.

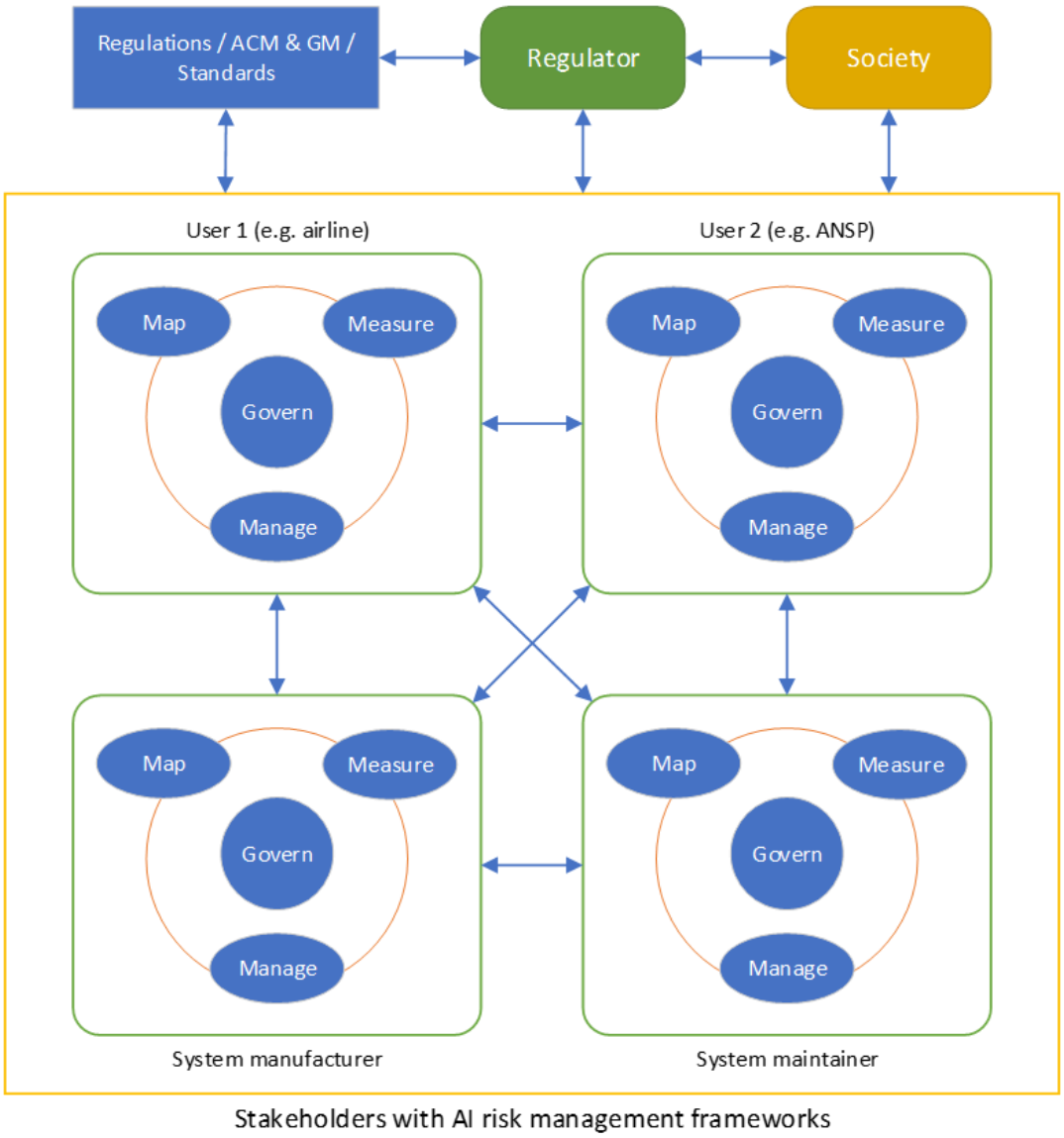


Figure 14. Schematic diagram of interacting system developers and users, which all apply AI risk management frameworks, and their relation with regulator and society.

A.3.2 Evaluation

Related KPAs: Accountability, Responsibility, Safety, Security, HF

Applicable readiness levels: NIST's AI RMF supports organisations designing, developing, deploying, or using AI systems. Applicable readiness levels are not explicitly indicated in (NIST, 2023), but the methods in the framework seem relevant at a wide range of readiness levels except for more basic research: TRL 4-9 and HRL 4-9. Emphasis is placed on the risk management by organisations employing AI in operation at TRL 9 and HRL 9.

Levels of automation: It can support all levels of automation LOA-0 to LOA-5.

Link with EASA objectives themes:

- *Characterization*: AI RMF includes methods for AI characterization.
- *Safety assessment*: AI RMF can be used in support of design, development, deployment or use of AI systems to manage risks of AI, including safety risks.
- *Information security*: AI RMF can be used in support of design, development, deployment, or use of AI systems to manage risks of AI, including information security risks.
- *Ethics-based assessment*: AI RMF includes methods for ethics-based assessment.
- *Development and post-ops AI explainability*: AI RMF includes methods for AI explainability.
- *Operational AI explainability*: AI RMF includes methods for AI explainability.
- *Human-AI teaming*: AI RMF includes methods for Human-AI teaming.
- *Modality of interaction and style of interface*: AI RMF includes methods for interfacing.
- *Error management*: AI RMF includes methods for error management.
- *Failure management*: AI RMF includes methods for failure management.
- *AI safety risk mitigation*: AI RMF includes methods for AI safety risk mitigation.
- *Organisations*: AI RMF has a strong focus on risk management processes in organisations.

Uncertainty: Dependent on the particular method in the AI RMF.

Technical complexity: Dependent on the particular method in the AI RMF.

Benefits: The framework provides a broad range of methods in the AI RMF Playbook (NIST, 2024) that can support safety management of AI-based systems. These provide suggested actions and documentation for all aspects of the Govern, Map, Measure, and Manage steps.

Limitations: The framework has a broad scope. It may be hard to find suitable methods in support of aviation and ATM.

A.4 BUSA (Bias, Uncertainty and Sensitivity Analysis)

If for a risk assessment it is not possible to collect data in reality, valuable feedback can be obtained by developing a model of reality and collecting data from that model. A risk model is a visual or mathematical representation of a system or critical scenario, and a useful tool for communication about the risk. By definition, such model differs from reality at various points and levels: assumptions have been adopted to simplify or generalise reality into a model structure, and parameters have been given values. The model contains various types of uncertainty, including Epistemic uncertainty (deficiencies due to lack of knowledge or information) and Aleatory uncertainty (intrinsic randomness in the data). Therefore, any observations or conclusions made about the model should include an assessment of the combined effect of these differences and uncertainties in terms of bias and uncertainty at the risk level.

The aim of BUSA (Bias, Uncertainty and Sensitivity Analysis) is to get detailed insight into the effect of all biases and uncertainties encountered during a model-based risk assessment. The method aims to assess all variations and assumptions on their bias and uncertainty effect on risk, and to combine the results to get an unbiased estimate of ‘true risk’ and a credibility interval for ‘true risk’. BUSA is an important step in Verification and Validation of model-based risk assessment (Beale, 2006), (Arnaldo Valdes et al., 2019).

In this context, an assumption is defined as any modelling choice with potential to imply a difference between model and 'reality', given the goal of the assessment, see Figure 15. Examples are choices for (quantitative) parameter values, particular hazards not covered by the model (e.g. assumed negligible), model structure choices, and numerical approximations. These assumptions apply to quantitative models as well as to qualitative models.

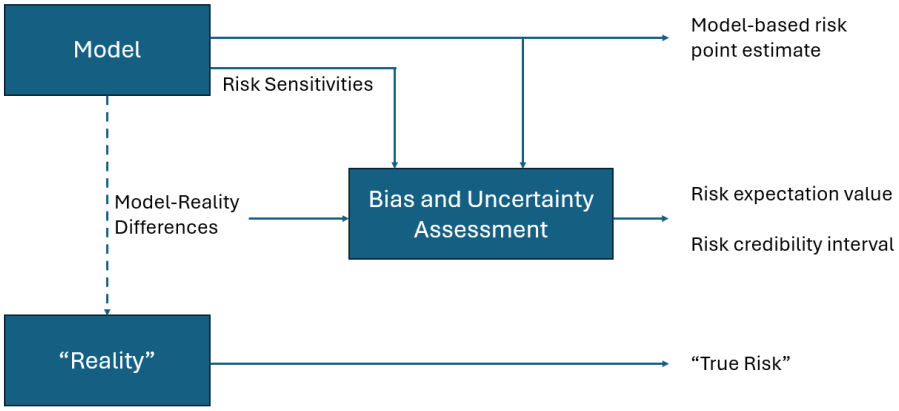


Figure 15. Bias, Uncertainty and Sensitivity assessment approach. Source: (Everdij et al., 2006).

The BUSA method follows four steps (Everdij et al., 2006).

Step 1: Identify all assumptions adopted in the model-based safety risk assessment. The assumptions are divided into two groups:

- Parameter value assumptions, i.e. assumptions that concern a quantitative or qualitative value for a parameter in the model. They are assessed on risk bias and uncertainty in Step 3 below.
- Non-parameter value assumptions, i.e. all other types, such as assumptions due to model structure choices, aspects not covered by the model, numerical approximations, etc. They are identified by systematically comparing the model with the operational design description and other inputs. They are assessed on risk bias in Step 2 below.

Step 2: Determine for each non-parameter value assumption whether its introduction has increased risk or decreased risk, and determine the factor of risk increase or decrease due to the assumption (conditional on all assumptions already assessed). Step 2 is first conducted qualitatively, using terms Negligible, Small, Minor, Considerable, and Major for the factor of risk increase/decrease. For quantitative models the assessments can next be refined quantitatively, using additional statistical data and expert judgment (and restricting to those assumptions that have a non-Negligible bias or uncertainty).

Step 3: Determine for each model parameter value assumption: a 95% credibility interval for the value of the parameter (using statistical information and expert judgment), and the risk log-sensitivity of the value of the parameter. The risk log-sensitivity is a measure for the change in risk, due to a change in a parameter value; this change is determined as a (multiplicative) factor and is typically determined using model simulations with various parameter settings. The credibility interval and risk sensitivity results are combined as input to Step 4. Step 3 is first conducted qualitatively, using terms Negligible, Small, Minor, Considerable, and Major for the size of the credibility interval and for the risk log-sensitivity. For quantitative models the assessments are next refined quantitatively, using additional

statistical data and model simulations (and restricting to those assumptions that have a non-Negligible bias or uncertainty).

Step 4: The outputs of steps 2 and 3 are combined to obtain an estimate for:

- Expected Risk, i.e. model-based risk compensated for bias and uncertainty in the model assumptions.
- A 95% credibility interval for the risk.

The results can be plotted e.g. as in Figure 16, where parameter m is a parameter of interest, such as a separation minimum, and where the risk point estimate is the model-based risk level before BUSA. For qualitative models, the bias and uncertainty is often expressed by an qualitative uncertainty band such as 'Negligible to minor increase'.

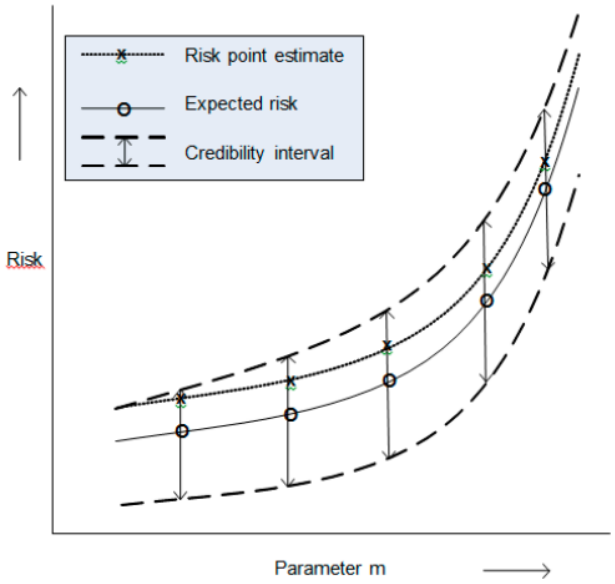


Figure 16. Example result of BUSA for a quantitative risk model.

For qualitative risk models, the sensitivity analysis part is more challenging. However, it is still important to maintain a list of all modelling assumptions adopted and (for KPA safety) maintain a list of all hazards identified.

Related KPAs: All. BUSA can be of value in any model-based risk assessment, irrespective of KPA.

Applicable readiness levels: All. BUSA can be of value at any TRL or HRL.

Levels of automation: It can support all levels of automation LOA-0 to LOA-5.

Link with EASA objectives themes:

- *Safety assessment*: BUSA can be used to assess uncertainty and sensitivity in safety risk results.
- *Information security*: BUSA can be used to assess uncertainty and sensitivity in security risk results, and for the impact on other KPAs, e.g. safety.

- *Ethics-based assessment*: BUSA can be used to assess uncertainty and sensitivity in environmental impact assessment.
- *Development and post-ops AI explainability*: BUSA-attained knowledge on uncertainty and sensitivity of KPAs, like safety, can be used to determine the need for explainability.
- *Operational AI explainability*: BUSA-attained knowledge on uncertainty and sensitivity of KPAs, like safety, can be used to determine the need for explainability.
- *Human-AI teaming*: BUSA-attained knowledge on uncertainty and sensitivity of KPAs, like safety, can be used for diagnosis of complex situations in human-AI interactions.
- *Modality of interaction and style of interface*: BUSA-attained knowledge on uncertainty and sensitivity of KPAs, like safety, can be used for diagnosis of interaction modes and interface style.
- *Error management*: BUSA-attained knowledge on uncertainty and sensitivity of KPAs, like safety, can be used to improve error robustness of the design.
- *Failure management*: BUSA-attained knowledge on uncertainty and sensitivity of KPAs, like safety, can be used to improve failure robustness of the design.
- *AI safety risk mitigation*: BUSA-attained knowledge on uncertainty and sensitivity in safety risk can be used to determine the need for safety risk mitigation, and to identify issues that contribute most to remaining risk.
- *Organisation*: BUSA-attained knowledge on uncertainty and sensitivity in safety risk can support organizations in continuous assessment of assumptions and conditions.

Uncertainty: The purpose of the method is to identify all sources of uncertainty and to analyse the effect on the outcome.

Technical complexity: The complexity of the method depends on the complexity of the underlying operation or model.

Benefits:

- Essential step in any KPA evaluation. Without an idea of the bias and uncertainty in the input and their effect on the output, the output is uncertain which is a risk in itself.
- It includes coverage of the effects of all model assumptions adopted, including their combinations.
- It generates both an expected risk result, and a 95% credibility interval for realistic risk.

Limitations:

- The technique relies partly on expert judgment, hence these results may be subjective. Resources required are: Operational experts who are able to judge (changes in) accident risks; Expert who is able to run the underlying risk model with different parameter settings; Statistical data (or expert judgment-based data) on suitable parameter values, including credibility intervals for these data.
- Some subject matter experts are uncomfortable with assessing assumptions on risk bias and combining the results in a quantitative measure.

A.5 Environmental Assessment of AI Ecosystem

In the development of new ATM operations and supporting systems, environmental impact due to changes in flight operations is assessed (SESAR JU, 2024a). In addition, for concepts using advanced automation with AI-based systems, an assessment of the environmental impact due to the AI ecosystem can be performed. This would allow to compare the possible reduction in environmental impact as a result of more efficient operations as enabled by AI-based systems with the environmental costs of the AI ecosystem.

In (Wu et al., 2022) a holistic perspective on the environmental impact of AI systems is advocated. This concerns a wide view on the AI ecosystem, including machine learning (ML) pipelines, as well as the life cycle of model development and system hardware, including manufacturing and operational use. The ML pipeline must be considered end-to-end, to assess the energy use for data collection, model exploration and experimentation, model training, optimization and run-time inference. The frequency of training and scale of each stage of the ML pipeline must be considered to understand salient bottlenecks to sustainable AI. Wu et al. (2022) provide various examples of gains in efficiency in the ML pipeline that have been achieved for large ML models at Meta. It is indicated that a sustainable mindset is needed where optimization goes beyond efficiency across software and hardware and where competitive model accuracy is achieved at a fixed or preferably reduced computational and environmental cost.

Related KPAs: Environmental sustainability

Applicable readiness levels: Feedback on the environmental impact of the AI ecosystem can be used from early development stages at TRL 3 to system operation at TRL 9.

Levels of automation: It can support all levels of automation LOA-0 to LOA-5.

Link with EASA objectives themes:

- *Ethics-based assessment.* Environmental Assessment of AI Ecosystem especially contributes to Obj.ET-06.

Uncertainty: Levels of uncertainty are not addressed in the results of (Wu et al., 2022), but assessment of uncertainty in energy needs could be added.

Technical complexity: Environmental assessment of the AI ecosystem is a new topic (Wu et al., 2022). It requires detailed knowledge of the environmental implications of its elements.

Benefits:

- Assessment of the impact of the AI ecosystem is a new element that has not yet been considered in ATM environmental impact assessments.

Limitations:

- It is a new approach without detailed guidelines.
- It requires considerable knowledge of technical details of the AI ecosystem.

Co-funded by the European Union

- Given the large energy demand of flight operations, it could well be that the environmental impact of the AI ecosystem is much smaller than the environmental impact enabled by an AI-based ATM system.

A.6 FMEA (Failure Modes and Effects Analysis)

FMEA (Failure Modes Effects Analysis) is a bottom-up reliability analysis technique that considers failures rather than hazards, and hence does not usually consider operating procedures, human factors, and transient conditions. It includes the identification of failure modes, the determination of the effect of the failure mode, a determination of how to detect the failure modes, and an assignment of a failure rate per failure mode (only in case of a quantitative FMEA).

FMEA is used for the analysis of technical systems, and is an established part of SAE ARP4761, the standard for safety assessment of civil airborne systems and equipment. EASA Guidance Material for the certification of ATM/ANS ground equipment (EASA, 2023) states that FMEA should be performed to evaluate the failure conditions of ATM/ANS ground equipment.

FMEA supports safety assessments at various levels, e.g. overall system, system part, and individual component. It is performed at a given level, and can either consider functions (functional FMEA), or actual pieces of equipment (piece-part FMEA). In either form of FMEA, the major steps include preparation, analysis, and documentation.

Preparation of the FMEA includes determining the customer requirements, obtaining current documentation, and understanding the operation of the function. Further information to be obtained includes specifications, current drawings on schematics, parts lists, functional block diagrams, explanatory material regarding the theory of operation, FMEA on a previous or similar function, et cetera.

The *analysis* phase of the FMEA includes the following activities:

- Gaining knowledge on the functions and the design being analysed. Includes reviewing and understanding the information collected during the preparation phase.
- Identification of failure modes. Components and functions that make up the given level are considered on how they may fail.
- Determination of the effect of the failure mode. Consider the effect at the given level and on higher levels. This activity includes the definition of ‘failure effect categories’, corresponding to a unique higher level effect. Use is made of worksheets (see also *documentation* step below).
- Determination of how to detect the failure modes. Usually, but not always included. Detection means are also included in the FMEA worksheet.
- Assignment of a failure rate per failure mode (only in case of a quantitative FMEA). Whenever possible, this is determined from failure data of similar systems already in use.

Documentation of the FMEA is done in a FMEA report, and includes describing the objectives, all inputs and all activities and results. Usually, a FMEA worksheet is used for documentation of the various activities, consisting of a table with column headings such as item, potential failure mode, potential effects of the failure mode, severity of the failure, potential causes of the failure, and likelihood that a potential cause will occur.

FMECA (Failure Modes Effects and Criticality Analysis) is an extension of FMEA that also includes criticality analysis, which allows charting the probability of failure modes against the severity. FMECA is more commonly used than FMEA and is more suited than FMEA for hazard control (FAA, 2000). The main activities to be performed in FMECA are largely in line with the description of the *analysis* step of FMEA above. The main difference is that after the identification of the detection means, the activity of Determining criticality is added. This is usually expressed in a criticality index, which is a combination of the severity of the effect and the probability of occurrence of the failure mode.

Numerous other extensions of FMEA are in use such as DMEA (Damage Mode and Effects Analysis), HEMECA (Human Error Mode, Effect and Criticality Analysis), HF PFMEA (Human Factors Process Failure Mode and Effects Analysis), HMEA (Hazard Mode Effects Analysis), SWFMEA (Software Failure Modes and Effects Analysis), and many others that do not end with "MEA".

AEA (Action Error Analysis) analyses interactions between machine and humans, and is used to study the consequences of potential human errors in task execution related to directing automated functions. It is similar to FMEA and HAZOP (and uses guidewords such as 'omitted', 'too late'), but is applied to the steps in human procedures rather than to hardware components or parts.

Related KPAs: Safety.

Applicable readiness levels: TRL 3 – 6.

Levels of automation: It can support all levels of automation LOA-0 to LOA-5.

Link with EASA objectives themes:

- *Safety assessment*: FMEA contributes to safety assessment.
- *Failure management*: FMEA analyses failure modes of systems and evaluates the impact on safety. Such knowledge provides a basis for failure management strategies.
- *AI safety risk mitigation*: FMEA assesses the safety impact of failure modes, potentially addressing unacceptable remaining safety risks, which need to be mitigated.

Uncertainty: FMEA usually does not consider interactions between system elements or failure modes, and usually does not account for uncertainties in the input data.

Technical complexity: Low

Benefits:

- FMEA is widely considered as a main reliability method for technical systems. The method is systematic and comprehensive, and is supported by standardised forms.
- The method can provide input to a Fault Tree Analysis or a similar numerical method. This can be done in a way that includes analysis of the detection of component failures and the identification of safety-critical equipment where a single failure would be critical for the system.
- FMEA is widely-used and well-understood, and can be performed by a single analyst.

Limitations:

- FMEA focuses on single-point failures of technical systems. It does not consider other hazards, e.g., associated with normal operations, and is not good at identifying hazards caused by humans or the environment. FMEA is specifically not well suited for analysis of human reliability, while in aviation most accidents have a significant human contribution. Also, procedures and processes, and the effects of human mistakes on the functioning of the system are not considered. Accordingly, FMEA is useful for safety-critical mechanical and electrical equipment, but should not be the only hazard identification method.
- The method is not very suitable for complex systems, especially systems that involve dynamic interactions between failures. The method is static, there are no temporal aspects.
- A comprehensive FMEA may be very time consuming and expensive. This specifically holds true for applications to larger systems, for which the use of some form of computer assistance is nearly always necessary. Further factors are that not all component failure modes affect safety on the same level, and that the method may be applied at a level too deep. Duplication of effort and significant amounts of redundant documentation are not uncommon.
- The method sometimes leads to inconsistencies, ambiguities, or difficulties in understanding. One reason for this is that there are sometimes so many failures that they are described in a very brief way. Also, the method does not provide a systematic approach for identifying failure modes or for determining their effects, and no real means for discriminating between alternate courses of improvement or mitigation. Effects that arise from multiple causes are generally not grouped. Information overload from the large but scattered data sometimes obscures the relations in a FMEA. Finally, the benefit of the method depends significantly on the experience of the analyst.
- The method requires a hierarchical system drawing as the basis for the analysis, which the analyst usually has to develop before the analysis can start.
- The method usually does not account for uncertainties in the input data.

A.7 FRAIA (Fundamental Rights and Algorithms Impact Assessment)

The Fundamental Rights and Algorithm Impact Assessment (FRAIA) helps to map the risks to human rights in the use of algorithms and to take measures to address these risks. Here, an algorithm is defined as a set of rules and instructions that a computer automatically follows when making calculations to solve a problem or answer a question, (Ministry of I&KR, 2022). FRAIA is built following international developments including the EU AI Act (EU, 2024) and the EASA AI guidelines (EASA, 2024). The AI Act contains a legal requirement to conduct a Fundamental Rights Impact Assessment (FRIA) when high-risk AI is involved. This legal requirement does not apply to all algorithms because not all algorithms contain AI. The FRAIA is suitable for both algorithms and AI (Straatman et al, 2024).

The aim of FRAIA is to prevent the premature deployment of an algorithm whose consequences have not been properly assessed, resulting in risks such as inaccuracy, ineffectiveness or violation of fundamental rights. It takes the form of a questionnaire with questions of legal, ethical, or technical perspectives, discussing fundamental rights, data and ethics. The discussion on the different questions should take place in a multidisciplinary team consisting of people with different specialisations and backgrounds. For each question, FRAIA indicates who should be involved in the discussion. This tool pays attention to all roles within a multidisciplinary team. It balances the expected positive impact of the algorithm against the expected negative impact on fundamental rights, after which an informed discussion can take place, and a decision can be made on whether or not to deploy the algorithm or whether modifications are necessary and desirable.

Related KPAs: Societal sustainability.

Applicable readiness levels: TRL 4-9, HRL 4-9

Levels of automation: It can support all levels of automation LOA-0 to LOA-5.

Link with EASA objectives themes:

- *Ethics-based assessment*: The main focus of FRAIA is on assessment of ethics and fundamental rights.
- *Organisation*: FRAIA supports organisations in continuous assessment of ethics-based aspects of applying AI-based systems.

Uncertainty: The method is qualitative; uncertainty is not explicitly taken into account

Technical complexity: Low

Benefits:

- FRAIA is seen as a useful tool for discussing fundamental rights, data and ethics. Users appreciate the different perspectives (legal, ethical, technical) and discussions it generates.

Limitations:

- The process may be considered time-consuming, some questions are less relevant and users may struggle to involve all necessary roles.
- Many organisations find it difficult to determine when to conduct a FRAIA. The FRAIA is especially intended for high-risk algorithms.

A.8 HAZOP (Hazard and Operability study)

The aim of HAZOP is to discover potential hazards, operability problems and potential deviations from intended operation conditions. It also establishes approximate qualitative likelihood and consequence of events. It is based on a group review, and is essentially a structured brainstorming using specific guidewords.

The basic notion is that the processes of a sociotechnical system design can be represented by a collection of connected nodes, which can be individually reviewed. A HAZOP study considers various aspects (or parameters) of the operation of nodes and flows between them. In particular, it considers deviations from the intended behaviour, prompted by guidewords, and consequences of these deviations.

The five elements of a HAZOP study are (Storey, 1996):

1. A team of multi-disciplinary ‘experts’, including chairperson, secretary, system designer, engineer, operator/controller, human factors expert.

2. A representation of processes of a sociotechnical system design, in terms of nodes/parameters and flows between them. For the role of a human operator this may be based on a task analysis diagram or a decision flow diagram.
3. A list of guide words, e.g.
   - NO or NONE, meaning a complete negation of the intention
   - REVERSE, meaning the clear opposite of the intention
   - LESS OF/MORE OF, meaning a quantitative decrease/increase
   - AS WELL AS/PART OF, meaning a qualitative increase/decrease
   - SOONER THAN/LATER THAN, meaning intention done sooner/later than required
   - Depending on the type of processes other or additional guidewords may be used like OTHER THAN, REPEATED, MIS-ORDERED, EARLY, LATE.
4. A list of property words. For a technical system these may be e.g. flow, temperature, pressure, concentration, reaction, transfer, contamination, corrosion/erosion, testing. For a human operator these property words could include e.g. Information, Management, Selection, Communication, Input.
5. A recording form to capture information, i.e. a table with the following column headings: Step, Deviation, Cause, Consequence, Indication, System defence, Recommendations.

Related KPAs: HF, Safety

Applicable readiness levels: Since HAZOP uses all types of process descriptions as input, it is best used late in design. However, a preliminary HAZOP can be applied on conceptual process descriptions early in the design stage to avoid later costly problems. A full HAZOP can then be done later in the design process, even if a preliminary HAZOP has already been done. TRL 3-6, HRL 3-6.

Levels of automation: It can support all levels of automation LOA-0 to LOA-5.

Link with EASA objectives themes:

- *Safety assessment*: HAZOP contributes to safety assessment.
- *Human-AI teaming:* HAZOP can be used to analyse decision-making flows between agents in a human-AI team.
- *Error management:* HAZOP can be used to analyse the impact of errors in sociotechnical systems.
- *Failure management*: HAZOP can be used to analyse the impact of failure modes.
- *AI safety risk mitigation*: HAZOP assesses the safety impact of hazards, potentially addressing unacceptable remaining safety risks, which need to be mitigated.

Uncertainty: HAZOP is a qualitative approach that does not explicitly account for uncertainties and assumptions.

Technical complexity: Low

Benefits:

- HAZOP is effective for both technical faults and human errors; it covers human operators in the loop.
- HAZOP can rapidly spot those functionalities whose failure mode effects can be remedied. It recognises existing safeguards and develops recommendations for additional ones.

- Unlike FMEA it does not require the systematic study of the failure modes of each part of the functionality and of their effects.
- It does not concentrate only on failures, but has the potential to find more complex types of hazardous events and causes.
- It provides a systematic coverage and can lead to the discovery of new hazards.
- It encourages creative thinking about all the possible ways in which hazards or operating problems may arise.
- HAZOP is useful in the analysis of complex systems or plants, with which there is yet little experience, and procedures that occur infrequently.
- It can identify design problems at an early stage.
- Only limited training required; HAZOP is an 'intuitive' method.
- It uses the experience of operating personnel as part of the team. The use of a team gives a range of viewpoints and the interaction of several disciplines or organisations provides results that are often overlooked by groups working in isolation.
- HAZOP has a good track record in certain industries; it is widely used and its disadvantages are well-understood.

Limitations:

- It is difficult to assign to each guideword a well-delineated portion of the system and failure causes.
- Due to the systematic approach used and the number of people involved, the method is often time-consuming, and therefore expensive.
- Its success heavily depends on the facilitation of the leader and the knowledge, experience, degree of co-operation and commitment of the team. GIGO (garbage in, garbage out) applies.
- HAZOP cannot easily model dependency between failures.
- It concentrates on single deviations, rather than on cases with multiple deviations or failures.
- It is optimised for process hazards, and needs modification to cover other types of hazards.
- It requires development of procedural descriptions, which are often not available in appropriate detail. However, the existence of these documents may benefit the operation.
- Documentation is lengthy (for complete recording).
- It analyses causes and effects with respect to deviations from expected behaviour, but it does not analyse whether the design, under normal operating conditions, yields expected behaviour or if the expected behaviour is what is desired.
- Deviations from within components or processes are not inspected directly; instead, a deviation within a component is assumed to be manifested as a disturbed flow. Process-related malfunctions and hazards may be neglected in favour of component-related causes and effects.

A.9 Heuristic Evaluations

A heuristic evaluation is a method for identifying design problems in a user interface. It can be used by Subject Matter Experts (SMEs) to foresee most usability issues. A set of best practices (guidelines, heuristics) is chosen that suits the system under analysis. Evaluators then record their observations while interacting with the user interface and rate how well it aligns with the chosen guidelines/heuristics. For any problem encountered, remedial measures are proposed. In detail:

1. Define a set of tasks or scenarios. Tasks representative for the planned use of the system should be identified for the evaluators to perform using the interface. A task list may be derived from a Hierarchical Task Analysis (HTA, see Appendix A.11).
2. Choose a set of heuristics. Multiple sets of general heuristics exist, such as the 10 usability heuristics developed by Nielsen (2024). As commonly used heuristics often do not consider specific domains as is the case with automation or AI, these may be supplemented by appropriate guidelines. Various Guidelines for AI system exist, such as the ALTAI (Assessment List for Trustworthy AI) requirements checklist (EASA, 2024), the Microsoft AI Design Guidelines (Amershi et al., 2019) or the guidelines for the design of human-autonomy systems (Endsley, 2017) that incorporate situation awareness considerations. EASA objectives may also be analysed for fulfilment by heuristic evaluators (refer to Link to EASA Objectives below for examples). For some product types, more specific heuristic checklists exist, such as the Augmented and Mixed Reality Usability Heuristic Checklist for AR/MR applications (Derby, 2023).
3. Perform the chosen tasks. Evaluators note their observations while interacting with the interface to complete the predefined tasks.
4. Compare heuristics against the interface. Evaluators judge how well the interface aligns with each heuristic chosen for analysis. When a heuristic is not met, problems should be recorded and remedial measures proposed.
5. Discuss and aggregate results. If multiple evaluators are involved, they should compare their observations and collate their findings after all have completed their heuristic evaluations.

Related KPAs: HF, Safety, Efficiency

Applicable readiness Levels: Best used early on when the first mock-up or prototype has been developed and iteratively, after changes have been implemented, i.e. HRL 3-6.

Levels of automation: It can support all levels of automation that include a human operator: LOA-0 to LOA-4.

Link with EASA objectives themes:

- *Development and post-ops AI explainability*: Heuristic evaluations may support the definition of operational data that needs to be recorded for post-ops analysis of interaction between AI-based system and end-users.
- *Operational AI explainability*: The prime purpose of heuristic evaluations is to analyse and improve the interface between AI-based systems and end-users, including AI operational explainability.
- *Human-AI teaming*: Heuristic evaluation can support analysis and improvement of interactions between humans and AI-based systems.
- *Modality of interaction and style of interface*: The prime purpose of heuristic evaluations is to analyse and improve the interface between AI-based systems and end-users, including interaction modes and interface style.
- *Error management*: Heuristic evaluations can support the analysis and design of fault tolerant interfaces and suitable information provision to users in the case of errors.
- *Failure management*: Heuristic evaluations can support the analysis and design of suitable information provision to users in the case of failures.

Uncertainty: Heuristic evaluations are based on subjective judgment with poor reliability and validity. Involving three to five SMEs to perform the heuristic evaluation is recommended to increase the quality of results. While many usability problems can be identified early on, heuristic evaluations are not exhaustive so they may be followed up with usability testing.

Technical complexity: Heuristics are simple to use and require very little training.

Benefits

- Heuristic evaluations are simple, requiring little training.
- Heuristics can be applied early on and iteratively in the design cycle to functional prototypes as well as static mock-ups or paper drawings.
- Many usability problems can be identified and addressed early in the design process.
- It is a cost-effective approach to interface assessment requiring few resources.
- Heuristic evaluations are quick to be performed and analysed.
- Findings of heuristic evaluations offer immediate and useful feedback on the interface's problems and potential solutions.

Limitations

- Heuristic evaluations have poor reliability and validity.
- Findings are not exhaustive.
- Evaluations are subjective.
- Results depend on the skills and knowledge of evaluators, so SMEs are needed.

A.10 HITL (Human-In-The-Loop) Simulations & Wizard of Oz

Human-in-the-Loop (HITL) studies involve a high-fidelity interactive simulation in which participants interact with the system as they would in real life and thereby influence the outcomes of events in the simulation. Both common as well as rare, but critical real-world scenarios, may be replicated in human-in-the-loop simulations. This way, scenarios can be tested in a controlled setting to identify issues that would otherwise only become apparent after the new system is deployed and the scenario arises in the real world. HITL simulations may involve a mature AI-based system that is tested in conjunction with the human participants or a prototype controlled by a human operator.

A common method employed in Human-in-the-Loop studies including AI features is the Wizard of Oz technique (Cooke, 2020). In Wizard of Oz studies, participants interact with a prototype that is presented as being an autonomous system, but is actually controlled by a hidden human operator (the "wizard"). This allows for early testing of how humans will interact with the autonomous system before it is fully developed. It also eliminates the risk of inconsistent or unwanted behaviour of the autonomous system. Depending on the goal of the study, the human operator may also deliberately cause unwanted system behaviour to assess how humans react when the autonomous system is not working as intended.

A procedure of a HITL simulation with optional Wizard of Oz technique includes:

1. Define a research question and task. The research question and studied task will guide the procedure of the simulation and metrics assessed. A high-level task may be derived from an HTA.

2. Develop use case scenarios. Scenarios are developed in which the task is to be completed. These may include day-to-day procedures as well as critical events representing conditions the user may encounter in the real world.
3. Create the simulation. The simulation setup should replicate all relevant components of the real-world environment. Additionally, the system to be studied should be prototyped allowing for interaction by the user and control by the wizard. The prototype should be integrated into the simulation set-up.
4. Develop a script of events. It should be noted under what conditions use case scenarios will occur in the simulation runs, e.g. at a specific point in time or triggered by a preceding event.
    a. Develop a script for the wizard. To ensure consistent responses of the wizard to participant's inputs, a detailed script should be written for the wizard outlining every potential user input and associated response by the wizard. This allows for consistent responses across participants matching the intended functionality of the autonomous system.
5. Obtain ethical approval from an ethics committee. Research involving human participants generally requires ethical approval.
    a. Obtain ethical approval for deception. It is important to note that the Wizard of Oz technique uses deception since participants are made to believe that they are interacting with an autonomous system instead of a human operator.
6. Prepare equipment. Next to the simulation setup, any equipment required to record selected metrics may be set up, including recording devices, selected questionnaires, and instructions given to the participants.
    a. Prepare instructions for wizard. The wizard should follow a script that should cover all scenarios of the simulation runs.
7. Pilot test. It is recommended to run a pilot study to test whether the simulation is working as planned and check clarity of instructions given to participants. This may also be done internally without the need to recruit end-users. Afterwards, the simulation and instructions may be refined accordingly. A second pilot study may be required when many changes were made to ensure that all identified issues have been addressed sufficiently.
    a. Pilot test the wizard script. In a pilot study, potential gaps in the wizard's script may be identified and the script improved iteratively.
8. Recruit end-users. Participants should represent potential end-users of the system.
9. Brief participants. The procedure and purpose of the study should be presented to participants including all metrics recorded. They may be given an overview of the new system as a demonstration, either live or as a video in which the functionalities of the system are presented. This can be followed by a training session in which they familiarise themselves with the simulation and the new system. Participants should be given time to ask questions before the simulation runs. Before the simulation runs, they should sign an informed consent form approved by the ethics committee.
    a. Present the "autonomous" system. During the presentation or demonstration of the system, participants are led to believe that the system they are going to interact with is autonomous.
10. Run the simulation. After participants are instructed, the simulation is run. It is recommended to provide no to minimal assistance or feedback while the participants progress through the simulation. Researchers may note observations or participant's comments during the simulation run. Depending on the research questions and their complexity, use cases may be presented one-by-one in separate runs or after one another during the same simulation run.

Some intrusive measures may be chosen that are administered during the run, such as Instantaneous Self-Assessment (ISA), Situation Present Assessment Method (SPAM). If participants complete multiple simulation runs (as in a within-subjects design), they may be given a break after each run to reduce fatigue and collect feedback through questionnaires.

11. Administer selected questionnaires. After one or all simulation runs are completed, participants may be asked to rate their experience on the chosen questionnaires. Depending on the goal of the study, participants may retrospectively rate their workload (e.g. NASA-TLX), situation awareness (e.g. Situation Awareness Rating Technique, SART). They may also be asked to rate the interface's usability (e.g. System usability Scale, SUS), their trust in the system (e.g. the Trust between People and Automation scale, TPA) or acceptance of the system (e.g. System Acceptance Scale, SAS). In conjunction with standardised questionnaires, participants may be asked open questions.
12. Debrief participants. Participants may be asked to reflect back on the simulation runs through semi-structured interview to gather qualitative insights.
    a. Debrief about use of deception. If deception was used as in the Wizard of Oz technique, participants should be debriefed about the elements of deception including the reasons for deception. Participants should be given the opportunity to withdraw their consent or re-consent after complete disclosure of the deception.
13. Analyse data and report findings. Both qualitative and quantitative data can be derived from usability studies. Questionnaire data may hint at potential problems and can be used to compare the interface to a benchmark or similar interfaces. Task completion times, search patterns, and errors or success rates also reveal strengths and weaknesses of the interface. Combining these findings with participant's comments or answers to open questions can help to understand why problems exist and what may need to change.

Related KPAs: Safety, HF, Efficiency

Applicable readiness levels: HRL 5-9

Levels of automation: It can support all levels of automation that include a human operator: LOA-0 to LOA-4.

Link with EASA objectives themes:

- *Safety assessment*: HITL simulations can support a safety assessment.
- *Operational AI explainability*: HITL simulations are a prime means to analyse and improve the interface between AI-based systems and end-users, including operational AI explainability.
- *Human-AI teaming*: HITL simulations can support analysis and improvement of interactions between humans and AI-based systems.
- *Modality of interaction and style of interface*: HITL simulations can support analysis and improvement of the interface between AI-based systems and end-users, including interaction modes and interface style.
- *Error management*: HITL simulations can support the analysis and design of fault tolerant interfaces and suitable information provision to users in the case of errors.
- *Failure management*: HITL simulations can support the analysis and design of suitable information provision to users in the case of failures.
- *Organisation*: HILT simulations can support the development of training processes for interacting with AI-based systems by end-users.

Uncertainty: Since usually a limited number of scenarios can be analysed, for risk assessments the level of uncertainty in the output can be high.

Technical complexity: High

Benefits:

- HITL simulation enables testing of an AI-based system before it is (fully) developed and it can reveal issues before deployment of the tested system.
- It provides in-depth analysis of how end-users will interact with the system.
- Controlling the AI-based system manually in the Wizard-of-Oz mode ensures replicable and desirable actions of the system.
- It provides the opportunity to study human responses in in critical scenarios, including non-nominal conditions of technical systems.

Limitations:

- HITL simulation infrastructure and tuning for a particular advanced automation operational concept are costly.
- Conducting a HITL simulation experiment is costly and time-consuming.
- Only a limited number of scenarios can be evaluated in HITL simulation runs. This means that potentially critical scenarios, types of variability in traffic conditions, and uncertainties in the input of AI-based systems are neglected.

A.11 HTA (Hierarchical Task Analysis)

Hierarchical Task Analysis (HTA) is a top-down task decomposition method that determines how a task is split into subtasks and in which order the subtasks are performed. The list of steps to be conducted in an HTA has minor variations across references; according to (Diaper & Stanton, 2004), the steps are as follows:

1. Decide the purpose(s) of the analysis. For example, to design a new system, to modify an existing system, or to develop operator training.
2. Definition of task goals. Get agreement between stakeholders on the definition of the performance goals of the task and on how one would know whether these goals have been attained. Stakeholders may include designers, managers, supervisors, instructors, and operators.
3. Identify data acquisition means. Identify sources of task information and select means of data acquisition. Available sources may include direct observation, walk-through, protocols, expert interviews, operating procedures and manuals, performance records, accident data, and simulations.
4. Acquire data and draft a decomposition table/diagram. Collect the data of step 3 and use it to develop a task decomposition. This can be done in a diagram or table, or both. At the top of the diagram is the primary task. Below it are subtasks that need to be executed to complete the primary task. For the subtasks the order in which they need to be addressed is indicated. At the next level, the subtasks are decomposed in the same way. The subtasks are appropriately numbered for easy reference.

5. Re-check the validity of the decomposition with stakeholders. Here, stakeholders may be invited to confirm the analysis. It is recommended to revert to step 4 until misinterpretations and omissions have been rectified.
6. Identify significant operations. Here, it is determined where to cut off further decomposition, based on the purpose of the analysis. The main stopping rule is to stop re-describing when further re-description will add no useful information for the analysis, given the scope of the analysis. A frequently used HTA stopping rule is P x C: Stop when the product of probability of unsatisfactory performance (P) times some cost of unsatisfactory performance (C) is judged acceptable (Diaper & Stanton, 2004).
7. Generate and, if possible, test hypotheses concerning task performance. The HTA analysis is used to generate hypotheses concerning the likely sources of actual or potential failure to meet overall task goals and to propose practical solutions. The solutions are to be regarded as hypotheses to be tested.

The required depth of the HTA diagram depends on the depth of analysis and the complexity of the task. Three 'levels' in the HTA diagram is usually the minimum, with seven as a practically-recommended maximum. Although diagrams are more easily assimilated by people, tables are more thorough, because detailed design notes can be added.

Related KPAs: HF, Efficiency

Applicable readiness levels: Task analysis is typically initiated at HRL 3 and may be updated until full maturation of the human system design at HRL 6.

Levels of automation: It can support all levels of automation that include a human operator: LOA-0 to LOA-4.

Link with EASA objectives themes:

- *Characterization*: HTA can support a functional decomposition of the sociotechnical system.
- *Human-AI teaming*: HTA can support analysis of task allocation in human-AI teams.

Uncertainty: HTA is a qualitative approach that does not explicitly account for uncertainties and assumptions.

Technical complexity: Low

Benefits:

- HTA decomposes complex tasks into subtasks and is useful for concurrent operations. The hierarchical structure of HTA enables the analyst to focus on crucial aspects of the task within the context of the overall task.
- HTA offers two distinct training benefits to people engaged in the analysis. First, analysts can use the technique rapidly to gain insight into processes and procedures in an organisation. Second, it has training benefits for people collaborating with the analyst, since they are required to express how they think tasks should be carried out, thereby articulating their understanding of systems.
- Separating a task into subtasks allows the design of supporting systems to offer new ways of performing parts of the task.

- HTA is helpful in the redesign of an existing product or process where tasks should follow a logical sequence.
- The HTA is commonly used and widely accepted in cognitive task analysis.
- It is applicable to human-computer interaction design and has been adopted by designers in software design.

Limitations:

- HTA focuses on processes, meaning that it may not pick up problems with the look, layout, or content of the interface.
- It does not account for system dynamics. HTA does not give a good sense of the length of time of various activities. As a result, inefficiencies due to "waiting" may be missed.
- It is difficult to represent goals which apply to every activity, interrupted activities or 'ad hoc' activities.
- The HTA applies only to procedural activities and not to heavily parallel activities.
- Real tasks may be very complex. HTA does not scale very well; the notation soon becomes unwieldy, making it difficult to follow.

Other task analysis methods:

It is noted that numerous Task Analysis methods are in use, with HTA one of the best known. Other well-known examples are[3]:

- ACT-R (Adaptive Control of Thought - Rational) aims to define the basic and irreducible cognitive and perceptual operations that enable the human mind. In theory, each task that humans can perform should consist of a series of these discrete operations.
- CFA (Cognitive Function Analysis) enables a design team to understand better the right balance between cognitive functions that need to be allocated to human(s) and cognitive functions that can be transferred to machine(s).
- CTA (Cognitive Task Analysis) is used to design human-system interaction and displays, assess job requirements, develop training, or evaluate teamwork. The framework consists of: (a) an analysis of the task that has to be carried out to accomplish particular goals; (b) an analysis of the knowledge and skills required to accomplish these tasks; and (c) an analysis of the cognitive (thought) processes of experienced and less experienced persons.
- GDTA (Goal-Directed Task Analysis) is a cognitive task analysis technique that focuses on the basic goals for each team role (which may change dynamically), the major decisions that should be made to accomplish these goals, and the situation awareness requirements for each decision.
- OFM (Operator Function Model) describes task-analytic structure of operator behaviour in complex systems, focusing on the interaction between an operator and automation in a highly

[3] But also: AET (Ergonomic Job Analysis), CAMEO/TAT (Cognitive Action Modelling of Erring Operator/Task Analysis Tool), Critical Path Method, Critical Task Analysis, Decision Tables, FPC (Flow Process Chart), HECA (Human Error Criticality Analysis), OSD (Operational Sequence Diagram), Operator Task Analysis, PERT (Program Evaluation Review Technique), TAFEI (Task Analysis For Error Identification), TALENT (Task Analysis Linked EvaluatioN Technique), Talk-Through Task Analysis, Team CTA, TTA (Tabular Task Analysis), Walk-Through Task Analysis. See e.g. (Everdij & Blom, 2020) for references.

proceduralised environment. Using graphical notation, OFM attempts to graph the high level goals into simpler behaviours to allow the supervision of the automation.

A.12 NSV-4 diagram (System Functionality and Flow model)

The NSV (NATO Systems Viewpoint)[4] describes systems and interconnections supporting NATO processes. The NSV associates systems resources to the NOV (NATO Operational Viewpoint) and/or the NSOV (NATO Service-Oriented Viewpoint). These systems are the resources that are used to construct the services of the NSOV. They support the operational activities and facilitate the exchange of information among operational nodes as defined in the NOV. The views are numbered NSV-1 through NSV-12, with NSV-1 being the System Interface Description, NSV-2 being the System Communications Description, see e.g. NoMagic (2025) for the full list.

NSV-4 is the System Functionality Description[5], which addresses human and system functionality. The primary purposes of NSV-4 are to develop a clear description of the necessary data flows that are inputs and outputs for each resource, to ensure that the functional connectivity is complete, and to ensure that the functional decomposition reaches an appropriate level of detail. The description provides detailed information regarding the allocation of functions to resources and flow of data between functions. NSV-4 can be represented using:

- NSV-4 diagram for Function hierarchies. This diagram is based on the UML Class diagram.
- NSV-4 diagram for Function flows. This diagram is based on the UML Activity diagram.
- UML Class diagram.
- UML Activity diagram.
- SysML Block diagram.
- SysML Activity diagram.

The functions are often related to Operational Activities captured in an NOV-5. Here NOV (NATO Operational Viewpoint) is a description of the tasks and activities, operational elements, and information exchanges required to accomplish NATO missions. The NOV describes how to identify the nodes, their assigned tasks and activities, and dependencies between nodes. It defines the types of information exchanged, which tasks and activities are supported by the information exchanges, and the operational details of information exchanges. The operational viewpoints are numbered NOV-1 through NOV-7, with NOV-5 being the Operational Activity Model.

The NSV-4 view consists of two diagrams, which are created in this order:

1. Functionality Description diagram. This diagram represents Functionality Description hierarchies and is created in 5 steps: 1) Create Functions. 2) Create or reuse (recommended) Resources from NSV-1, NSV-2. 3) Draw Is Capable To Perform relationship between the Resources and Functions. 4) Draw Compositions (whole-part relationships) between the Functions if necessary. 5) Draw an Implements relationship between the Functions and Operational Activities from NOV-5.
2. Functionality Description Flow diagram. This diagram represents Functionality Description flows and is created in 3 steps: 1) Either create Function Actions or drag the Functions from

[4] https://docs.nomagic.com/display/UAF12P2024x/NATO+Systems+Viewpoint
[5] https://docs.nomagic.com/display/UAF12P2024x/NSV-4+System+Functionality+Description

the Containment tree directly to the diagram. 2) Connect the Function Actions using the Function Edges. 3) Display the possible Resource Exchange s on every Function Edge.

Use of NSV-4 within SESAR

In SESAR guidance material (2018), the NSV-4 diagram is used for deriving design characteristics of the ATM/ANS functional system to ensure that the system operates as specified (the 'success approach'). These design characteristics are the functional, performance and interfacing properties of the ATM/ANS functional system design elements that are affected by the Change and which have a safety implication. There is focus on those design elements that are modified or new. First, the user identifies which ATM/ANS functional system design elements would concur to the fulfilment of a given Safety Objective. Next, Safety Requirements are derived for the previously identified design elements, aiming at specifying the new or modified functionality or performance. As a general rule, a design element might be allocated to one or multiple Safety Requirements and a Safety Requirement should not address more than one design element. Since the design will be further evolving, the corresponding Safety Requirements will be further refined as well.

The NSV-4 can be created starting from the NOV-5 diagram which describes the scenarios/use cases (Operational processes) defined for the Solution. The NSV-4 diagram supports the description of how the resources (human and technical) are contributing to the use case/operational process. The Activities of the operational process are achieved by Functions, which are provided by a Capability Configuration, i.e. a combination of Human Roles and technical resources. The latter are represented as either Technical Systems or Functional Blocks.

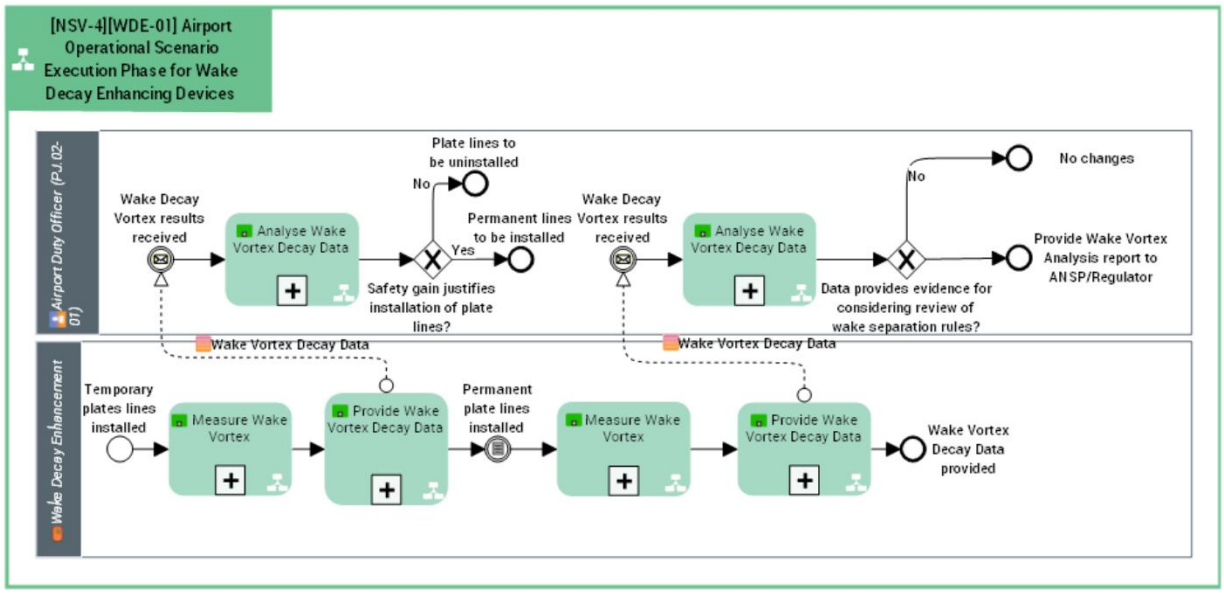


Figure 17. SESAR NSV-4 diagram example, Airport Operational Scenario Execution Phase for Wake Decay Enhancing Devices. Source: https://www.sesarju.eu/sites/default/files/documents/solution/PJ02-01-01%20TS%20IRS.pdf

Related KPAs: Safety.

Applicable readiness levels: According to SESAR guidance material, the NSV-4 diagram is most useful in the Safety Requirements derivation in both V2 and V3 phases, which translates to TRL2-6.

Levels of automation: It can support all levels of automation LOA-0 to LOA-5.

Link with EASA objectives themes:

- *Characterization*: NSV-4 diagrams can support functional decomposition of the sociotechnical system.
- *Safety assessment*: NSV-4 diagrams can support safety assessment.
- *Human-AI teaming*: NSV-4 diagrams can support analysis of task allocation in human-AI teams.
- *AI safety risk mitigation*: NSV-4 diagrams can be used in support of deriving safety requirements.

Uncertainty: The models are qualitative; uncertainty is not really accounted for.

Technical complexity: Medium

Benefits:

- The models are used within SESAR Reference Material.

Limitations:

- The modelling approach is not easily consistently applied. Models made for the same subject but by different experts are often very different.

A.13 Responsibility & Liability Analysis

Responsibility generally means that persons or teams in charge of tasks accept the consequences of their actions/decisions to undertake the tasks, whether the result to be eventually right or wrong. Liability is the state of being legally responsible for something. Accountability means that a person or institution being responsible for a set of duties is required to give account of their fulfilment, i.e. explain their aims, motivations and reasons. Authority is the ability to make decisions and take actions without the need for approval from another member involved in the operations. In (AI-based) advanced automation concepts there can be shifts in the level of authority. In particular, as explained in Section 3.1.2, human operators can have full, partial, limited, or no authority, depending on the level of automation. Such shifts in authority imply shifts in responsibility from human operators to the developers, producers and maintainers of the (AI-based) systems that attain shared or (almost) full authority in operations. Furthermore it can be argued that even at levels of automation where the full authority lies with the human operator (such as LOA-1,2), there may be some level of responsibility with those who developed/produced/maintained the (AI-based) system that support human operators in achieving decisions and implementing actions. The interrelations between stakeholders imply that it may be difficult to identify with precision where responsibility lies in decisions in complex situations, like incidents and accidents. Analysis of responsibility and liability as part of safety management can provide clarity and support the adoption of advanced automation.

The purpose of a responsibility & liability analysis is to identify scenarios and to determine associated risks for issues regarding responsibility and liability of stakeholders in operational concepts that include (AI-based) advanced automation. A responsibility & liability analysis includes the following steps.

1. *Scope & objectives.* The scope of the study concerns the boundaries of the operations, the equipment and the stakeholders. The equipment includes the new AI-based systems as well as other relevant systems in the sociotechnical system. The stakeholders may include end users, end user organisations, developers and producers, trainers, maintainers, and authorities and regulatory bodies. The scope also includes the identification of a legal framework for the liability analysis. The objectives concern a description of the types of results than need to be achieved. This can, for instance, be the identification of conditions where stakeholders may be responsible/liable, it can be assessment of changes in responsibility/liability, or it can be risks (outcomes and likelihoods) concerning responsibility/liability.
2. *Describe sociotechnical system and accountability.* In this step the sociotechnical system, including the AI-based system and advanced automation, is described. It involves the operational context, environmental conditions, the functioning and interface of the AI-based system, the functioning and interaction of other technical systems, the roles, tasks and responsibilities of human operators and their interaction with all relevant technical systems (including the AI-based systems). It also describes the accountability structure of the organisation in line with the scope of the study, i.e. who is accountable for what to whom. This may for instance concern operators, management, system developers, training, and maintenance.
3. *Identify critical scenarios.* The purpose of this step it to identify scenarios that can have a negative impact on key performance areas of an organisation in the scope of the study, e.g. safety, security, environmental impact. Such scenarios may for instance be based on a safety or security assessment that accounts for hazards, interacting stakeholders, and contextual conditions.
4. *Assess responsibility and liability of stakeholders in critical scenarios.* In this step a qualitative assessment is made of the responsibility and liability of stakeholders in the identified critical scenarios. This can be based upon a risk assessment for other KPAs (like safety, security), where severities of operational outcomes and likelihoods of attaining such severity levels have been assessed. For each of these cases the responsibility of directly involved operators as well as of stakeholders is assessed. Based on the assessed responsibilities and the legal regime, the potential liability of each stakeholder is assessed. As defined in the scope, the liability may be assessed in a risk-based way, meaning possible punishment measures and likelihoods of these severity levels.
5. *Conclusions and feedback*. The results of the responsibility and liability assessment provides a structured oversight over responsibilities and possible liability risks for stakeholders in advancedly automated operations. It may be concluded that particular responsibilities are not well defined and that additional operational procedures or organisational changes are advised. It may be concluded that particular liability risks are too high and that changes to the use of the advanced automation or the organisational embedding are advised.

Related KPAs: Liability, responsibility, accountability

Applicable readiness levels: It can be applied if the system design and operational concept have used a sufficient level of maturity, like HRL 4/TRL 4 and higher. At lower readiness levels feedback to the

development can be more easily managed. Responsibility & liability analysis can also support safety management at HRL 9/TRL 9.

Levels of automation: It can be applied at all levels of automation: LOA-0 to LOA-5.

Link with EASA objectives themes:

- *Characterization:* One of the elements of Responsibility & Liability Analysis is identification of the end users, their tasks and responsibilities.
- *Safety assessment*: It can be used to assess the responsibility and liability of stakeholders in safety-critical scenarios.
- *Information security:* It can be used to assess the responsibility and liability of stakeholders in critical scenarios including information security scenarios.
- *Human-AI teaming*: It can be used to assess responsibility of agents in human-AI teams.

Uncertainty: Uncertainty may be handled like in other risk assessment approaches, i.e. by providing ranges of possible severity and likelihood classes.

Technical complexity: The technical complexity of the approach is limited, as it is based on structured qualitative reasoning. Nevertheless, as the complexity of the argumentation increases with the scope of the operational concept and the set of stakeholders, the resulting complexity of a study may be considerable.

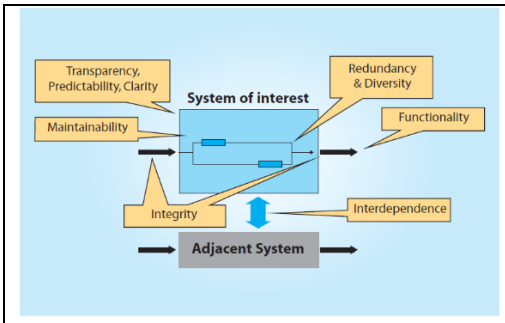
Benefits: Responsibilities and potential liability issues have not yet been systematically analysed for the introduction of advanced automation in ATM and air transport. It is needed to support acceptance of advanced automation and just culture. A limited initial application was reported in (HUCAN D4.2, 2024).

Limitations: The approach has not yet been applied in detail and may need to be finetuned.

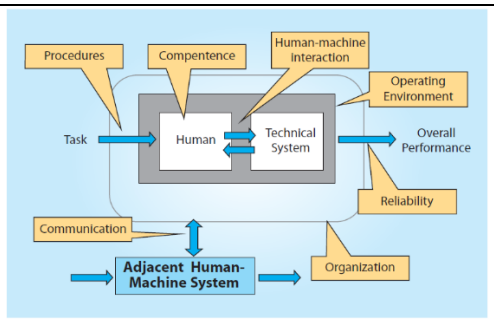
A.14 Safety Scanning and Security Scanning

The purpose of Safety Scanning and its Safety Scanning Tool (SST) is to scan an air transport operational concept regarding all aspects important for safety. The SST is a self-assessment tool, which shows stakeholders the loose ends that require further attention from safe concept development, safety oversight, legislation, regulation, safety management, operational safety, and technology. Oversight officials can use the tool to identify if the safety argumentation offered by the service provider is complete and/or mature enough to accept a notified change. (SCAN TF, 2010, 2010a, 2010b).

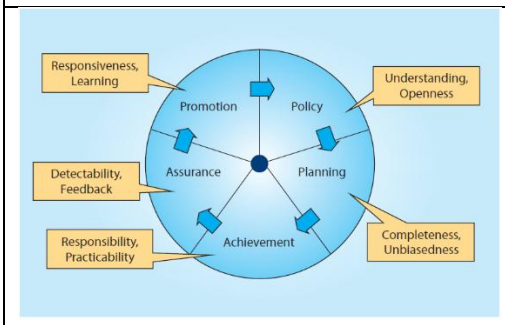
The SST is built on a set of twenty-two ‘Safety Fundamentals’, which are basic design criteria for safe operations. The Safety Fundamentals are organised into four groups (Regulation framework, Safety management, Operational safety, Safety architecture). See Figure 18 for the overview.



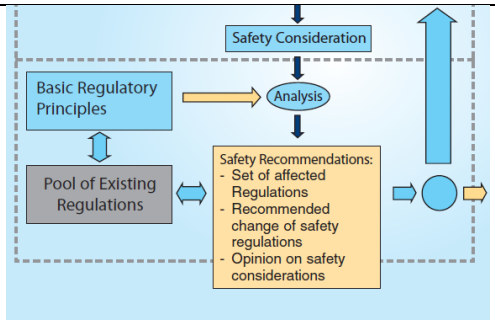
Safety Architecture perspective



Operational Safety perspective



Safety Management perspective



Safety Regulation perspective

Figure 18. Four perspectives of Safety Scanning. The Safety Fundamentals are in yellow.

The tool guides the user systematically through these Safety Fundamentals, by asking for each fundamental between one and five related questions. The answers to be given are multiple-choice (Yes, Partially, No), and a written justification of each choice is required. When all questions have been answered, the user receives a qualitative overview of the Safety Fundamentals that require further attention, as well as an automatically generated report of all answers and justifications provided.

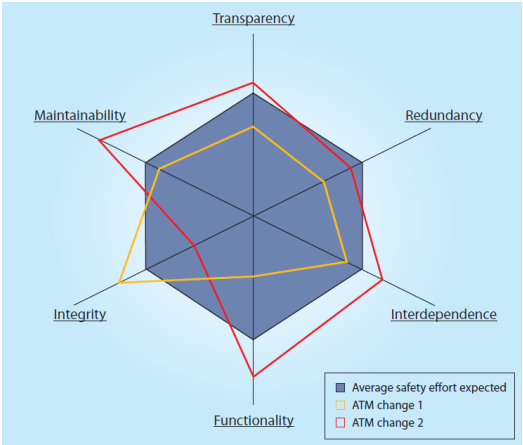


Figure 19. Example output of safety scan (Safety Architecture perspective) of two ATM changes.

Security Scanning is a method based on the same principles; it is supported by a Security Scanning Tool (SecST). The Security fundamentals are provided in Figure 20.

Security Regulation: • Regulatory principles • Structural needs – Legal mandate • Structure needs – Ability • Implementation needs • Need for new regulations	Security Management: • Security Policy • Security Planning • Security Achievement • Security Assurance • Security Awareness	Operational Security: • Procedures • Operating Environment • Competence • Human-machine Interaction • Organisation • Communication • Resilience	Security Architecture: • Transparency • Redundancy • Interdependence • Functionality • Confidentiality, Integrity, Availability • Maintainability

Figure 20. Security Fundamentals used for Security Scanning Tool (SecST).

Related KPAs: Safety, Security

Applicable readiness levels: TRL1-6 and HRL1-6. The method can be used in all lifecycle stages of a proposed change, but is most effective during earlier stages.

Levels of automation: It can support all levels of automation LOA-0 to LOA-5.

Link with EASA objectives themes:

- *Characterisation:* Safety scanning and Security scanning show stakeholders the loose ends that require further attention from safe concept development, safety oversight, legislation, regulation, safety management, operational safety, and technology.
- *Safety assessment*: Safety scanning can support safety assessment.
- *Information security*: Security scanning can support security assessment.
- *Ethics-based assessment:* Safety Scanning and Security Scanning address confidentiality and integrity considerations such as unauthorised disclosure of or access to data.
- *Organisation:* Safety Scanning and Security Scanning show stakeholders the loose ends that require further attention from safe concept development, safety oversight, legislation, regulation, safety management, operational safety, and technology.

Uncertainty: Uncertainty is not really accounted for.

Technical complexity: Low

Benefits:

- In initial phases of the life-cycle of an operational concept, before safety assessments are usually executed, the SST can be used to coordinate and communicate awareness and understanding of safety needs between stakeholders, and to give a level of confidence of how safety is being addressed in a project. The later stages of development therefore aim to be better informed about safety issues and safety assessments are more likely to meet the required safety targets.
- In later phases, it supports oversight officials in developing acceptance criteria for safety evidence.

- The method promotes early consideration of the potential impacts of safety regulation applicable to ATM and its interdependencies with the total aviation activity. This should provide a reduction in project risk regarding safety, and the preparation of a sound safety plan.
- It enables an inclusion of safety into cost benefit analyses (CBA).

Limitations:

- Safety scanning does not provide a quantitative safety assessment.

A.15 SecRAM (Security Risk Assessment Methodology)

In (SESAR JU, 2024b) a cybersecurity risk assessment methodology is described for innovation in ATM. Cybersecurity risk assessment is a process to identify and mitigate the consequences of a cyberattack. It defines a set of security requirements to ensure that if an attack takes place the consequences have been estimated and can be managed and may contribute to the recovery of normal operations in a reasonable time. The steps in the SESAR security risk assessment are (see also Figure 21):

- Define the scope of the risk assessment (description of involved roles, equipment, systems...) and the identification of dependencies on other systems and infrastructure. To perform this step, specialist operational or design knowledge of the system is required.
- Identify assets and valuate possible impacts on assets: assets form the targets of security attacks, and the identification of possible impacts is concerned with evaluating the harm resulting from each asset being compromised by an attack.
- Identify vulnerabilities, threats and likely threat combinations: it comprises the identification of possible (or credible) threat sources and related threat scenarios. Each threat is associated to vulnerabilities of the system that can be exploited by an attacker. This group of activities aims at providing an insight into all routes through the system (threat scenarios) that a threat may use to access an asset.
- Identify a set of security controls that act upon the supporting assets, that will reduce the impact on primary assets, and evaluate the impact on primary assets after implementation of the security controls. Note: a first iteration of the risk evaluation may be conducted with controls limited to those already in operations (e.g. environment assumptions) and generic organisational controls (e.g., from a Minimum Set of Security Controls set by SESAR) to focus only on the identification of controls mitigating risks that do not meet the programme generic security objective.
- Determine the likelihood of the impact on primary assets to occur.
- Assess the security risk.
- Determine whether the security risk is within the acceptable level set by the cyber-security objectives – if not, it is necessary to go back in the process to identify how the situation can be improved.

The need for prioritisation should be reevaluated at each maturity gate.

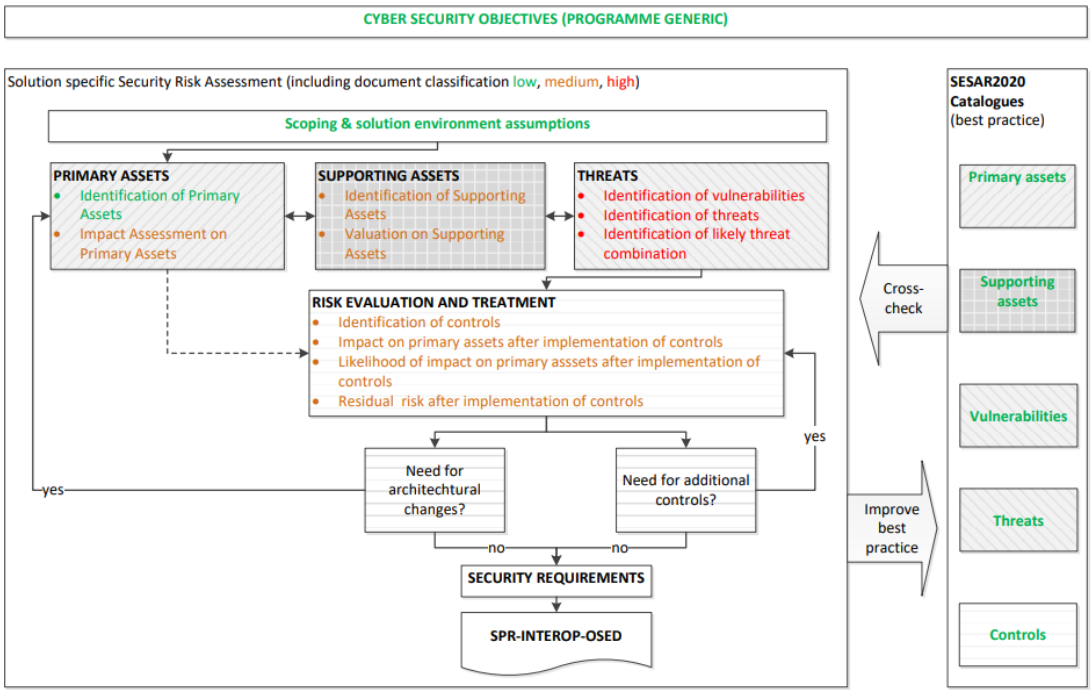


Figure 21. The SecRAM methodology. Source: (SESAR JU, 2024b).

Related KPAs: Security (Cyber)

Applicable readiness levels: TRL2-8.

Levels of automation: It can support all levels of automation LOA-0 to LOA-5.

Link with EASA objectives themes:

- *Information security*: SecRAM can support security assessment.
- *Organisations*: SecRAM can support organisations in continuous assessment of information security risks.

Uncertainty: The likelihoods are provided qualitatively, but ranges of uncertainty can be given.

Technical complexity: Medium

Benefits:

- SecRAM is a general and complete framework for the assessment of cybersecurity risk in ATM.

Limitations:

- The methodology does not seem to deeply analyse vulnerabilities.
- Additional effort is required to address a continuous assessment as the threat landscape evolves.

A.16 Usability Testing

After refining the prototype based on usability issues identified through heuristic evaluations, usability tests may be performed with end-users. They may be asked to think aloud while performing a set of tasks on the interface and provide feedback on clarity, functionalities, and layout. Usability tests can also be supplemented with standardised usability questionnaires to compare the new systems' usability against the current system or a benchmark. A common procedure is:

1. Define tasks and metrics under analysis. It should be specified which aspects of the interface and user interaction should be assessed and which tasks the users will perform using the interface. Metrics assessed may include subjective usability, task performance, workload, situation awareness, user reactions or error rates. These may be assessed using objective measures such as task completion times or subjective measures such as post-run questionnaires. Standardised questionnaires may be supplemented with additional questions, e.g. to record participants' familiarity with similar already existing systems. A set of key tasks representative of the interface's functions are chosen for analysis. It may be helpful for the analyst to note the sequence of task steps including the relevant interface components of each step to later compare the intended procedure with user behaviour. These may be derived from a HTA conducted previously.
2. Formulate instructions for users. Instructions should include a description of the studied system and the purpose of the study as well as a list of tasks the users should perform using the interface. These may be given to the users as they progress through the interface in written or oral format. Questions asked throughout the study may also be noted, e.g. to ask for feedback after a task has been completed.
3. Obtain Ethical Approval from an Ethics Committee. Research involving human participants generally requires ethical approval.
4. Pilot test. It is recommended to run a pilot study to test the clarity of instructions and questions as well as the timeframe of the planned study. This can also be done internally without the need to recruit end-users. Afterwards, test plan and instructions may be refined accordingly.
5. Recruit end-users. Participants of the usability study should represent potential end-users of the system.
6. Brief participants. Participants may be given an overview of the interface and purpose of the study. Depending on the goal of the study, this may include a demonstration, either live or as a video or minimal visual presentation of the interface prior to the study. Before the start of the study, participants should sign an informed consent form approved by the ethics committee.
7. Run usability study. Instructions are presented to the participants to begin the first task. It is recommended to provide no to minimal assistance or feedback while the participants progress through the tasks. Video and audio may be recorded for further analysis and analysts may note observations or participant's comments during the run. Participants may also be instructed to think-aloud, voicing anything that comes to mind while interacting with the interface. Questions may be asked after task completion and before progressing to the next task.
8. Administer selected questionnaires. After all tasks are completed, participants may be asked to rate their experience on the chosen questionnaires. Depending on the goal of the study, participants may retrospectively rate their workload (e.g. NASA-TLX), situation awareness (e.g. SART), and the interface's usability (e.g. SUS). Open questions may be added, e.g. to verify the fulfilment of EASA objectives.

9. Analyse data and report findings. Both qualitative and quantitative data can be derived from usability studies. Questionnaire data may hint at potential problems and can be used to compare the interface to a benchmark or similar interfaces. Task completion times, search patterns, and errors or success rates also reveal strengths and weaknesses of the interface. Combining these findings with participant's comments or answers to open questions can help to understand why problems exist and what may need to change.
10. Propose design recommendations. Based on the findings of the usability study, recommendations for design changes can be developed to mitigate the problems identified. Then, the proposed changes should be implemented to improve the system.

Related KPAs: HF, Safety, Efficiency

Applicable readiness levels: Best used early on when a prototype has been developed and iteratively, after changes have been implemented, i.e. HRL 3-6. Usability studies may have different purposes depending on the readiness level, e.g. layout may be assessed as early as possible whereas workload assessment may be more helpful once the prototype has reached a higher fidelity.

Levels of automation: It can support all levels of automation that include a human operator: LOA-0 to LOA-4.

Link with EASA objectives themes:

- *Characterisation:* It supports the identification of the end users, and identifies the tasks the users will perform using the user interface.
- *Development and post-ops AI explainability*: Usability testing may support the definition of operational data that needs to be recorded for post-ops analysis of interaction between AI-based system and end-users.
- *Operational AI explainability*: A main purpose of usability testing is to analyse and improve the interface between AI-based systems and end-users, including AI operational explainability.
- *Human-AI teaming*: Usability testing can support analysis and improvement of interactions between humans and AI-based systems.
- *Modality of interaction and style of interface*: A main purpose of usability testing is to analyse and improve the interface between AI-based systems and end-users, including interaction modes and interface style.
- *Error management*: Usability testing can support the analysis and design of fault tolerant interfaces and suitable information provision to users in the case of errors.
- *Failure management*: Usability testing can support the analysis and design of suitable information provision to users in the case of failures.

Uncertainty: Usability tests greatly rely on self-reported feedback from participants who may be biased towards current systems. Analysts may also interpret participant's responses according to their own expectations or biases. The validity and reliability of findings varies depending on the design of the study. Ideally, the participant sample should represent the target population of end-users and be large enough to derive statistically meaningful results.

Technical complexity: Conducting usability tests requires working knowledge of the associated techniques. Without prior knowledge, extensive training is required, whereas for an experienced

analyst, technical complexity is rather low. Including the initial task definition and analysis of all recorded data, usability studies are also time-consuming.

Benefits:

- Usability tests can be run using prototypes before the interface is fully developed to identify usability issues early on.
- Multiple metrics can be assessed for various purposes including performance, error rates, trust, acceptance, mental workload, and situation awareness.
- Usability tests can offer detailed insights through rich amounts of qualitative and quantitative data.
- End-users are involved in the design process, which increases the fit of the interface to the target population and future acceptance of the final system.
- Usability tests reveal how users will interact with the system, which may differ from intended interactions.
- Simple to conduct with appropriate personnel.

Limitations:

- Usability tests including associated methods are time-consuming to conduct.
- Data analysis, especially for qualitative data, may be lengthy.
- Acquiring enough participants for reliable results may be difficult, depending on the target population.

Appendix B Objectives EASA AI guidelines

Table 7 (also provided in Appendix A of HUCAN D4.2 (2024)) lists all Objectives from Section C (AI Trustworthiness guidelines) in the EASA AI guidelines, in their Concept Paper with guidance for level 1&2 ML applications (EASA, 2024). These Objectives are organised into:

- C.2: AI trustworthiness analysis
- C.3: AI assurance
- C.4: Human factors for AI
- C.5: AI safety risk mitigation

Also added (C.6) are provisions that are anticipated to apply to the organisations developing or deploying AI-based systems.

Table 7 provides the Levels of Automation (LoA) for which the objectives are applicable, and indicates whether EASA anticipates Means of Compliance (MOC). Here,

- Objectives in White are relevant for Levels of Automation 1A-2B,
- Objectives in Green are relevant for 1B-2B,
- Objectives in Yellow are relevant for 2A-2B,
- Objectives in Blue are relevant for 2B only.

Levels of Automation 3A and 3B (i.e. advanced automation) are out of scope of the EASA AI guidelines. Also note that the EASA AI guidelines cover supervised learning or unsupervised learning, but not other types of learning such as reinforcement learning, and it covers offline learning processes where the model is 'frozen' at the time of approval, but not online learning processes.

C2. Trustworthiness analysis		
LoA	**EASA Objectives**	**Anticipated MOC**
C2.1(CO/CL). Characterisation and classification of the AI application		
1A - 2B	Obj.CO-01: The applicant should identify the list of end users that are intended to interact with the AI-based system, together with their roles, their responsibilities (including indication of the level of teaming with the AI-based system, i.e. none, cooperation, collaboration) and expected expertise (including assumptions made on the level of training, qualification and skills).	-
1A - 2B	Obj.CO-02: For each end user, the applicant should identify which goals and associated high-level tasks are intended to be performed in interaction with the AI-based system.	Ant.MOC CO-02
1A - 2B	Obj.CO-03: The applicant should determine the AI-based system taking into account domain-specific definitions of 'system'.	Ant.MOC CO-03
1A - 2B	Obj.CO-04: The applicant should define and document the ConOps for the AI-based system, including the task allocation	Ant.MOC CO-04* *Dependencies:

	pattern between the end user(s) and the AI-based system. A focus should be put on the definition of the OD and on the capture of specific operational limitations and assumptions.	Obj.CO-01 Obj.CO-02
1A - 2B	Obj.CO-05: The applicant should document how end users' inputs are collected and accounted for in the development of the AI-based system.	Ant.MOC CO-05
1A - 2B	Obj.CO-06: The applicant should perform a functional analysis of the system, as well as a functional decomposition and allocation down to the lowest level.	Ant.MOC CO-06
1A - 2B	Obj.CL-01: The applicant should classify the AI-based system, based on the levels presented [by EASA], with adequate justifications.	Ant.MOC CL-01-1* *Dependencies: Obj.CO-02 Ant.MOC CL-01-2
C2.2(SA). Safety assessment of ML Applications		
1A - 2B	Obj.SA-01: The applicant should perform a safety (support) assessment for all AI-based (sub)systems, identifying and addressing specificities introduced by AI/ML usage.	Ant.MOC-SA-01-1 Ant.MOC-SA-01-2 Ant.MOC-SA-01-3 Ant.MOC-SA-01-4 Ant.MOC-SA-01-5* *Dependencies: Objs.LA Ant.MOC-SA-01-6 Ant.MOC-SA-01-7 Ant.MOC-SA-01-7 Ant.MOC-SA-01-8 Ant.MOC-SA-01-9
1A - 2B	Obj.SA-02: The applicant should identify which data needs to be recorded for the purpose of supporting the continuous safety assessment.	Ant.MOC SA-02 *Dependencies: Ant.MOC EXP-04-2
1A - 2B	Obj.SA-03: In preparation of the continuous safety assessment, the applicant should define metrics, target values, thresholds and evaluation periods to guarantee that design assumptions hold.	Ant.MOC SA-03
C2.3(IS). Information security risks management		
1A - 2B	Obj.IS-01: For each AI-based (sub)system and its data sets, the applicant should identify those information security risks with an impact on safety, identifying and addressing specific threats introduced by AI/ML usage.	Ant.MOC IS-01

1A - 2B	Obj.IS-02: The applicant should document a mitigation approach to address the identified AI/ML-specific information security risk.	Ant.MOC IS-02
1A - 2B	Obj. IS-03: The applicant should validate and verify the effectiveness of the security controls introduced to mitigate the identified AI/ML-specific information security risks to an acceptable level.	Ant.MOC IS-03
C2.4(ET). Ethics-based assessment		
2A – 2B	Obj.ET-01: The applicant should perform an ethics-based trustworthiness assessment for any AI-based system developed using ML techniques or incorporating ML models.	-
2A – 2B	Obj.ET-02: The applicant should ensure that the AI-based system bears no risk of creating overreliance, attachment, stimulating addictive behaviour, or manipulating the end user's behaviour.	Ant.MOC ET-02 *Dependencies: Obj.ET-01 Obj.IMP-09
2A – 2B	Obj.ET-03: The applicant should comply with national and EU data protection regulations (e.g. GDPR), i.e. involve their Data Protection Officer, consult with their National Data Protection Authority, etc.	Ant.MOC ET-03
2A – 2B	Obj.ET-04: The applicant should ensure that the creation or reinforcement of unfair bias in the AI-based system, regarding both the data sets and the trained models, is avoided, as far as such unfair bias could have a negative impact on performance and safety.	Ant.MOC ET-04
2A – 2B	Obj.ET-05: The applicant should ensure that end users are made aware of the fact that they interact with an AI-based system, and, if applicable, whether some personal data is recorded by the system.	Ant.MOC ET-05
2A – 2B	Obj.ET-06: The applicant should perform an environmental impact analysis, identifying and assessing potential negative impacts of the AI-based system on the environment and human health throughout its life cycle (development, deployment, use, end of life), and define measures to reduce or mitigate these impacts.	Ant.MOC ET-06 *Dependencies: Obj.ET-01
2A – 2B	Obj.ET-07: The applicant should identify the need for new skills for users and end users to interact with and operate the AI-based system, and mitigate possible training gaps	Ant.MOC ET-07 *Dependencies: Obj.ET-01 Prov.ORG-07 Prov.ORG-08

2A – 2B	Obj.ET-08: The applicant should perform an assessment of the risk of de-skilling of the users and end users and mitigate the identified risk through a training needs analysis and a consequent training activity	Ant.MOC ET-08 *Dependencies: Obj.ET-01 Prov.ORG-07 Prov.ORG-08

C3. AI Assurance

LoA	EASA Objectives	Anticipated MOC
C3.1(DA). Learning assurance		
1A-2B	Obj.DA-01: The applicant should describe the proposed learning assurance process, taking into account each of the steps described in Sections C.3.1.2 to C.3.1.14, as well as the interface and compatibility with development assurance processes.	Ant.MOC DA-01
1A-2B	Obj.DA-02: Based on (sub)system requirements allocated to the AI/ML constituent, the applicant should capture the following minimum for the AI/ML constituent requirements: — safety requirements allocated to the AI/ML constituent (e.g. performance, reliability, resilience); — information security requirements allocated to the AI/ML constituent; — functional requirements allocated to the AI/ML constituent; — operational requirements allocated to the AI/ML constituent, including AI/ML constituent ODD monitoring and performance monitoring (to support related objectives in Section C.3.2.6), detection of OoD input data and data-recording requirements (to support objectives in Section C.3.2.7); — other non-functional requirements allocated to the AI/ML constituent (e.g. scalability); and — interface requirements.	*Dependencies: Obj.CO-04
1A-2B	Obj.DA-03: The applicant should define the set of parameters pertaining to the AI/ML constituent ODD, and trace them to the corresponding parameters pertaining to the OD when applicable.	Ant.MOC DA-03 *Dependencies: Obj.CO-04
1A-2B	Obj.DA-04: The applicant should capture the DQRs for all data required for training, testing, and verification of the AI/ML constituent, including but not limited to: — the data relevance to support the intended use; — the ability to determine the origin of the data; — the requirements related to the annotation process; — the format, accuracy and resolution of the data; — the traceability of the data from their origin to their final operation through the whole pipeline of operations;	Ant.MOC DA-04

	— the mechanisms ensuring that the data will not be corrupted while stored, processed, or transmitted over a communication network; — the completeness and representativeness of the data sets; and — the level of independence between the training, validation and test data sets.	
1A-2B	Obj.DA-05: The applicant should capture the requirements on data to be pre-processed and engineered for the inference model in development and for the operations.	-
1A-2B	Obj.DA-06: The applicant should describe a preliminary AI/ML constituent architecture, to serve as reference for related safety (support) assessment and learning assurance objectives.	-
1A-2B	Obj.DA-07: The applicant should validate each of the requirements captured under Objectives DA-02, DA-03, DA-04, DA-05 and the architecture captured under Objective DA-06.	Ant.MOC DA-07 *Dependencies: Obj.DA-02 Obj.DA-03 Obj.DA-04 Obj.DA-05 Obj.DA-06
1A-2B	Obj.DA-08: The applicant should document evidence that all derived requirements generated through the learning assurance processes have been provided to the (sub)system processes, including the safety (support) assessment.	*Dependencies: Obj. DA-03 Obj. DA-04 Obj. DA-05 Obj. LM-01 Obj. LM-02 Obj. LM-04 Obj. IMP-01
1A-2B	Obj.DA-09: The applicant should document evidence of the validation of the derived requirements, and of the determination of any impact on the safety (support) assessment and (sub)system requirements.	*Dependencies: Obj. DA-03 Obj. DA-04 Obj. DA-05 Obj. LM-01 Obj. LM-02 Obj. LM-04 Obj. IMP-01
1A-2B	Obj.DA-10: Each of the captured AI/ML constituent requirements should be verified.	

C3.1(DM). Data management

1A-2B	Obj.DM-01: The applicant should identify data sources and collect data in accordance with the defined ODD, while ensuring satisfaction of the defined DQRs, in order to drive the selection of the training, validation and test data sets.	
1A-2B	Obj.DM-02-SL: Once data sources are collected and labelled, the applicant should ensure that the annotated or labelled data in the data set satisfies the DQRs captured under Objective DA-04.	*Dependencies: Obj.DA-04
1A-2B	Obj.DM-03: The applicant should define the data preparation operations to properly address the captured requirements (including DQRs).	
1A-2B	Obj.DM-04: The applicant should define and document pre-processing operations on the collected data in preparation of the model training.	Ant.MOC DM-04
1A-2B	Obj.DM-05: When applicable, the applicant should define and document the transformations to the pre-processed data from the specified input space into features which are effective for the performance of the selected learning algorithm.	Ant.MOC DM-05-1 Ant.MOC DM-05-2 Ant.MOC DM-05-3
1A-2B	Obj.DM-06: The applicant should distribute the data into three separate data sets which meet the specified DQRs in terms of independence (as per Objective DA-04): — the training data set and validation data set, used during the model training; — the test data set used during the learning process verification, and the inference model verification.	*Dependencies: Obj.DA-04 Obj.DA-07
1A-2B	Obj.DM-02-UL:Once data sources are collected and the test data set labelled, the applicant should ensure that the annotated or labelled data in this test data set satisfies the DQRs captured under Objective DA-04.	*Dependencies: Obj.DA-04
1A-2B	Obj.DM-07: The applicant should ensure verification of the data, as appropriate, throughout the data management process so that the data management requirements (including the DQRs) are addressed.	Ant.MOC DM-07-1 Ant.MOC DM-07-2 Ant.MOC DM-07-3 Ant.MOC DM-07-4 Ant.MOC DM-07-5
1A-2B	Obj.DM-08: The applicant should perform a data verification step to confirm the appropriateness of the defined ODD and of the data sets used for the training, validation and verification of the ML model.	Ant.MOC DM-08 *Dependencies: Obj.EXP-02 Obj.EXP-03

C3.1(LM). Learning process management

1A-2B	Obj.LM-01: The applicant should describe the ML model architecture.	Ant.MOC LM-01
1A-2B	Obj.LM-02: The applicant should capture the requirements pertaining to the learning management and training processes, including but not limited to: — model family and model selection; — learning algorithm(s) selection; — explainability capabilities of the selected model; — activation functions; — cost/loss function selection describing the link to the performance metrics; — model bias and variance metrics and acceptable levels (only in supervised learning); — model robustness and stability metrics and acceptable levels; — training environment (hardware and software) identification; — model parameters initialisation strategy; — hyper-parameters and parameters identification and setting; — expected performance with training, validation and test data sets.	Ant.MOC LM-02
1A-2B	Obj.LM-03: The applicant should document the credit sought from the training environment and qualify the environment accordingly.	
1A-2B	Obj.LM-04: The applicant should provide quantifiable generalisation bounds.	Ant.MOC LM-04
1A-2B	Obj.LM-05: The applicant should document the result of the model training.	Ant.MOC LM-05 *Dependencies: Obj.SA-01
1A-2B	Obj.LM-06: The applicant should document any model optimisation that may affect the model behaviour (e.g. pruning, quantisation) and assess their impact on the model behaviour or performance.	Ant.MOC LM-06
1A-2B	Obj.LM-07-SL: The applicant should account for the bias-variance trade-off in the model family selection and should provide evidence of the reproducibility of the model training process.	Ant.MOC LM-07-SL
1A-2B	Obj.LM-08: The applicant should ensure that the estimated bias and variance of the selected model meet the associated learning process management requirements.	Ant.MOC LM-08 *Dependencies: Obj.DM-02-UL

1A-2B	Obj.LM-09: The applicant should perform an evaluation of the performance of the trained model based on the test data set and document the result of the model verification.	Ant.MOC LM-09 *Dependencies: Obj.SA-01 Obj.LM-04
1A-2B	Obj.LM-10: The applicant should perform requirements-based verification of the trained model behaviour.	Ant.MOC LM-10 *Dependencies: Obj.LM-02 Obj.DA-02
1A-2B	Obj.LM-11: The applicant should provide an analysis on the stability of the learning algorithms.	Ant.MOC LM-11
1A-2B	Obj.LM-12: The applicant should perform and document the verification of the stability of the trained model, covering the whole AI/ML constituent ODD.	Ant.MOC LM-12
1A-2B	Obj.LM-13: The applicant should perform and document the verification of the robustness of the trained model in adverse conditions.	Ant.MOC LM-13
1A-2B	Obj.LM-14: The applicant should verify the anticipated generalisation bounds using the test data set.	Ant.MOC LM-14 *Dependencies: Obj.LM-04
1A-2B	Obj.LM-15: The applicant should capture the description of the resulting ML model.	
1A-2B	Obj.LM-16: The applicant should confirm that the trained model verification activities are complete.	Ant.MOC LM-16
C3.1(IMP). Model implementation		
1A-2B	Obj.IMP-01: The applicant should capture the requirements pertaining to the ML model implementation process.	Ant.MOC IMP-01
1A-2B	Obj.IMP-02: The applicant should validate the model description captured under Objective LM-15 as well as each of the requirements captured under Objective IMP-01.	*Dependencies: Obj.LM-15 Obj.IMP-01
1A-2B	Obj.IMP-03: The applicant should document evidence that all derived requirements generated through the model implementation process have been provided to the (sub)system processes, including the safety (support) assessment.	
1A-2B	Obj.IMP-04: Any post-training model transformation (conversion, optimisation) should be identified and validated for its impact on the model behaviour and performance, and the	Ant.MOC IMP-04-1 Ant.MOC IMP-04-2 *Dependencies: Obj.LM-06

	environment (i.e. software tools and hardware) necessary to perform model transformation should be identified.	Obj.IMP-01
1A-2B	Obj.IMP-05: The applicant should plan and execute appropriate development assurance processes to develop the inference model into software and/or hardware items.	Ant.MOC IMP-05
1A-2B	Obj.IMP-06: The applicant should verify that any transformation (conversion, optimisation, inference model development) performed during the trained model implementation step has not adversely altered the defined model properties.	Ant.MOC IMP-06 *Dependencies: Obj.IMP-01
1A-2B	Obj.IMP-07: The differences between the software and hardware of the platform used for model training and those used for the inference model verification should be identified and assessed for their possible impact on the inference model behaviour and performance.	Ant.MOC IMP-07
1A-2B	Obj.IMP-08: The applicant should perform an evaluation of the performance of the inference model based on the test data set and document the result of the model verification.	Ant.MOC IMP-08 *Dependencies: Obj.SA-01 Obj.LM-09
1A-2B	Obj.IMP-09: The applicant should perform and document the verification of the stability of the inference model.	Ant.MOC IMP-09
1A-2B	Obj.IMP-10: The applicant should perform and document the verification of the robustness of the inference model in adverse conditions.	Ant.MOC IMP-10
1A-2B	Obj.IMP-11: The applicant should perform requirements-based verification of the inference model behaviour when integrated into the AI/ML constituent.	Ant.MOC IMP-11 *Dependencies: Obj.IMP-01 Obj.DA-02 Obj.DM-02-UL
1A-2B	Obj.IMP-12: The applicant should confirm that the AI/ML constituent verification activities are complete.	Ant.MOC IMP-12
C3.1(CM). Configuration management		
1A-2B	Obj.CM-01: The applicant should apply all configuration management principles to the AI/ML constituent life-cycle data, including but not limited to: — identification of configuration items; — versioning; — baselining; — change control;	Ant.MOC CM-01

	— reproducibility; — problem reporting; — archiving and retrieval, and retention period.	
C3.1(QA). Quality and process assurance		
1A-2B	Obj.QA-01: The applicant should ensure that quality/process assurance principles are applied to the development of the AI-based system, with the required independence level.	
C3.1(RU). Reuse of AI/ML models		
1A-2B	Obj.RU-01: The applicant should perform an impact assessment of the reuse of a trained ML model before incorporating the model into an AI/ML constituent. The impact assessment should consider: — alignment and compatibility of the intended behaviours of the ML models; — alignment and compatibility of the ODDs; — compatibility of the performance of the reused ML model with the performance requirements expected for the new application; — availability of adequate technical documentation (e.g. equivalent documentation depending on the required assurance level); — possible licensing or legal restrictions on the reused ML model (more particularly in the case of COTS ML models); and — evaluation of the required development level.	Ant.MOC RU-01 *Dependencies: Obj.DA-01
1A-2B	Obj.RU-02: The applicant should perform a functional analysis of the COTS ML model to confirm its adequacy to the requirements and architecture of the AI/ML constituent.	*Dependencies: Obj.DA-02
1A-2B	Obj.RU-03: The applicant should perform an analysis of the unused functions of the COTS ML model, and prepare the deactivation of these unused functions.	*Dependencies: Obj.DA-03 Obj.DA-04 Obj.DA-05 Obj.DA-10 Obj.DM-01 Obj.DM-05 Obj.DM-06 Obj.DM-07 Obj.LM-01 Obj.LM-02 Obj.LM-03 Obj.LM-08 Obj.LM-09 Obj.LM-10

		Obj.LM-11 Obj.LM-12 Obj.LM-15 Obj.IMP-01 Obj.IMP-05 Obj.IMP-06 Obj.IMP-11 Obj.CM-01 Obj.QA-01 Obj.EXP-03
C3.1(SU). Surrogate modelling		
1A-2B	Obj.SU-01: The applicant should capture the accuracy and fidelity of the reference model in order to support the verification of the accuracy of the surrogate model.	
1A-2B	Obj.SU-02: the applicant should identify, document and mitigate the additional sources of uncertainties linked with the use of a surrogate model.	
C3.2(EXP). Development and post-ops AI explainability		
1A-2B	Obj.EXP-01: The applicant should identify the list of stakeholders, other than end users, that need explainability of the AI-based system at any stage of its life cycle, together with their roles, their responsibilities and their expected expertise (including assumptions made on the level of training, qualification and skills).	*Dependencies: Obj.CO-01
1A-2B	Obj.EXP-02: For each of these stakeholders (or groups of stakeholders), the applicant should characterise the need for explainability to be provided, which is necessary to support the development and learning assurance processes.	Ant.MOC EXP-02
1A-2B	Obj.EXP-03: The applicant should identify and document the methods at AI/ML item and/or output level satisfying the specified AI explainability needs.	
1A-2B	Obj.EXP-04: The applicant should design the AI-based system with the ability to deliver an indication of the level of confidence in the AI/ML constituent output, based on actual measurements or on quantification of the level of uncertainty.	
1A-2B	Obj.EXP-05: The applicant should design the AI-based system with the ability to monitor that its inputs are within the specified ODD boundaries (both in terms of input parameter range and	

	distribution) in which the AI/ML constituent performance is guaranteed.	
1A-2B	Obj.EXP-06: The applicant should design the AI-based system with the ability to monitor that its outputs are within the specified operational AI/ML constituent performance boundaries.	
1A-2B	Obj.EXP-07: The applicant should design the AI-based system with the ability to monitor that the AI/ML constituent outputs (per Objective EXP-04) are within the specified operational level of confidence.	Ant.MOC EXP-07 *Dependencies: Obj.EXP-04
1A-2B	Obj.EXP-08: The applicant should ensure that the output of the specified monitoring per the previous three objectives are in the list of data to be recorded per MOC EXP-09-2.	*Dependencies: Ant.MOC EXP-09-2
1A-2B	Obj.EXP-09: The applicant should provide the means to record operational data that is necessary to explain, post operations, the behaviour of the AI-based system and its interactions with the end user, as well as the means to retrieve this data.	Ant.MOC EXP-09-1 Ant.MOC EXP-09-2 Ant.MOC EXP-09-3 Ant.MOC EXP-09-4 Ant.MOC EXP-09-5
C4. Human factors for AI		
LoA	**EASA Objectives**	**Anticipated MOC**
C4.1(EXP). AI operational explainability		
1B-2B	Obj.EXP-10: For each output of the AI-based system relevant to task(s) (per Objective CO-02), the applicant should characterise the need for explainability.	*Dependencies: Obj.EXP-03 Obj.CO-02
1B-2B	Obj.EXP-11: The applicant should ensure that the AI-based system presents explanations to the end user in a clear and unambiguous form.	Ant.MOC EXP-11
1B-2B	Obj.EXP-12: The applicant should define relevant explainability so that the receiver of the information can use the explanation to assess the appropriateness of the decision/action as expected.	Ant.MOC EXP-12
1B-2B	Obj.EXP-13: The applicant should define the level of abstraction of the explanations, taking into account the characteristics of the task, the situation, the level of expertise of the end user and the general trust given to the system.	Ant.MOC EXP-13

1B-2B	Obj.EXP-14: Where a customisation capability is available, the end user should be able to customise the level of abstraction as part of the operational explainability.	Ant.MOC EXP-14
1B-2B	Obj.EXP-15: The applicant should define the timing when the explainability will be available to the end user taking into account the time criticality of the situation, the needs of the end user, and the operational impact.	Ant.MOC EXP-15/16
1B-2B	Obj.EXP-16: The applicant should design the AI-based system so as to enable the end user to get upon request explanation or additional details on the explanation when needed.	Ant.MOC EXP-15/16
1B-2B	Obj.EXP-17: For each output relevant to the task(s), the applicant should ensure the validity of the specified explanation.	
1A–2B	Obj.EXP-18: The training and instructions available for the end user should include procedures for handling possible outputs of the ODD monitoring and output confidence monitoring.	
1A–2B	Obj.EXP-19: Information concerning unsafe AI-based system operating conditions should be provided to the end user to enable them to take appropriate corrective action in a timely manner.	
C4.2(HF). Human-AI teaming		
2A-2B	Obj.HF-01: The applicant should design the AI-based system with the ability to build its own individual situation representation.	Ant.MOC HF-01
2A-2B	Obj.HF-02: The applicant should design the AI-based system with the ability to reinforce the end-user individual situation awareness.	Ant.MOC HF-02
2B only	Obj.HF-03: The applicant should design the AI-based system with the ability to enable and support a shared situation awareness.	Ant.MOC HF-03
2A-2B	Obj.HF-04: If a decision is taken by the AI-based system that requires validation based on procedures, the applicant should design the AI-based system with the ability to request a cross-check validation from the end user.	Ant.MOC HF-04
2A-2B	Obj.HF-05: For complex situations under normal operations, the applicant should design the AI-based system with the ability to identify a suboptimal strategy and propose through argumentation an improved solution.	Ant.MOC HF-05

2A-2B	Corollary Obj.HF-05: The applicant should design the AI-based system with the ability to process and act upon a proposal rejection from the end user.	
2B only	Obj.HF-06: For complex situations under abnormal operations, the applicant should design the AI-based system with the ability to identify the problem, share the diagnosis including the root cause, the resolution strategy and the anticipated operational consequences.	Ant.MOC HF-06 *Dependencies: Obj.HF-05
2B only	Corollary Obj.HF-06: The applicant should design the AI-based system with the ability to process and act upon arguments shared by the end user.	
2B only	Obj.HF-07: The applicant should design the AI-based system with the ability to detect poor decision-making by the end user in a time-critical situation, alert and assist the end user.	Ant.MOC HF-07
2B only	Obj.HF-08: The applicant should design the AI-based system with the ability to propose alternative solutions and support its positions.	Ant.MOC HF-08
2B only	Obj.HF-09: The applicant should design the AI-based system with the ability to modify and/or to accept the modification of task allocation pattern (instantaneous/short-term).	Ant.MOC HF-09
C4.3(HF). Modality of interaction and style of interface		
2A-2B	Obj. HF-10: If spoken natural language is used, the applicant should design the AI-based system with the ability to process end-user requests, responses and reactions, and provide an indication of acknowledgement of the user's intentions.	Ant.MOC HF-10
2B only	Obj.HF-11: If spoken natural language is used, the applicant should design the AI-based system with the ability to notify the end user that he or she possibly misunderstood the information.	Ant.MOC HF-11
2B only	Obj.HF-12: If spoken natural language is used, the applicant should design the AI-based system with the ability to identify through the end user responses or his or her action that there was a possible misinterpretation from the end user.	Ant.MOC HF-12
2B only	Obj.HF-13: In case of confirmed misunderstanding or misinterpretation of spoken natural language, the applicant should design the AI-based system with the ability to resolve the issue.	Ant.MOC HF-13
2A-2B	Obj.HF-14: If spoken natural language is used, the applicant should design the AI-based system with the ability to not	Ant.MOC HF-14

	interfere with other communications or activities at the end user's side.	
2B only	Obj.HF-15: If spoken natural language is used, the applicant should design the AI-based system with the ability to provide information regarding the associated AI-based system capabilities and limitations.	Ant.MOC HF-15
2A-2B	Obj.HF-16: If spoken procedural language is used, the applicant should design the syntax of the spoken procedural language so that it can be learned and applied easily by the end user.	
2A-2B	Obj.HF-17: If gesture language is used, the applicant should design the gesture language syntax so that it is intuitively associated with the command that it is supposed to trigger.	Ant.MOC HF-17
2A-2B	Obj.HF-18: If gesture language is used, the applicant should design the AI-based system with the ability to disregard non-intentional gestures.	Ant.MOC HF-18
2B only	Obj.HF-19: If gesture language is used, the applicant should design the AI-based system with the ability to recognise the end-user intention.	
2B only	Obj.HF-20: If gesture language is used, the applicant should design the AI-based system with the ability to acknowledge the end-user intention with appropriate feedback.	Ant.MOC HF-20
2A-2B	Obj.HF-21: If spoken natural language is used, the applicant should design the AI-based system so that this modality can be deactivated for the benefit of other modalities.	Ant.MOC HF-21
2B only	Obj.HF-22: If spoken (natural or procedural) language is used, the applicant should design the AI-based system with the ability to assess the performance of the dialogue.	
2B only	Obj.HF-23: If spoken (natural or procedural) language is used, the applicant should design the AI-based system with the ability to transition between spoken natural language and spoken procedural language, depending on the performance of the dialogue, the context of the situation and the characteristics of the task.	Ant.MOC HF-23
2B only	Obj.HF-24: The applicant should design the AI-based system with the ability to combine or adapt the interaction modalities depending on the characteristics of the task, the operational event and/or the operational environment.	Ant.MOC HF-24

2B only	Obj.HF-25: The applicant should design the AI-based system with the ability to automatically adapt the modality of interaction to the end-user states, the situation, the context and/or the perceived end user's preferences.	Ant-MOC HF-25
C4.4(HF). Error management		
2A-2B	Obj.HF-26: The applicant should design the AI-based system to minimise the likelihood of design-related end-user errors.	Ant.MOC HF-26
2A-2B	Obj.HF-27: The applicant should design the AI-based system to minimise the likelihood of HAIRM-related errors.	Ant.MOC HF-27
2A-2B	Obj.HF-28: The applicant should design the AI-based system to be tolerant to end-user errors.	Ant.MOC HF-28 *Dependencies: Obj.HF-25 Obj.HF-26 Obj-HF-27
2A-2B	Obj.HF-29: The applicant should design the AI-based system so that in case the end user makes an error while interacting with the AI-based system, the opportunities exist to detect the error.	Ant.MOC HF-29
2A-2B	Obj.HF-30: The applicant should design the AI-based system so that once an error is detected, the AI-based system should provide efficient means to inform the end user.	
C4.5(HF). Failure management		
2B only	Obj.HF-31: The applicant should design the system to be able to diagnose the failure and present the pertinent information to the end user.	Ant.MOC HF-31
2B only	Obj.HF-32: The applicant should design the system to be able to propose a solution to the failure to the end user.	Ant.MOC HF-32
2B only	Obj.HF-33: The applicant should design the system to be able to support the end user in the implementation of the solution.	Ant.MOC HF-33
2B only	Obj.HF-34: The applicant should design the system to provide the end user with the information that logs of system failures are kept for subsequent analysis.	Ant.MOC HF-34

C5. AI safety risk mitigation

LoA	EASA Objectives	Anticipated MOC
C5(SRM). AI safety risk mitigation concept and top-level objectives		
1A-2B	Obj.SRM-01: Once activities associated with all other building blocks are defined, the applicant should determine whether the coverage of the objectives associated with the explainability and learning assurance building blocks is sufficient or whether an additional dedicated layer of protection, called hereafter safety risk mitigation, would be necessary to mitigate the residual risks to an acceptable level.	Ant.MOC SRM-01
1A-2B	Obj.SRM-02: The applicant should establish safety risk mitigation means as identified in Objective SRM-01.	Ant.MOC SRM-02 *Dependencies: Obj.SRM-01

C6. Organisations

LoA	EASA Objectives	Anticipated MOC
C6.1(ORG). High level provisions and anticipated AMC		
1A-2B	Prov.ORG-01: The organisation should review its processes and adapt them to the introduction of AI technology.	
1A-2B	Prov.ORG-02: In preparation of the Commission Delegated Regulation (EU) 2022/1645 and Commission Implementing Regulation (EU) 2023/203 applicability, the organisation should continuously assess the information security risks related to the design, production and operation phases of an AI/ML application.	Ant AMC ORG-02
1A-2B	Prov.ORG-03: Implement a data-driven 'AI continuous safety assessment' process based on operational data and in-service events.	Ant.AMC ORG-03 *Dependencies: Obj.EXP-09
1A-2B	Prov.ORG-04: The organisation should establish means (e.g. processes) to continuously assess ethics-based aspects for the trustworthiness of an AI-based system with the same scope as for Objective ET-01.	Ant.AMC ORG-04 *Dependencies: Obj.ET-01
1A-2B	Prov.ORG-05: The organisation should adapt the continuous risk management process to accommodate the specificities of AI, including interaction with all relevant stakeholders.	Ant.AMC ORG-05

1A-2B	Prov.ORG-06: The organisation should ensure that the safety-related AI-based systems are auditable by internal and external parties, including especially the approving authorities.	
C6.2(ORG). Competence considerations		
1A-2B	Prov.ORG-07: The organisation should adapt the training processes to accommodate the specificities of AI, including interaction with all relevant stakeholders (users and end users).	Ant.AMC ORG-07 *Dependencies: Prov.ORG-06 Prov.ORG-07
1A-2B	Prov.ORG-08: The organisations operating the AI-based systems should ensure that end users' licensing and certificates account for the specificities of AI, including interaction with all relevant stakeholders.	

Table 7. Objectives from EASA Concept Paper with guidance for level 1&2 ML applications (EASA, 2024).