safety

MDPI

Article

Development of a Holistic Assessment Framework for the Design of AI-Based Automation

Sybert Stroeve [1,*], Barry Kirwan [2] and Mariken Everdij [1]

[1] Safety and Human Performance, Royal Netherlands Aerospace Centre NLR, 1059 CM Amsterdam, The Netherlands; mariken.everdij@nlr.nl
[2] Future Human Factors, Westbury-on-Trym, Bristol BS9 3HQ, UK; barrykirwan@bkhf.onmicrosoft.com
* Correspondence: sybert.stroeve@nlr.nl

Abstract

There is a need to ensure that the application of artificial intelligence (AI) in increasingly automated operations is safe, human-centric, and trustworthy, and respects ethical principles. To this end, this paper presents an innovative holistic assessment framework to support certification-aware design of AI-based sociotechnical systems with a range of levels of automation along multiple design stages from low to high technology and human readiness levels (TRLs/HRLs). The holistic scope considers a range of relevant key performance areas (KPAs): safety, resilience, security, Human Factors, accountability, responsibility, liability, efficiency, societal sustainability, and environmental sustainability. The core of the framework is a seven-step cycle that assesses the KPAs for critical scenarios and evaluates the combined performance, including uncertainty and trade-offs. This provides feedback to either adapt the design at the same TRL/HRL or refine it at higher TRLs/HRLs. The framework enacted by a toolbox of assessment methods for the KPAs. The framework has been developed in the aviation domain, but it is formulated in a generic manner, enabling application to various AI techniques and operational domains. Its application is illustrated in detail for an air traffic management use case that employs an AI-based system to support air traffic controllers in sequencing aircraft. It is concluded that the framework provides a viable approach for holistic assessment of AI-based sociotechnical systems.

Keywords: artificial intelligence; advanced automation; certification; feedback to design; technology readiness level; human factors; safety; assessment framework

Academic Editor: Raphael Grzebieta

Received: 10 February 2026
Revised: 2 July 2026
Accepted: 2 July 2026
Published: 7 July 2026

1. Introduction

1.1. Foundations for Integration of AI-Based Technologies in Sociotechnical Systems

Artificial intelligence (AI) and its machine learning (ML) component are key drivers of innovation, enabling higher levels of automation (LOAs) across multiple domains, improving operational efficiency for complex tasks and supporting human operators and organisations. A primary concern of stakeholders involved in the transition to higher LOAs is to establish the necessary conditions and standards to ensure that the solutions are aligned with regulations and can be certified.

As a basis, the AI Act of the European Union (EU) [1] has laid down a legal framework for AI systems that aims to foster human-centric and trustworthy AI, ensuring health, safety, and fundamental rights. The AI Act has adopted a wide definition of AI system as "a 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." The regulation provides a range of requirements for so-called high-risk systems. An AI system is considered high-risk (Article 6 of [1]) if it serves as a safety component, if it is a product covered by a range of specific EU laws and must pass a third-party conformity assessment, or if it is on a specified list. The indicated EU laws (Annex I of [1]) include legislation for marine equipment, motor vehicles, rail systems, and civil aviation.

The AI Act provides amendments to these laws specifying that the requirements of the AI Act for high-risk systems shall be taken into account (Chapter XIII of [1]). For instance, for civil aviation (for which the framework described in this paper has been developed), articles in [2] concerning airworthiness, air traffic management (ATM)/air navigation service providers (ANSPs), and unmanned aircraft were extended with such reference to the AI Act. The requirements for high-risk AI systems concern the establishment of a risk management system; rules for data governance, technical documentation and record-keeping; demands for transparency, informing end-users and human oversight; and rules regarding accuracy, robustness and cybersecurity. An important foundation for the development of the AI Act has been the ethics guidelines for trustworthy AI [3] developed by the High-Level Expert Group on AI set up by the European Commission. Herein, a broad range of requirements are posed for human agency and oversight, technical robustness and safety, privacy and governance, transparency, diversity, non-discrimination and fairness, societal and environmental well-being, and accountability.

Various researchers and committees have characterised the complexity of the innovations needed for advanced automation and increasingly autonomous AI-enabled systems, developing research agendas to address identified barriers. A Committee on Autonomy Research for Civil Aviation [4] identified as key challenges certification, verification, and validation processes, as well as trust and decision-making for increasingly autonomous systems. This led to a broad-ranging research agenda, from modelling and simulation of adaptive/nondeterministic systems to certification processes, roles of personnel and systems, and stakeholder trust. Adopting a more narrow engineering scope, a recent Committee on using ML in safety-critical applications [5] identified that the ML and safety engineering communities have different methods, standards and cultures. They concluded that the integration of ML components in safety-critical applications requires research towards closing these gaps and towards new standards, regulations, and testing methods.

In parallel, a Committee on Human–System Integration (HSI) [6] addressed the state-of-the-art and research needs for interaction and cooperation between humans and AI systems, considering human–AI team models, situation awareness, transparency and explainability, trust, decision bias, and training. Building on Bainbridge's seminal paper "Ironies of automation" [7], Endsley [8] discusses ironies of AI, such as 'the more intelligent the AI, the more obscure it is, and the less able people are to determine its limitations and biases and when to use the AI' and 'the more natural the AI communications, the less able people are to understand the trustworthiness of the AI'. In [9], test and evaluation challenges in AI-enabled systems are addressed, focusing on the central question: how can sufficient confidence in AI-enabled systems be achieved? It discusses how human–AI interfaces present new challenges as responsibilities shift between humans and intelligent machines and new concepts of operations emerge. More prominence is argued for human readiness levels (HRL) and the user interface/user experience (UI/UX) for AI-enabled systems. Measures of performance and effectiveness, including assessments of user trust and justified confidence, must be formulated during system design and development, and assessed throughout test and evaluation and after system fielding.

The European Union Aviation Safety Agency (EASA), the principal European regulatory authority for European aviation, published an AI Roadmap [10] calling for a human-centric approach to AI in aviation. EASA issued initial guidance [11] for particular ML applications, which builds on the ethical guidelines of [3]. The guidance also provides objectives and anticipated means of compliance that consider a holistic scope, encompassing AI trustworthiness, AI assurance, Human Factors (HF), and safety risk mitigation. Similarly, Díaz-Rodríguez et al. [12] promote a multi-disciplinary vision of trustworthy AI, where the concept of responsible AI plays a key role, effected by auditability and accountability during its design, development and use, and supported by regulatory 'sandboxes'.

In the development of AI-supported systems and operations, there will be a shift in the level of authority towards increasingly autonomous systems with decreasing in-the-loop roles of human operators. This implies that a part of the intelligent contributions of human operators is taken over by contributions of AI-based systems and that the roles and responsibilities of human operators change. Such a shift has considerable legal implications for the allocation of responsibility and liability for product and service developers versus the organisations and persons providing services and using the products. Overall, the above papers and reports address the need for a broad-scope, holistic approach for the development and evaluation of advanced automation with AI-based systems. Such an approach must emphasise the role of Human Factors in understanding uncertainty and safety risks in sociotechnical systems with diverse levels of automation. It must also address the impact on responsibility and accountability in design and operations, data governance policies, and new testing methods.

1.2. Certification and Safety Assurance Issues for AI-Based Systems

Certification concerns the legal recognition by a certification authority that products, services, persons or organisations comply with requirements and standards. A gap analysis of a range of existing development assurance standards [13] regarding their limitation for application to AI-based systems highlights that the data-driven paradigm of ML may not be adequately addressed by the existing standards. Identified issues include requirements traceability, differences in ML development and system life cycles, and lack of verification methods for datasets. Furthermore, while safety assessment is a key element of standards towards certification, it should be realised that safety is not an intrinsic property of an AI-based system. The safety impact of a particular AI component in a system depends on the dynamic interactions with other systems, humans working with operational procedures, and contextual conditions. Traditional safety assessment approaches such as fault trees and failure mode and effect analysis have their origins in assurance schemes for physical components, which may fail/break and for which statistical quality control approaches can be applied. These approaches are known to have limitations for assessing and controlling the safety impact of software, AI-based systems, and Human Factors. While these limitations also apply to systems and operations without AI components, their implications may be aggravated due to the new aspects brought by ML and due to the shift in contributions of humans and AI-based systems in the operations.

In system design and in system life cycle processes in general, iteration and recursion are important for the progressive refinement of processes, system elements, and systems. In particular, 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 are especially important for high-LOA concepts with key roles and responsibilities granted to AI-based systems. In particular, the effective use of AI requires an agile and cyclic development methodology (especially given the high rate of development of AI). Aligning concepts

with compliance objectives from the early stages of design is vital, as it helps to address the challenges of contextualising these objectives within the development pipeline and to ensure the system meets regulatory requirements throughout its life cycle. Whilst 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. Furthermore, the development and approval of advanced automation and AI-based systems require collaboration with a diverse set of stakeholders and experts/practitioners (e.g., safety, legal, data science, Human Factors, etc.) during the life cycle stages. As such, proactive integration of compliance objectives, coordination with multiple stakeholders and experts, and incorporation of a holistic scope accounting for all relevant performance areas are all needed for effective design cycles. Failing to do so from an early stage of the design may lead to costly and time-consuming redesigns, missed hazards, or even to an overall failure of the development, since key requirements cannot be achieved or the impact on (possibly conflicting) key performance areas cannot be sufficiently balanced. Overall, failure to address all the performance areas in an integrated fashion with all key stakeholders and experts may lead to adoption failure of AI-based systems.

1.3. Objective and Structure of the Paper

To address the above needs, this paper presents an innovative holistic assessment framework to support certification-aware design of AI-based sociotechnical systems with advanced automation along multiple design stages (from low to high TRL/HRL). The framework was developed in [14] during the HUCAN project [15], which aims to pioneer approval processes in the ATM domain. The scope of the framework excludes adaptive and generative AI systems. The framework is formulated in a generic manner, allowing for application in various domains and AI techniques. For the purpose of this paper and as a first step towards empirical validation, it is illustrated by a use case on an AI-based Intelligent Sequence Assistant (ISA) for air traffic control, which was developed in the HAIKU project [16].

The objectives and structure of the paper are as follows:

I. To present an overview of the ten key performance areas (KPAs) of the holistic scope identified for the framework. These represent the high-level areas (e.g., safety, security, sustainability, etc.) that need to be addressed to ensure a safe, resilient, effective and ethical integration of AI-based systems into operations [Section 2].
II. To describe in detail the holistic assessment framework, including an assessment 'compass' and a holistic assessment cycle for maturing system design, alongside a toolbox of associated methods. This section shows 'what needs to be done when', translating the framework into an active and development-stage-linked process [Section 3].
III. To illustrate the application of the holistic assessment framework to the ISA use case for air traffic control. Since at present there is no medium LOA AI-based system active in aviation operations, a use case at TRL 5–6 is utilised. This use case nevertheless provides a suitably challenging testbed for many of the holistic framework methods and most KPAs [Section 4].
IV. To discuss the holistic framework in association with its application. Since AI integration into the cockpit or ATM operations room is currently at the research and development stage (e.g., TRLs 1–6) rather than in live operations, the insights arising necessarily remain qualitative at this stage [Section 5].
V. To conclude on the utility of the holistic framework and its constituent methods, and to note further research needs and limitations of the study [Section 6].

VI. To provide more detail of toolbox method application, both for practitioners in the aviation domain and for experts in other domains, allowing them to gain insights into how well the methods could transpose into their application industry [Appendices A–G].

2. Holistic Scope for Assessment of AI-Based Systems

Given the above justified need for a holistic scope in support of design, validation, and approval of AI-based sociotechnical systems, a broad range of KPAs needs to be accounted for in associated assessments. The KPAs for the holistic scope addressed in the framework are shown in Figure 1 and explained in Table 1. The set has been identified in the HUCAN project [14], building on the regulations in the AI Act [1], the associated ethics guidelines [3] and the EASA AI roadmap and guidance [10,11]. As such, it is considered a sufficiently complete and comprehensible set, which could be further extended if needed for emerging applications. For each of these KPAs, the outcome of the evaluation can be Acceptable (green), expressing a good level of performance for the KPA, Tolerable (orange), expressing some medium level that may be tolerated given trade-offs with other KPAs, or Unacceptable (red), expressing a poor and insufficient performance.

Figure 1. Prime KPAs for AI-based systems in the holistic assessment framework, which can be evaluated for Acceptable (green), Tolerable (orange), or Unacceptable (red) performance.

Table 1. Definition of prime KPAs for AI-based systems in the holistic assessment framework.

KPA	Definition
Safety	Safety has been and remains the prime KPA in certification, also for advanced automation and AI-based systems. Traditional functional hazard assessment (FHA) approaches lead to ranges of requirements for system design and development using allocation of assurance levels (AL), 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. 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.

Table 1. *Cont.*

KPA	Definition
Resilience	In a holistic framework the level of resilience of AI-based 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. Resilience Engineering is the discipline that focuses on developing principles and practices to support the resilience of sociotechnical systems [17], so as to support the safety and efficiency of its operations. In a review of resilience papers [18], 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, pose a risk to safety and the continuity of operations. Especially, the use of (AI-based) advanced automation in highly connected sociotechnical systems poses new cybersecurity threats, like data poisoning, adversarial attacks, and AI model theft. Evaluation and control of security risks is an important component in the development of AI-based systems for advanced automation.
Human Factors	Topics for validating 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, and training [6]. HF has 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 to other stakeholders are key aspects for successful introduction of higher LOAs. In [11], the importance of HF for AI is recognised and many objectives are provided for AI operational explainability, human–AI teaming, human–AI interfaces, and error, failure and workload management. HRL [19] can be used as a structure to validate the HSI in advanced automation concepts using AI-based systems.
Accountability	Accountability is one of the seven requirements for trustworthy AI systems as defined in [3] and explained in [12]. Accountability is linked to the principle of fairness and as such is closely related to risk management, 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 is 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 that address aspects such as the interpretability and 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 methodical manner. Trade-offs should be explicitly acknowledged and documented as part of risk management, supporting the continuous review of their appropriateness. Furthermore, accountability concerns the possibility of redressing an AI-based system if it has contributed to biased or adverse outcomes.

Table 1. *Cont.*

KPA	Definition
Responsibility	Responsibility generally means that persons in charge of tasks accept the consequences of their actions/decisions to undertake the tasks, whether they turn out 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 regulations of the domain of practice in which the AI system is to be used [12]. 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 implies a shift in responsibility from human operators to shared responsibility between human operators and the 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 [20]. 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 such as ANSPs and airlines [21]. 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 [1]. 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 [12]. Such sandboxes support pre-market auditability and conformity checks, as well as post-market monitoring and accountability.
Efficiency	Efficiency relates to the costs and benefits for stakeholders. In ATM, efficiency also addresses matters like punctuality and the 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, while the volume of airspace remains the same, a higher number of conflicts between those flights can be expected. In a similar way, a lower number of flight movements generally leads to fewer conflicts.
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 [1]. It includes AI ethics issues such as avoidance of bias and discrimination, third-party safety, fairness, privacy, transparency, job meaningfulness and human oversight/agency.
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 [1]. It entails that the system's development, deployment and use process, as well as its entire supply chain, are assessed regarding their environmental impact, such that the most environmentally friendly choices can be made. Existing guidelines for environmental assessment, such as [22], 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 the impact of the AI ecosystem on the environment.

3. Holistic Assessment Framework for Design of AI-Based Automation

3.1. Introduction

A high-level overview of the holistic assessment framework for the design of AI-based automation supporting development to a higher TRL/HRL is shown in Figure 2. It uses a cyclical approach to support the maturation of AI-based automation using the following main elements:

- *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 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 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 that should be adapted at the same TRL/HRL, or else it identifies/refines requirements or assurance levels towards a more mature design. The types of feedback are explained in Section 3.4.

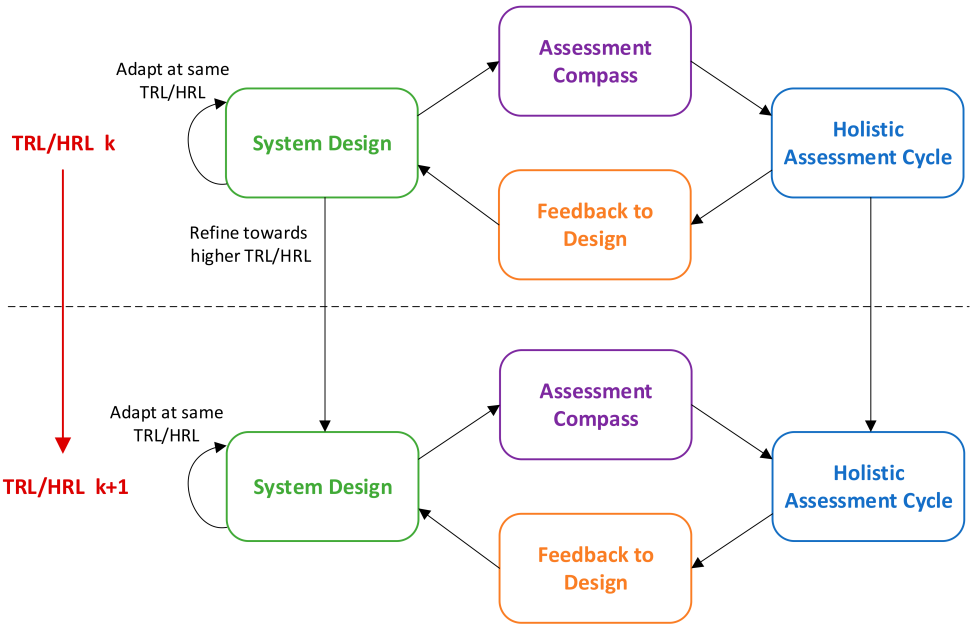


Figure 2. High-level overview of holistic assessment framework for design of AI-based automation supporting development to a higher TRL/HRL.

3.2. Assessment Compass

The assessment compass provides the basis for the holistic assessment cycle via the following elements:

- Determine level of automation/AI (Section 3.2.1);
- Determine technology and human readiness levels (Section 3.2.2);

- Determine key performance areas (Section 3.2.3);
- Identify certification objectives (Section 3.2.4).

3.2.1. Determine Levels of Automation/AI

Automation may be defined as "The use of control systems and information technologies reducing the need for human input, typically for repetitive tasks" [11], and a level of automation (LOA) then refers to the extent to which the need for human input has been reduced. Various taxonomies for LOAs exist [23], typically ranging from manual operation without any assistance to fully autonomous operations. In [11], AI applications are classified into six categories from AI Level 1A to AI Level 3B, which describe the level of human involvement in the applications. In LOA taxonomies for ATM R&D [24,25] up to six levels of automation are defined, which can be associated with these AI levels. 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, as described in Table 2. While these LOAs stem from aviation, their descriptions are generic and may be applied in other domains. Alternatively, one may adopt a domain-specific LOA taxonomy [23].

Table 2. Levels of automation (SESAR [24,25]) and AI application levels (EASA, [11]).

Level of Automation (SESAR)		AI Appl. Level (EASA)	Description
0	Low automation	1A	The human operator has full authority, takes all decisions and implements actions with or without execution support from the AI-based system.
1	Decision support	1B	The human operator has full authority, receives decision support from the AI-based system, and implements actions with or without support from the system.
2	Resolution support	2A	The human operator has full authority, receives resolution support from the AI-based system, and validates the proposed solution or comes up with a different solution. The system implements the actions under the direction of the operator.
3	Conditional automation	2B	The human operator has partial authority and supervises the performance of the AI-based system. The system selects solutions and implements the actions. The operator overrides or improves solutions that are not deemed appropriate.
4	Confined automation	3A	The human operator has limited authority, and only supervises and possibly intervenes if requested by the AI-based system. The system takes all decisions and implements all actions, except when it comes out of a predefined scope.
5	Full automation	3B	There is no human operator. The AI-based system decides and acts on its own.

3.2.2. Determine Technology and Human Readiness Levels

Technology and human readiness levels (TRLs/HRLs) support programme management for the development of new technology and ways of working. TRLs provide a nine-level scale to describe the maturity of technology, which has been developed at NASA [26] and is widely used. Recognising that many system development programmes 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 [19]. Here, human readiness is the readiness of a technology for use by the intended human

users in a specified intended operational environment. Table 2 gives an overview of HRLs and TRLs as provided in [19].

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. Transition from lower to higher levels is based on (funding) decisions/approvals of R&D organisations and/or system designers/manufacturers. Approval by a certifying authority especially relates to TRL/HRL 8, but as part of cyclic development, the regulator should be involved at lower TRLs/HRLs to ensure 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 [19], TRLs and HRLs 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 may provide objective evidence 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; see also the guidance in [19,27].

Table 3. Human and technology readiness levels, based on Table 5-1 of [19].

Level	Human Readiness Level (HRL)	Technology Readiness Level (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

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 which KPAs should be addressed for a particular application. The framework includes the KPAs addressed in Section 2, namely safety, resilience, security,

Human Factors, accountability, responsibility, liability, efficiency, societal sustainability, and environmental sustainability.

3.2.4. Identify Certification Objectives

In this step of the assessment compass, relevant objectives are identified that have been defined by a certifying authority. For example, for AI-based systems and advanced automation in aviation, a preliminary set of objectives is listed in [11]. For the identification of certification objectives in the assessment compass, it is worthwhile to consider the relevance of all objectives. 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.

3.3. Holistic Assessment Cycle

A detailed overview of the processes in the holistic assessment framework is shown in Figure 3. It shows the design of an AI-based system and the associated concept of operations (ConOps), the assessment compass, the holistic assessment cycle, and the feedback processes for the design, where a transition to a higher TRL/HRL of the system and ConOps is supported. The holistic assessment cycle applies methods from a toolbox, which is explained in Section 3.5. The cycle consists of the following subsequent steps:

1. Identify objectives, scope, and 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).

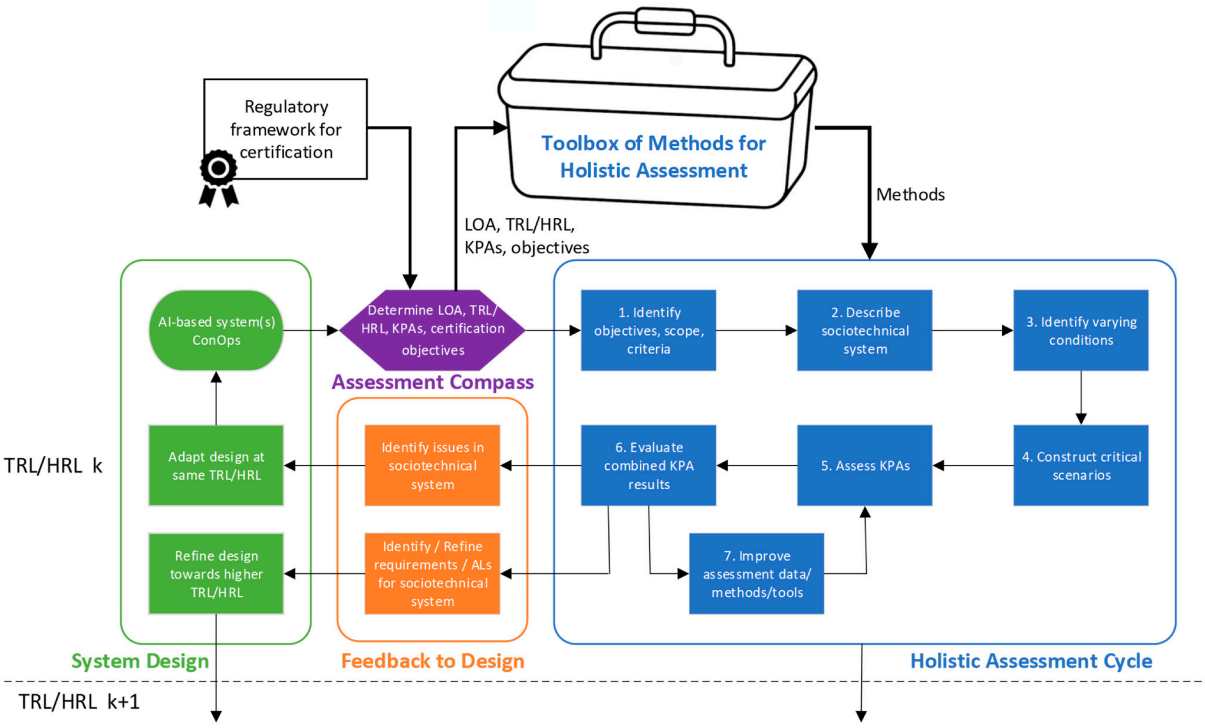


Figure 3. Overview of relations between system design and feedback to design from a holistic assessment cycle in support of transitioning to next technology/human readiness levels.

These steps have been inspired by the assessment cycles for safety risk of [28,29], while they have been adapted for other KPAs and for development towards higher TRL/HRL. The steps are explained next.

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 results of the assessment compass, such as the definition of the KPAs that will be addressed (Section 3.2.3) and relevant certification objectives (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 Figure 4. 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					
B					
C					
D					
E					

Figure 4. Example of a generic risk matrix with criteria for acceptability of likelihood–severity combinations: Unacceptable (red), Tolerable (orange), or Acceptable (green).

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 con-

text, 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 in the ConOps and the considered AI-based system(s). 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 the operations of the sociotechnical system. These can include frequently occurring conditions, including normal sensor errors, normal transmission delays, typical reaction times of human operators, differences in interpretations by humans, normal weather variability, but can also include rarer conditions such as system failures, extreme weather, and 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 brainstorming 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 a short 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 falls 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.

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 the most effort of the steps in the holistic assessment cycle. It is supported by a variety of methods and tools (see Section 3.5), 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 result of the KPA assessment is an overview of the assessed KPAs, including the key contributions or sensitivities impacting the 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 forms a key basis for this evaluation. The point of a holistic assessment is to see the whole, in this case all system KPAs together, to illustrate where the system is performing well, and where it may yet need attention, or indeed if, taken as a whole, it is not worth pursuing. Thus, while a new system concept may perform well on certain KPAs, unless it performs to a certain degree on a range of them, it is unlikely to go ahead into deployment and operation. Holistic views help decision-makers, e.g., senior management in an organisation considering implementing an AI-based system, to decide whether the benefits outweigh the risks and costs.

To obtain the holistic view on the range of KPAs, the assessment results (Section 3.3.5) need to be evaluated for each KPA by comparison to acceptability criteria defined in step 1 of the holistic assessment framework (Section 3.3.1). These criteria are specific for each KPA and they may depend on the TRL/HRL, e.g., being more high-level in early design and more detailed in later design stages. They may be qualitative or quantitative performance requirements, they may concern absolute or relative performance, or they may refer to required quality assurance processes that have been followed in the design. The acceptability criteria may discern various levels. For the argumentation in this paper, we assume three levels for each KPA and the associated results of the KPA assessment:

- *Acceptable*. The performance of the sociotechnical system is at a good level for the KPA. This good performance is supported by a number of specified factors. The performance may be sensitive to some factors, implying that it may deteriorate if these factors are changed in the design.
- *Tolerable*. The performance of the sociotechnical system is at a medium level for the KPA, which may be tolerated given trade-offs with other KPAs. This medium performance is influenced by a number of specified positive and negative factors. The performance may be sensitive to some factors; i.e., associated changes in the design would have a considerable impact on the KPA (either positively or negatively).
- *Unacceptable*. The performance of the sociotechnical system is at a low level for the KPA, which is not acceptable. This low performance stems from specified factors. The performance may be sensitive to some factors, implying that it may improve if these factors are changed in the design.

If the uncertainty in the assessment result of a KPA is too high to distinguish between two acceptability levels, both levels are assigned. For instance, the acceptability of a KPA is either Unacceptable or Tolerable.

This evaluation is done for each KPA in the holistic view, which can be represented by a graph as shown in Figure 5, indicating the level of acceptability of each KPA, or the lack of an assessment at a particular design stage.

Figure 5. Example of a holistic view on the acceptability levels of a particular design during the development, where green is Acceptable, orange is Tolerable, and red is Unacceptable. Having multiple acceptability levels for a KPA indicates uncertainty in the assessment, whereas having no acceptability levels means that the KPA has not been assessed at this design stage.

In combination, the holistic evaluation may lead to the following types of conclusions and bases for feedback to design.

1. It may be concluded with sufficient certainty that the performance of the AI-supported sociotechnical system is Unacceptable for one or more KPAs, or there are too many KPAs with Tolerable performance that are not sufficiently traded off with other 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 or Tolerable for all KPAs with a suitable trade-off for KPAs with Tolerable acceptability. 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 be 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 methods used in the assessment(s) may be extended, such that the uncertainty in the assessment results may be reduced.
3. Other assessment methods or tools 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 evaluation of the combined KPA results, feedback to design is provided either by identifying issues in the sociotechnical system, or by identifying/refining requirements or assurance levels (ALs), as explained next.

3.4.1. Identify Issues in Sociotechnical System

If the performance of the AI-supported sociotechnical system is (potentially or certainly) Unacceptable for one or more KPAs or if it is only Tolerable for too many 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 poor performance for one or more KPAs are shared with the developers. Furthermore, information on the sensitivity of the performance for aspects in the design is shared for all assessed KPAs (with better or worse performance). The information on these issues supports the designers in improving the AI-based system(s) and ConOps, and in considering trade-offs between KPAs. Aspects that significantly contribute to poor performance on particular KPAs need to be improved, while the changes to the design should not lead to considerable deterioration for other KPAs.

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. 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 model forms 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 specific agents (e.g., 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 these agents is within an expected range, but that it is sensitive to performance variations, such that it may become unacceptable for agents' performance outside of the expected range. In line with such a level of criticality, assurance levels can be set for the performance of agents, which define the level of rigour in ensuring that the performance of the agents remains within ranges that are compatible with the overall acceptable performance of the sociotechnical system.

The requirements and assurance levels determined at a particular readiness level TRL/HRL 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.

3.5. Toolbox of Assessment Methods

The framework includes a toolbox of methods to support the elements of the holistic assessment cycle. In this context, a method is any technique, method, standard, methodology, database, or model that can be used in support of the evaluation of a KPA in an operation with (AI-based) advanced automation. A detailed overview of such methods is available in [14], which describes to which KPAs the method can contribute, at what maturity levels (TRLs/HRLs) the method can be applied, for what LOAs the method can be applied, for which types of objectives [11] the method can be a potential means of compliance, how the method deals with uncertainty, what the level of technical complexity of the method is, and what benefits and limitations of the method are. The toolbox includes established as well as more innovative methods, and it is adaptable, allowing extension with additional methods. An overview of the set of methods, such as that identified in [14], is provided in Table 4. It shows for each method the associated KPAs, TRL/HRL, and LOA. Furthermore, selected methods that have been applied in the use case are presented in Appendices A–G.

Table 4. List of assessment methods in the toolbox and associated KPAs, TRL/HRL, and LOAs.

Method	KPAs	TRL/HRL	LOA
Agent-Based Modelling & Simulation (ABMS)	Safety, Security, HF, Resilience, Efficiency	TRL 2–9 HRL 2–8	0–5
AI Risk Management Framework (AI-RMF)	Accountability, Responsibility, HF, Safety, Security	TRL 4–9 HRL 4–9	0–5
Bias, Uncertainty and Sensitivity Analysis	All (quantitative)	TRL 2–9 HRL 2–9	0–5
Environmental Assessment of AI Ecosystem	Environmental sustainability	TRL 3–9	0–5
Failure Modes and Effects Analysis (FMEA)	Safety	TRL 3–6	0–5
Fundamental Rights and Algorithms Impact Assessment	Societal sustainability	TRL 4–9 HRL 4–9	0–5

Table 4. *Cont.*

Method	KPAs	TRL/HRL	LOA
Hazard and Operability study (HAZOP)	HF, Safety	TRL 3–9 HRL 3–9	0–5
Heuristic Evaluation	HF, Safety, Efficiency	HRL 1–6	0–4
Human-in-the-Loop (HITL) Simulation	HF, Safety, Efficiency	HRL 5–9	0–4
Responsibility, Accountability and Liability Analysis	Responsibility, Accountability, Liability	TRL 4–9 HRL 4–9	0–5
Safety and Security Scanning	Safety, Responsibility, Accountability, HF, Resilience, Security	TRL 1–6 HRL 1–6	0–5
Security Risk Assessment Methodology (SecRAM)	Security	TRL 2–8	0–5
System Functionality and Flow model (NSV-4)	Safety, HF	TRL 3–6 HRL 3–6	0–5
Task Analysis	HF, Efficiency	HRL 3–6	0–4
User Requirements Analysis	HF, Safety, Efficiency	HRL 3–6	0–4

4. Use Case: Intelligent Sequence Assistant (ISA)

4.1. Use Case Choice

Currently, AI-based systems are not in operational use in air traffic ops rooms or cockpits, though some AI tools can be used for non-safety-critical aspects such as enhanced weather prediction in ATM. However, a number of research projects are researching their development and application, e.g., the HAIKU [16] and JARVIS [30] projects in European aviation research. In particular, the recently completed HAIKU project had six use cases, all of which involved operational end users in the design and development phases:

- UC1—Cockpit startle response detection and directed situation awareness, to assist single pilots in recovering from sudden disorienting events such as lightning strike, and severe wake vortex [TRL 5, AI Level 1B, using the extreme gradient boosting AI technique].
- UC2—Cockpit re-routing assistant, to assist pilots in finding an alternative airport following a major re-routing requirement due to a major weather disturbance or unexpected military activity [TRL 4, AI Level 2A/2B, using a supervised learning AI algorithmic approach].
- UC3—Urban Air Mobility (UAM) traffic coordinator, for future traffic coordination of drones, drone swarms and sky taxis in major cities [TRL 2, AI Level 3A, no formal AI approach utilised as TRL 2, 'wizard of Oz' approach adopted instead].

- UC4—Single-runway arrival/departure sequencing assistant ISA to assist tower controllers in busy periods at a single-runway airport [TRL 5–6, AI Level 2A, neural network-based/XG-Boost AI].
- UC5—Safety Dashboard for airport incident analysis, to derive new actionable safety improvement insights for London Luton Airport [TRL 7, AI Level 1A; Feature extraction AI plus Small Language Model for text-based query and answer function].
- UC6—Chatbot for passengers evading airborne pathogen outbreaks in airports (TRL 6, AI Level 1A, Google-Vertex AI approach].

Since the authors had significant access to the HAIKU project, it was decided to select one of the six use cases for the application of the holistic framework. UC3 was discounted because it was low-TRL and did not involve any actual AI. UC5 and UC6 were similarly discounted, as although they were higher-TRL, the AI Level for both use cases was only 1A, therefore offering very limited 'challenge' in terms of the KPAs. UC1 was also considered but is a 'single pilot operation' which is currently not in operational practice, and thus would add more uncertainty into any evaluation of the framework. UC2 was also initially considered (moderate TRL and high AI Level), but there were key aspects of the use case that were commercially confidential; therefore, it was not a viable potential candidate.

Of these use cases, UC4 was therefore selected for application of the holistic framework as it was AI Level 2A and TRL 5–6, thus serving as one of the more mature and higher-LOA test case possibilities. UC4 also had significant engagement with senior management, rendering evaluation of KPAs such as Accountability and Responsibility more amenable. UC4 was also developed and evaluated in a very realistic training simulator with licensed end users (tower controllers), and so was the closest to the real operational environment of all the use cases (except UC5).

The remainder of this section shows the application of the holistic assessment framework to the ISA use case in ATM. First, in Section 4.2, ISA and its development are presented. Next, Sections 4.3–4.5 illustrate the application of the assessment framework, addressing the assessment compass, the holistic assessment cycle, and the feedback to design. Section 4.6 gives an overview of the performance of the applied assessment methods. Details of the assessment methods and illustrative results for the ISA use case are gathered in Appendices A–G.

4.2. ISA Development

ISA is an AI-based tool that aims to support and enhance decision-making for tower air traffic controllers (ATCOs) by optimising runway utilisation in a single-runway airport. It provides sequence suggestions for both arriving and departing aircraft, aiming to streamline operations and improve efficiency. ISA was developed as a high-fidelity prototype for operations at Alicante Airport (Spain) in the HAIKU project [16]. The prototype has been validated in Human-in-the-Loop simulation [31], but it has not gone into operation at Alicante Airport at this time.

The development of ISA was intended to provide the following benefits:

- Enhance overall performance, in terms of the number of arrivals and departures safely managed.
- Reduce mental workload for ATCOs during peak traffic hours.
- Improve situation awareness by keeping track of the number of arrivals and departures, and by generating and maintaining data logs on everything that has happened.

4.2.1. Architecture of ISA

The ISA prototype for sequencing advisories consists of the following modules [32]:

- *Core Processing Module.* Responsible for the way messages and information are exchanged, it ensures data is ingested in a timely fashion, processes are triggered, raw data and results are stored and communicated to the other modules that need them, and, in general, for the coordination of data processing.
- *ETA (Estimated Time of Arrival) Calculator Module.* Receives input (aircraft characteristics, positions and speed) and applies a machine learning (XGBoost regressor) model to predict the remaining seconds until an arriving aircraft lands. The system chooses between two trained models depending on the active runway (westward or eastward). The models have been trained on the trajectories of the aircraft included in approximately 1000 h of training simulations (which translates to around 10 k departures and 10 k landings) specific to the Alicante airport approach patterns.
- *Sequence Calculation Module.* Responsible for generating the proposed sequence in which aircraft will use the runway and the assigned (foreseen) times each one is expected to do so (use the runway). It receives as input aircraft and airport static and dynamic information, including departure gates and taxi times, ETA from the ML-based module, CTOT (Calculated Take-Off Time) for departures, aircraft model and wake turbulence category. The sequence is determined using CPLEX and the GNU Linear Programming Kit (GLPK) for solving large-scale linear programming and mixed-integer programming in the Python Pyomo version 6.8.0 package.
- *Explanation Generation Module.* Responsible for generating the explanations shown in the HMI. In general, its input is a number of sequence calculations and its output is a set of comprehensive textual explanations addressing different ATCO needs. The details of the input and output depend on the selected level in Construal Level Theory (CLT) [33] from a minimalist 'headline' for a busy air traffic controller, to a very detailed explanation more suited to a data scientist/analyst (CLT has six levels of explanation).
- *Human–Machine Interface.* Responsible for the interaction of the ATCO with the system, providing information to the ATCO and enabling the ATCO to perform the actions needed, ranging from a request for more information to switching off the sequence recommendations.

4.2.2. ISA-Supported Runway Operations

ISA constantly monitors departing and arriving traffic and adapts to situations in real-time. Taking into account the traffic conditions and the airport's operational constraints, such as its runway–taxiway connections, parking stand configurations and ground movement patterns, it proposes an ordering of the aircraft sequence for runway use. ISA can be of particular use when an aircraft in the inbound stream causes a perturbation in the sequence, such that re-sequencing is advisable. ISA can detect such an anomaly, e.g., an inbound aircraft coming in too fast, farther out than an ATCO would normally detect such an overspeed, and present a new sequence to the ATCO. The ATCO can either accept the new sequence, or reject it and take alternative action (e.g., contacting the 'speeding' aircraft).

Figure 6 shows the electronic flight strips in the ATCO display, which show the sequence of the aircraft departing (in green) and arriving (in orange). When a change is advised by ISA, the slot numbers (1, 2, 3, etc.) for landing/take-off aircraft begin to blink and a magenta box highlights the change. In Figure 6, the number '2' in the magenta box, along with the upward arrow, indicates that the aircraft BAW412 has been moved up the landing sequence, to number 2 (it was formerly third in the overall sequence). The arrival/landing sequence advised is now RYR landing first, then BAW landing, then KLM taking off. Originally the KLM would have taken off after the RYR had landed, but because the BAW was approaching too fast, ISA judged that the separation between the KLM taking

off and the BAW landing was too close and could result in a go-around for the BAW and an aborted take-off for the KLM. In support of the decision-making by the ATCO, ISA also gives a brief explanation for each re-sequencing, e.g., 'BAW412 approaching too fast'. This explainability in ISA is multi-levelled according to CLT (with a maximum of six levels), from 'headline' to more detailed [33].

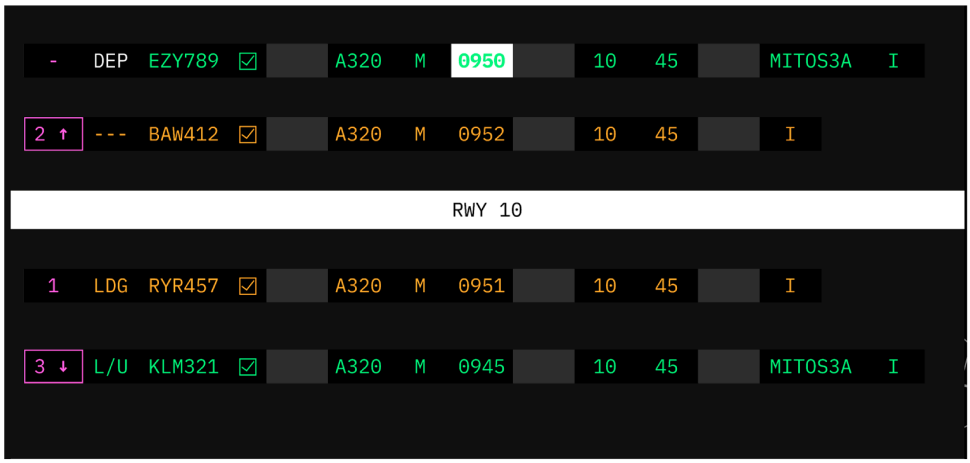


Figure 6. ATCO interface displaying the aircraft sequence and ISA induced changes [32].

Accepting or rejecting a new sequence proposed by ISA is done by tacit interaction, meaning that ATCOs implicitly accept or reject ISA's proposals through their interaction with their HMI and the cockpit crews. For instance, in the example of Figure 6, if the ATCO does not agree with the suggestion by ISA to place BAW before KLM in the sequence, then the ATCO clicks on KLM321 in the HMI and uses voice communication to grant the departure clearance. ISA then recognises the input in the HMI and automatically updates the sequence accordingly (KLM321 back to number 2 and BAW412 back to number 3). If the ATCO accepts the suggested new sequence, then the ATCO simply follows the new sequence by allowing BAW to land first (using HMI input and voice communication of the clearance). If the ATCO does nothing when ISA suggests a new sequence, the blinking of the strips remains for an attraction time of ten seconds, after which the new sequence remains in the flight strips.

4.3. Assessment Compass for ISA Use Case

4.3.1. Level of Automation

ISA endorses a LOA-2 "Resolution support" sociotechnical system, equivalent to EASA AI application level 2A. The ATCO remains in full authority on the sequencing and receives specific new solutions for the sequence planning by ISA, which may be rejected by the ATCO. The ATCO remains in full control of handling the aircraft in the sequence using traditional voice communication with the cockpit crews.

4.3.2. Technology and Human Readiness Levels

ISA was developed from scratch as one of the use cases for human-centred AI-based intelligent assistants in the HAIKU project [16] along a number of design phases [31]. For the purpose of illustrating the holistic assessment framework, we consider the design at the initial and final development stages, D1 and D2, of the concept of operation and prototype. At the initial stage D1, the prototype included initial versions of the sequencing algorithms and the HMI with the explanations. This prototype was evaluated by comparison of results of sequencing exercises by ATCOs with those generated by ISA and by obtaining feedback from ATCOs on the HMI and explanations. The final prototype D2 was developed

using feedback from evaluations of the initial version, resulting in a real-time sequencing tool with an enhanced HMI. The final prototype was evaluated in HITL simulations in a high-fidelity simulator [31], which resulted in a number of recommendations for further improvement of the design.

The initial prototype D1 can be considered to be at TRL 4 and HRL 4, considering the validation and expert evaluation of ISA components. The final prototype D2 completed TRL 5 and HRL 5, and it was evaluated in a high-fidelity simulated environment supporting the option to obtain TRL 6 and HRL 6. Given the feedback to the design following these last simulations, it is argued that TRL 6 and HRL 6 were not yet achieved.

4.3.3. Key Performance Areas

For the purpose of this paper, as an illustration of the holistic framework, the focus in the use case is on a restricted set of seven KPAs: safety, HFs, efficiency, responsibility, accountability, liability, and resilience.

4.3.4. Certification Objectives

As part of the ISA development, several objectives from the EASA guidelines [11] on AI operational explainability and human–AI teaming were identified as relevant, even though these guidelines are only part of the development towards a formal certification framework. These were formulated as specific objectives for ISA, which were evaluated in validation exercises. Table 5 shows some examples of such objectives.

Table 5. Examples of objectives in EASA guidelines and formulated ISA validation objectives.

Objective in EASA Guidelines [11]	ISA Validation Objective [31]
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.	HF-05: ISA must be able to recognise a potential suboptimal sequence generated by user's interaction and suggest a better alternative to the ATCO.
HF-28: The applicant should design the AI-based system to be tolerant to end-user errors.	HF-28: ISA must tolerate and adjust to manual user inputs, by maintaining an ongoing sequence recommendation capability.
EXP-11: The applicant should ensure that the AI-based system presents explanations to the end user in a clear and unambiguous form.	EXP-11: ISA HMI must show explanations related to sequence generation and sequence changes.
EXP-15: The applicant should define the timing when the explainability will be available to the end user taking into account the time criticality of the situation, the needs of the end user, and the operational impact.	EXP-15: Explanations about sequence changes must be available to the user with minimum delay and must permit progressive levels of detail keyed to user needs.

4.4. Holistic Assessment Cycles for ISA Use Case

4.4.1. Identify Objectives, Scope, Criteria

The objectives for the ISA use case stem from the assessment compass and they include assessing the seven indicated KPAs and validating the design with respect to identified certification objectives. The scope of the assessment concerns the ISA sociotechnical system, as described concisely in Section 4.2. Relative performance criteria are used, so as to compare the performance on the KPAs with respect to current operations.

4.4.2. Describe Sociotechnical System

This concerns the objectives, the airport context, the roles and responsibilities of the tower controllers, supervisor, pilots, the procedures, the ISA functioning and its HMI, and the other technical systems for employing ISA in the sequencing and tower operations. For this paper, we refer to the description in Section 4.2, while more details are in [31].

4.4.3. Identify Varying Conditions

There is a broad range of varying conditions that may influence the approach and runway operations in the ISA use case. The identification of such varying conditions is supported by brainstorming, the HAZOP, and the preparation and findings of the HITL simulations. Examples of varying conditions include the following:

- Variability in reaction time of ATCO to resequencing advisory by ISA;
- ATCO is distracted or deals with a contingent complex situation, and does not observe the ISA HMI for a prolonged time;
- Variability in reaction time of pilot to line-up and take-off following ATCO clearance;
- Pilots misunderstand change in sequence and enter the runway without clearance;
- ETA prediction error due to circumstances that were not in the training set, such as wind conditions, aircraft with unusual speed profiles (e.g., military);
- Handover of shifts between ATCOs;
- Temporary R/T blockage affecting communication between ATCO and pilots;
- ISA suddenly breaks down.

4.4.4. Construct Critical Scenarios

Based on the identified varying conditions, various scenarios can be identified that may have a critical impact on the KPAs in the assessment. As an illustration in this paper, we consider the following scenarios involving a departing aircraft DEP and an arriving aircraft ARR:

1. *Distraction of ATCO at proposed sequence change*. Initially, ARR is first in the sequence and DEP second. DEP is waiting for line-up at the runway entrance position. Next, based on the ISA ETA estimate of ARR, it is considered by ISA that there is sufficient time for DEP to depart before ARR arrives at the runway, and it proposes a change in the sequence. The ETA estimate may be (A) correct, but, e.g., ARR was delayed during the approach, or (B) incorrect due to an estimation error—e.g., the data used for the training does not well capture the actual wind conditions. When the sequence change is presented on the HMI, the ATCO is distracted by some event near the runway and does not notice the indication of the sequence change. Following the distraction, the ATCO is somewhat confused, leading to some delay, and then provides a take-off clearance to DEP in line with the sequence on the HMI. Next, due to delay in the provision of the take-off clearance, possibly in combination with an error in the ETA estimate (case B), ARR is already close to the runway, and its crew initiates a go-around to avoid a collision on the runway.
2. *ATCO disagrees with proposed sequence change*. Initially, DEP is first in the sequence and ARR is second. Next, based on the ISA ETA estimate of ARR, it is considered by ISA that there is no longer sufficient time for DEP to depart before ARR arrives, and it proposes a change in the sequence. The ETA estimate may be (A) correct—e.g., ARR has cut a corner during the approach—or (B) incorrect, such that ARR will not arrive earlier. The ATCO does not agree with the proposed sequence change, and thinks there is sufficient time for DEP to remain first in the sequence, and provides a take-off clearance to DEP. Next, in case A, ARR may need to make a go-around to avoid a collision on the runway, since the ATCO was too optimistic, while in case B,

DEP effectively departs on time, since the ATCO correctly judged that there was ample space.

Scenario 1 leads to the initiation of a go-around to avoid a runway collision due to distraction and confusion regarding a sequence change, which may be aggravated by an incorrect ETA estimate. Scenario 2 may lead to a go-around due to the ATCO disagreeing with the ISA proposal, while such disagreement may also support the efficiency of the operation if the ETA estimate is wrong. In both scenarios, there is a potential risk of a collision on the runway if the critical situation is not recognised in time.

4.4.5. Assess KPAs

A range of illustrative assessments for the ISA use case are presented in Appendices A–G. The main insights stemming from these assessments, their relations with the TRL/HRL, and their role in feedback to the design are presented in this section.

Safety Scanning

At an early stage of the design, the Safety Scanning approach (Appendix A) provides a high-level systematic assessment of aspects requiring attention in the regulation framework, safety management, operational safety, and safety architecture. In its application to ISA, various topics were found that required focus in ISA's further development and eventual implementation, such as changes in competence of ATCOs, detection of erroneous advisories by ATCOs, avoiding complacency, and consideration of potentially harmful conditions in the operating environment. Given the broad scope of the assessment method, it provides attention points not only for safety, but also for responsibility, accountability, resilience, and Human Factors. The approach does not support an assessment of the acceptability of a design, but its derived attention points provide useful feedback for the design at an early stage.

User Requirements Analysis

User requirements analysis involves designers working with future operators at an early stage of system design, with the aim of maturing a concept idea into a working system likely to be used and found useful by its intended operators (Appendix B). In the ISA development, this was done by Focus Groups to determine role allocation between ATCOs and ISA. This provided early feedback to the design regarding high-level tasks, responsibilities, and system interaction for ISA and the human operators. Prototyping was then used to develop the concept and initial requirements using static ISA HMI prototypes in normal and non-nominal traffic scenarios.

Task Analysis

There exist various task analysis approaches, which can describe the workflow and interactions of human operators with AI-based systems, including Hierarchical Task Analysis (HTA), Operator Sequence Diagram (OSD), Tabular Task Analysis, and Link Analysis (Appendix C). For the ISA development, an OSD was developed for steps in operational scenarios, including goals, situation awareness states, and decisions/actions by the human operator and AI-based system. This task analysis did not provide feedback to design on its own, but it was set up as a basis for a HAZOP analysis.

Hazard and Operability Analysis (HAZOP)

The aim of HAZOP is to discover potential hazards, operability problems and potential deviations from intended operating conditions. It applies a structured brainstorming session with a multi-disciplinary team of experts using specific guidewords (Appendix D). For the ISA development, a HAZOP session with ATCOs was done, considering scenarios

for normal use, shift handover, and go-around. Feedback to design derived from the HAZOP concerned implications for training, to promote ATCO understanding of ISA and its limitations, and the design of the HMI.

Heuristic Evaluation

A Heuristic Evaluation approach considers the design in the context of guidelines on Human Factors and usability (Appendix E). In support of Heuristic Evaluation for human–AI teaming, a formal review system was developed called HAIQU [34,35] for evaluating the degree to which ISA satisfied a set of HF requirements in eight HF areas (e.g., Human-Centred Design, roles and responsibilities, Sense-Making, etc.). In the ISA's development, an evaluation was carried out during a one-day meeting with the design team and Human Factors people, evaluating ISA against 95 HF design requirements, designating them either as met, not met, or to be developed. The evaluation showed that ISA had generally good compliance with HF requirements, but in a number of cases it led to further design recommendations for ISA, both for usability and for safety.

Human-in-the-Loop Simulation

Human-in-the-Loop (HITL) studies involve a high-fidelity interactive simulation in which participants interact with the system and influence the outcomes of events in the simulation (Appendix F). Common, as well as rare-but-critical, real-world scenarios may be replicated and tested in the simulations. In the ISA development, HITL simulation was done in a full-scope ATC tower training simulator, involving eight sessions with tower ATCOs. The results led to the validation of defined objectives and the identification of a number of user-interaction improvements [31].

Responsibility, Accountability and Liability Analysis

The purpose of a responsibility, accountability and liability analysis is to identify scenarios and to evaluate associated risks for issues regarding accountability, responsibility and liability of stakeholders in operational concepts that include (AI-based) advanced automation (Appendix G). Building on a range of topics concerning policy and objectives, risk management, training and communication, and culture, the responsibilities and accountabilities of stakeholders are evaluated, and the potential liability of the stakeholders is assessed. For the assessment of accountability and responsibility for ISA, a set of questions partly derived from AI-RMF [36] was used in an interview with a company director familiar with ISA. Overall, the results were positive, though there are some areas where ISA needed to consider new approaches to fully accommodate AI-based systems integration into the company's safety governance policies and processes.

4.4.6. Evaluate Combined KPA Results

This step provides a holistic view of the acceptability levels of the KPA results. Figure 7 shows the results of the acceptability levels for the ISA use case at the two design stages, showing improvements and advancement in six of the KPAs. At stage D1, initial assessments were done for safety, resilience, HF, accountability and responsibility, which provided inconclusive and yet uncertain results (illustrated by KPA assessments covering all three colours). At stage D2, additional assessments were done for the KPAs efficiency and liability, and the uncertainty in the acceptability levels was reduced for the earlier assessed KPAs. These illustrative results show that the KPAs responsibility and liability may be at unacceptable levels, whereas the performance for the KPAs safety, resilience, HF, accountability, and possibly responsibility may be acceptable. However, there still is considerable uncertainty, such that each KPA may still only be at a tolerable acceptability level.

Examples of factors that have a positive or negative impact on the acceptability level for a KPA, and factors for which the performance is uncertain, are listed in Table 6. These results provide feedback for improving assessments (see Section 4.4.7) and for changing the design while being aware of the trade-offs between the KPAs (see Section 4.5.1).

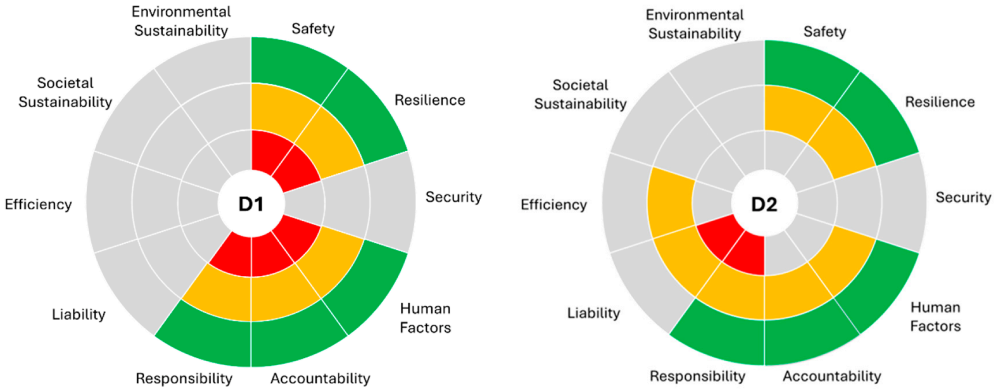


Figure 7. Illustrative results of the holistic view of the acceptability levels for the ISA use case at two design stages, D1 and D2, showing improvement of KPAs (red zones gone for safety, resilience, Human Factors and accountability, plus more advanced information now available on liability and efficiency).

Table 6. Examples of factors with positive, negative and/or uncertain impact on the KPAs for ISA at design stage D2.

KPA	Factor	Impact
Safety	ATCOs found ISA useful as an additional safety net for early detection of potential conflicts.	Positive
	Sequencing errors by ISA may not be detected by ATCO and contribute to a runway incursion.	Negative & Uncertain
Resilience	Early conflict detection supports system resilience	Positive
	ETA module was trained using simulator data and it may not be robust for variations and special conditions in real operations.	Negative & Uncertain
Human Factors	ATCOs appreciated the clarity of the explanations provided by ISA's HMI.	Positive
	Explanations are not always provided in a timely manner.	Negative
	Possible overreliance by ATCO on ISA advisories.	Negative & Uncertain
Accountability	A safety plan exists for the system and the operational unit, with a mixture of reactive and proactive measures. Every time a new threat is identified, it is added, along with new parameters to monitor and track threat evolution.	Positive
	Approaches for risk monitoring and learning from AI-supported operations have not yet been defined.	Negative
Responsibility	AI responsibilities and reporting lines are clearly laid out in the organigram, up to and including the Executive level. In case of an incident/accident involving the AI-based system, the Safety Director will take the lead.	Positive
	Lack of definition of the responsibility of system developers of the ISA resolution support tool.	Negative
	Possible overreliance on ISA, while ATCO remains responsible.	Negative
Liability	In case of an incident following an inappropriate ISA advisory, the ATCO may be liable.	Negative
Efficiency	ATCOs felt ISA was a bit conservative and sometimes inefficient.	Negative
	ISA does not adapt to ATCO's individual working styles.	Negative

4.4.7. Improve Assessment

The illustrative results in Section 4.4.6 show that there exists uncertainty in the assessment results of the KPAs. Such uncertainty may be reduced by improving assessments of the KPAs at the same design stage. For instance, referring to the uncertain results indicated in Table 6, the following improved assessments of ISA may be performed:

- Introduce erroneous sequencing advisories in HITL simulations and to evaluate the effectiveness of detection of such inappropriate advisories by ATCOs. Such extended simulations would improve the assessments of safety and HF.
- Evaluate the performance of the ETA module using data from real operations and special conditions (e.g., abnormal wind, military aircraft). This would improve the assessment of resilience for varying conditions.
- Develop an agent-based model of runway operations with a coupling to ISA advisories based on simulated aircraft positions and variability in ATCO and pilots' performance. Monte Carlo simulation by such a model can provide quantification of runway incursion probability (safety) and the relation with the traffic density [37–39].

4.5. Feedback to Design for ISA Use Case

4.5.1. Issues Requiring System Redesign

Based on the results of the holistic assessment cycle, there are various aspects in the design of ISA at the current TRL/HRL that could be improved to attain a better performance for the KPAs. Examples of such advice to improve the design are listed in Table 7. These were all accepted by the ISA design and development team and the management team.

Table 7. Examples of feedback to design for the ISA use case at design stage D2.

KPA	Possible Design Issue
Safety	• Develop ATCO training for understanding conditions where sequencing advisories may be less reliable • Develop level of certainty indicator in ISA advisories
Resilience	Develop ETA module using data from real operations
Human Factors	• Make sequence changes more visually prominent • Add optional audio signals for critical warnings • Reduce lag in sequence updates and recalculations • Provide explanations earlier to allow more reaction time • Maintain ATCO refresher training and occasional regular operation without ISA advisories (to avoid skill fade)
Accountability	Develop methods for risk monitoring and learning for ISA-supported operations
Responsibility	Define responsibilities of system developers, system maintenance engineers, ATCOs, and management to support high performance of ISA
Liability	Define liability risks in incident types for system developers, system maintenance engineers, ATCOs, and management
Efficiency	• Reduce conservatism in ISA to improve efficiency • Consider individualising sequencing advisories to ATCO style

One interesting aspect of the holistic approach is that it helps consider how changes to one KPA might influence other KPAs, such that there can be a trade-off. For example, improving the efficiency by individualising the sequence advisories could have strong ripple effects on other KPAs: it could increase liability risks of ATCOs and the whole

organisation, and reduce the safety value of ISA. Also, improving efficiency might affect the HF KPA, as this seemingly small change would likely require significant additional study, for example, on how it would affect teamwork and training when different ATCOs are 'calibrating' risk in different ways.

4.5.2. Requirements and Assurance Levels

As explained above, there are various aspects in the design of ISA that would need to be improved at the current maturity level. This means that while the system was tested at TRL/HRL 5/6, it cannot transfer yet to the next TRL/HRL 7, which would imply testing in an operational environment. As such, no requirements are defined at this stage for the implementation in the operational environment.

4.6. Evaluation of the Assessment Methods

Table 8 summarises the eight methods applied to the ISA concept, with an approximate resource estimate. As can be seen, several of the methods are relatively light on resources, though once an operational concept migrates from research (TRL 6) to industrialisation (TRL 7–9), the level of resources spent across KPAs is likely to increase significantly.

Table 8. Overview of methods, types of results and effort for the ISA use case.

Method	Types of Results	Effort
Safety Scanning	Alignment with Safety Fundamentals	1 day/2 people
User Requirements Analysis	Requirements (Human–AI role allocation, HMI)	5 days/10 people
Task Analysis	Human–AI interaction analysis (basis for HAZOP)	2 days/4 people
HAZOP	Hazards and mitigations (Concept, Operational, Validation)	1 day/7 people
Heuristic evaluation	Adherence to HF requirements	1 day/6 people
HITL Simulation	Evaluated requirements New requirements	90 days/8 people [1]
Responsibility, Accountability and Liability Analysis	Alignment with AI Governance Best Practice	1 day/2 people

[1] Preparing a real-time simulation can take anywhere from weeks to months with several people involved, plus those managing the simulation platform. Actual running of such a simulation with licensed pilots/controller may then take a week. Analysis of the results can take from weeks to months by several people.

5. Discussion

5.1. Purpose of the Holistic Assessment Framework

The aim of this paper is to present a viable assessment framework to support the design of AI-based systems against a broad range of KPAs implied by international edicts such as the EU Act on AI. Such a framework aims to ensure that future AI-supported systems remain safe, secure, and beneficial to stakeholders, including the host organisation, operators and society. This encompasses aiming to help protect organisations involved in the development and deployment of such systems from future litigation and reputation damage, or simple loss of investment due to failure of AI adoption or poor performance. While various techniques and nascent AI regulations already exist, they tend to be piecemeal, dealing with one or a few KPAs at a time. The presented framework is a first attempt

at a holistic approach to AI assurance. Since the use case was only at TRL 5–6, any evaluation of the holistic framework and its methods can at best be only partial or provisional. However, given the scale of the potential impact of high-LOA AI on system resilience, safety and performance, and the rapidity of AI's development, it is advisable to begin such holistic analysis early. As AI-based prototypes mature to TRL 7 and beyond, further evaluation can take place and additional or adapted methods can be added as required. This helps to maintain safety and other KPAs on a proactive rather than reactive footing.

The holistic scope builds on the broadly recognised notion that the development and successful introduction of AI in operations cannot solely focus on the performance and reliability of the AI system itself, but needs to consider the overall AI-based sociotechnical system, including interactions with and variability of technical systems, humans, and operating conditions; responsibilities and accountabilities; and connections to society and the environment. While the need for a broad scope has also been acknowledged for more traditional system design without AI, the complexity and potentially poor transparency of AI-based systems and the considerable shifts in accountability, responsibility, and human oversight due to the higher LOAs in operations imply that it is even more important for the design and approval of AI-based systems.

5.2. Aspects Contributing to a Holistic Scope

The holistic scope in the produced framework is achieved (1) by addressing a broad set of ten prime KPAs, (2) by applying the same cyclical approach for the assessment and evaluation of combined results for the KPAs, and (3) by incorporating a toolbox of methods for assessment of the KPAs. The set of KPAs addresses safety, the traditional and current focus in certification; security and especially new cybersecurity threats; Human Factors concerning the broad landscape of human–AI interaction and human oversight; accountability and responsibility at individual and organisational levels for operations and governance; liability for legal responsibility of (problems with) increasingly autonomous systems; resilience for AI system robustness and sustaining operations in all kinds of varying conditions; societal and environmental sustainability for broader assessment of the ethical and environmental impact of AI-based systems; and last but not least, the efficiency of introducing the AI-based system, concerning costs and benefits for stakeholders. Obviously, there are many interrelations between the KPAs, including supporting or counteracting the impact of system aspects on various KPAs. For instance, safety of operations typically requires proper handling of Human Factors and being resilient to disturbances; it can be supported by measures against security attacks, but particular security measures may also negatively impact safety (e.g., hiding safety-relevant information), and there may be a trade-off with efficiency. Various relations exist between accountability, responsibility, liability, Human Factors and safety of AI-based sociotechnical systems, which are at the heart of potential hindrances towards introducing AI-based systems. All in all, the set of KPAs provides a solid foundation for holistic assessment of AI-based systems.

The application of a unified assessment cycle for the various KPAs supports the consideration of all relevant KPAs and criteria for their acceptable performance, the use of a range of similar critical scenarios for their assessment, and the evaluation of the acceptability of KPAs in combination, including possible supporting or counteracting effects of the AI-based sociotechnical system. Defining acceptable performance criteria can be a difficult task. While for some KPAs, such as safety, there may be absolute and possibly quantitative criteria on acceptable versus unacceptable risks, for other KPAs, such criteria are less common, and relative criteria with respect to an acceptable (e.g., current) status may be applied. Also, a KPA may be multidimensional on its own; e.g., HF in the HAIQU approach [34] contains topics like communication, sense-making and teamworking,

which may lead to multiple criteria for acceptable performance. Notwithstanding such difficulties, explicit formulation of criteria for acceptable performance of each KPA provides a needed basis for structured argumentation in a holistic assessment. The use of similar critical scenarios in an assessment cycle supports the evaluation of the scenarios from the perspective of multiple KPAs. Considering, for instance, a scenario in the ISA use case where erroneous advice contributes to an incident, this may be studied from the perspectives of HFs, safety, responsibility and liability. Such combined use of scenarios supports obtaining clear and concrete results, aligned with the various KPAs. The circular diagrams for KPA acceptability levels support a holistic view of all relevant KPAs and of the uncertainty in their assessment along the design stages. Associated tables, evidenced with positive, negative and uncertain factors, provide an explanation of the indicated levels and may indicate relations between KPAs. As such, the diagrams and tables provide feedback to the design team of the AI-based sociotechnical system for improving the overall balance in KPA performance, as well as feedback to the assessment team for reducing uncertainty in the assessment results. While the diagrams thus provide a useful overall view, the argumentation and assessment details that provide their basis are crucial and may require substantial effort.

5.3. Methods in the KPAs Assessment Toolbox

The toolbox of methods in support of the holistic assessment framework covers a broad set of KPAs over considerable ranges of TRL/HRL, and a good part of them are illustrated by the ISA use case. This toolbox is not considered to be fixed and complete. Additional methods may be added, or methods may be adapted to make them more suitable for a wider variety of AI-based systems.

Most methods in the current toolbox can address multiple KPAs (see Table 4). In particular, HF and safety are assessed in combination frequently, which is in line with the important role of human operators for safeguarding operations. Such methods as user requirements analysis, Heuristic Evaluation, HAZOP and HITL simulation all involve confronting future users with the system design, whether statically or in dynamic simulation, and evaluating human interactions, responses, situation awareness and workload in (potentially hazardous) operational scenarios.

Whereas the results of such methods are mostly qualitative, quantitative safety risk assessment may be achieved by agent-based modelling and simulation. In addition to the above operationally focused methods, others such as Safety Scanning and AI-RMF are usefully aimed at safety governance, i.e., how the organisations involved (system developer and operational organisation) manage safety throughout the design-to-deployment-and-operation process. These latter methods also address accountability and responsibility intra- and inter-organisationally, which may be extended to liability risks by a responsibility, accountability and liability analysis. Such methods address questions that may go beyond minimum requirements set by regulation, e.g., considering ethics, impact on jobs and staff wellbeing, etc. This is important not only for governance but also for organisational return on investment, as minimum requirements may determine only that the system is safe to use, not whether it will be profitable and sustainable.

While there exist methods for all KPAs in the toolbox, for KPAs beyond safety and HF, the number of methods and the practical experience in applying them for design, development and approval are limited. And even for applying safety and HF assessment methods, AI characteristics such as poor transparency and adaptability pose new challenges. Security (predominantly cybersecurity) is probably the most urgent, as although some methods exist, application to AI-based systems is relatively recent, and as already noted, AI has a rapid evolution process, suggesting that any security method may be playing ‘catch-up’

given the speed at which AI applications evolve. There are also new threats with AI, such as data poisoning, model obfuscation, indirect prompt injection, and vulnerabilities created by complex data management and operational practices, as noted in a recent standard on AI cybersecurity of the European Telecommunications Standards Institute [40]. It uses a life cycle approach, and its focus on hazard identification and enabling human responsibility (oversight) for AI-based systems means it is in line with the holistic assessment framework of this paper. Additionally, its attention to AI data, training, infrastructure and detailed interactions with users means it is skewed towards higher TRLs and beyond, into monitoring behaviour in operations and end of life.

Given changes in accountability and responsibility due to higher LOAs, the development and application of methods for analysis of liability risks is urgently needed. Whilst liability has mostly been developed incrementally via case law, there is a need to consider potential legal argumentation, evidence, and legal outcomes via testing in legal sandboxes [12]. Otherwise, there is a risk that decades of safety culture and safety learning in domains like aviation could be upended by one or two accidents in AI-based sociotechnical systems and resultant prosecutions of involved operators (pilots, ATCOs).

Also, enhanced methods for societal and environmental sustainability are needed. AI ethics topics, including interpretability, bias, transparency, trust and human oversight, overlap with HF studies, but they also concern societal topics such as equal opportunities, fairness, privacy, job meaningfulness and labour protection. These topics are important for maintaining the overall ecosystem of transport modalities like aviation [41]. The environmental impact in the overall supply chain (including AI data factories) needs to be evaluated and traded off with other KPAs such as societal sustainability and efficiency, which requires sufficiently detailed methods for each KPA.

5.4. Support of Range of TRL/HRL Maturity Levels

The methods in the toolbox generally cover the entire development and deployment process, starting from TRL/HRL 1 through to TRL/HRL 9, with early application methods being superseded by more intensive methods from TRL/HRL 4 onwards. Whilst many methods indicate a large spread of activity over readiness levels, in practice they are more often applied during a narrower phase, or else are utilised periodically as the system develops and matures. This latter approach is fundamental to the TRL/HRL philosophy [19,26], where methods are explicitly linked to the stages and should be applied before moving on to the next phase. Such an episodic approach is normal in risk assessment and is pragmatic given assessment resources. Hence, for example, HAZOP could be applied at TRL/HRL 4 to identify critical issues, then at level 6 to see if those issues have been mitigated and if new issues have arisen, and then at level 7 or 8 when the AI-based system has been coded and integrated into the operational architecture, and procedures and training have been developed for the users. The backbone to this episodic approach would be hazard assessment and HF logs, charting all hazards and HF issues identified and the efficacy of the mitigations put in place to resolve/mitigate them.

While many R&D projects, including the ISA use case in this paper, reach readiness levels up to TRL/HRL 6, a particular concern regards the subsequent scale-up to operations at TRL/HRL 7–9. For conventional hardware or automation, such scaling up follows standardised approaches, and whilst new issues may arise, these can normally be dealt with during the commissioning phase by engineering ingenuity. For AI-based systems, however, new types of variance and unexpected behaviour may occur that are harder to deal with. Additionally, at this stage, human operators must interact with the AI-based system in real operations as opposed to simulations, and there may be fear of job losses, usability frustrations, or AI-based biases or errors leading to safety events. This may lead

to AI adoption failure at a late stage, if such impacts on HF/usability, responsibility and societal sustainability have not been sufficiently considered during the prior development stages. In addition to the adoption of a holistic scope during the development phases, there is a need for methods that focus specifically on TRL/HRL 7–9 phases for transitioning the designed system to deployment, and for ensuring that the performance for the various KPAs is maintained at acceptable levels during operations.

5.5. Types of AI Systems and Applications

The holistic assessment framework and its associated methods have been described generically, meaning that they are expected to be useful for a variety of AI techniques as well as for a variety of application domains. The steps in the assessment cycle can, in large part, be applied alike for a broad range of AI techniques, such as symbolic AI (including rule-based systems), machine learning (including supervised learning, reinforcement learning, deep learning), evolutionary algorithms, and hybrid approaches (e.g., the combination of ML and traditional optimisation techniques in the ISA use case). An AI-based system can be considered as an agent with particular input–output characteristics in the overall sociotechnical system, and the assessment cycle as such does not depend on the particular AI technique that determined its input–output characteristics. The details of the assessment steps can, however, depend on specificities of the applied AI techniques. For instance, knowledge of particular problems or error types of an AI technique (including misclassifications, hallucinations, regression errors and edge cases) can be incorporated as varying conditions (or hazards) in step 3 and included in critical scenarios in step 4 as a basis for the assessment.

Although the inference logic of many AI systems is deterministic (it always provides the same output given a particular input), the behaviour of the overall AI-based system is principally stochastic due to sensor errors in its input. The level of stochasticity may be aggravated if the AI logic includes randomness by itself. In general, a sociotechnical system is stochastic and its random elements can be incorporated in the assessment cycle as varying conditions. For example, ABMS for assessment of traditional and AI-based airborne collision avoidance systems explicitly represents sensor error and pilot response variability for evaluation in Monte Carlo simulation [42]. In a similar way, the impact of stochastic AI-based systems on the overall performance of the sociotechnical system may be assessed and provide feedback on the acceptability of uncertainty levels in system elements.

A prerequisite of the assessment framework is that the AI-based system is non-adaptive, meaning that it does not update the AI logic during operation. The use of an adaptive AI-based system would lead to considerably new types of varying conditions and hazards, and most importantly, it would mean that there is no stable configuration in the system design stages. Application of an adaptive AI-based system in operation (at TRL/HRL 9) would demand advanced ways of risk management to track the impact on KPAs and to maintain acceptable performance. During the design of an adaptive AI-based system, the capability of such risk management to keep the performance at acceptable levels should also be assessed. Clearly, such meta-analysis may be complex and uncertainty-prone.

While the holistic assessment framework and the illustrative ISA use case have been developed in the context of ATM research, it is expected that they can be adapted and applied in other domains as well. Such application may be achieved most seamlessly in other transport domains like maritime and railway transport, where there is overlap with the types of conditions impacting KPAs and associated evaluation methods [43,44].

Application in other domains may be achieved with potential tuning of the relevant KPAs and extension of the toolbox with domain-specific assessment methods.

6. Conclusions, Limitations and Future Research

A schematic overview summarising the relations between the key requirements, the components and KPAs of the holistic assessment framework, the application in the use case, and future research is shown in Figure 8. Next, conclusions, limitations and future research directions are provided.

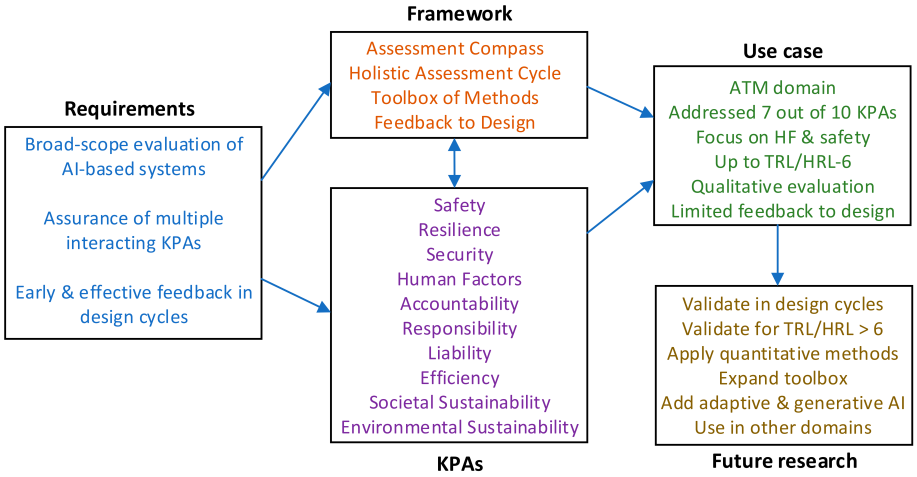


Figure 8. Schematic overview of key requirements, framework, KPAs, use case and future research.

6.1. Conclusions

A viable holistic assessment framework for AI-based sociotechnical systems has been developed and illustrated in depth via results from a practical use case in the ATM domain. The individual methods applied appeared to be fit-for-purpose and resource-efficient, generating usable insights in their respective KPAs and often leading to new design requirements that were adopted by the use case project management. Because the most advanced use case available was limited to TRL/HRL6, quantitative methods could not be applied, and empirical validation was not feasible. Nevertheless, the holistic framework and its associated toolbox represent a practical way forward to evaluating AI-based projects in aviation and other domains across a broad spectrum of KPAs. Such a holistic yet in-depth approach should increase the likelihood of safe, successful and sustainable AI integration into future operational systems.

The key takeaways from the study are the following:

- The holistic assessment framework addresses a broad set of KPAs, providing a unified assessment cycle for their assessment and evaluation via a toolbox of assessment methods. Within the limited confines of the use case application, the KPA framework, assessment cycle and toolbox appear fit for purpose (mainly for 7 of the KPAs), and applicable for a range of non-adaptive AI techniques.
- Regulations and a broad era of HF studies indicate that development and successful introduction of AI in operations requires due consideration of the entire AI-based sociotechnical system. This is preferable to piecemeal or single-KPA approaches that may miss critical or even systemic issues that may ultimately compromise AI adoption.
- The assessment framework supports maturing of the AI-based sociotechnical system to higher technology and human readiness levels by identification of design issues and specification of requirements or assurance levels. It is important to begin assessment of AI-based systems early, allowing identification of significant issues at a time when concept design change is not yet prohibitive.

6.2. Limitations of the Study

Assessment and approval of AI-based sociotechnical systems is a broad and complex topic and various aspects have not been addressed in the current study, in particular because the types of AI-based systems of interest remain at the research stage in aviation at this time. The main limitations include:

- As the use case was only at maturity level TRL/HRL 6, mainly using qualitative assessment approaches, it could not be shown how quantitative methods can be applied, nor how their results can serve as a basis to specify quantitative requirements for designs at the higher maturity levels and for learning assurance of AI-based systems.
- Currently, there are a limited number of methods in the toolbox to assess all KPAs (notably including security and liability risks) and the interrelations between the methods have not been studied. Furthermore, the results and effectiveness of the developed assessment framework have not been compared with those of other assessment approaches.
- As the use case and the development of the framework were both in the aviation domain, additional KPAs and methods may be needed for other application domains.
- The framework only supports the design of non-adaptive AI systems, leaving out systems that could update their behaviour during operation. Furthermore, generative AI systems (e.g., ChatGPT, Claude, etc.) have not been studied, and these are likely to require additional specific approaches.

6.3. Future Research

In line with the above recognised limitations, we pose the following recommendations for future research:

- To validate the effectiveness of feedback to design of the holistic assessment framework and the methods in the toolbox by incorporating them in design cycles of AI-based systems, and to compare them with results of other assessment approaches.
- To apply quantitative methods, such as ABMS, for the specification of quantitative requirements of systems at high TRL, and to validate their use and effectiveness in the framework.
- To extend the toolbox with additional methods and techniques, once favourably tested in the context of AI use cases, providing a range of approaches that can be suitably adapted to assess all relevant KPAs within an application domain. This could include innovative systems engineering analysis approaches, such as System-Theoretic Process Analysis (STPA) [45].
- To investigate the effectiveness of the framework for domains other than aviation, and possibly extend it with other KPAs and methods as appropriate.
- To investigate the implications of generative AI systems for the framework, and possibly extend it with new approaches or adapted techniques, particularly in the areas of HAT, liability, and hazard identification and mitigation.
- To extend the framework for adaptive AI-based systems and to study the implications for risk management of operations at TRL/HRL 9.

Author Contributions: Conceptualisation, S.S. and M.E.; methodology, S.S. and M.E.; validation, S.S., B.K. and M.E.; investigation, S.S., B.K. and M.E.; resources, B.K.; writing—original draft preparation, S.S., B.K. and M.E.; writing—review and editing, S.S., B.K. and M.E.; visualisation, S.S., B.K. and M.E.; project administration, M.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the SESAR 3 Joint Undertaking under grant agreement No 101114762 for the HUCAN project under the European Union's Horizon Europe research and innovation programme.

Institutional Review Board Statement: The illustrative HF results were achieved in accordance with the Declaration of Helsinki, following the procedures documented in report D1.3 "Social, ethical, legal, privacy issues identification and monitoring" (21 December 2022) of the HAIKU project.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data for the illustrative HF results is available in deliverables at https://haikuproject.eu/deliverables/, accessed 30 January 2026.

Acknowledgments: We thank all members of the HUCAN project under the coordination of Paola Lanzi of Deep Blue for their stimulating views on the development of the framework, and the members of the HAIKU project involved in the ISA development for their explanations of the results that were used for the illustration of the framework, in particular Roberto Venditti of Deep Blue and Miguel Villegas Sanchez of Skyway. We thank the anonymous reviewers for their stimulating comments, which helped us to further improve the paper.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ABMS	Agent-Based Modelling and Simulation
AI	Artificial Intelligence
AI-RMF	AI Risk Management Framework
AL	Assurance Level
ANSP	Air Navigation Service Provider
ATCO	Air Traffic Controller
ATM	Air Traffic Management
ConOps	Concept of Operations
CTOT	Calculated Take-Off Time
CLT	Construal Level Theory
EASA	European Union Aviation Safety Agency
ETA	Estimated Time of Arrival
EU	European Union
FHA	Functional Hazard Assessment
FMEA	Failure Modes and Effects Analysis
HAZOP	Hazard and Operability study
HAIQU	Human–AI teaming Questionnaire
HAT	Human–AI Teaming
HF	Human Factors
HFES	Human Factors & Ergonomics Society
HITL	Human-in-the-Loop
HMI	Human–Machine Interface
HSI	Human–System Integration
HRL	Human Readiness Level
HTA	Hierarchical Task Analysis
ISA	Intelligent Sequence Assistant
KPA	Key Performance Area
LOA	Level Of Automation
ML	Machine Learning

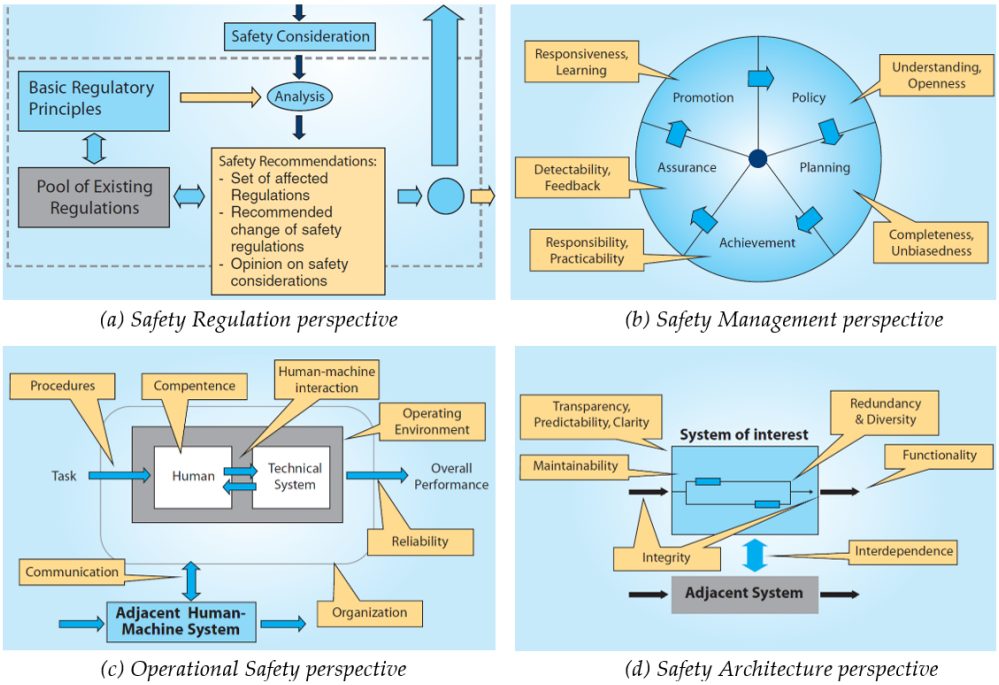
OSD	Operations Sequence Diagram
R&D	Research & Development
SST	Safety Scanning Tool
STPA	System-Theoretic Process Analysis
TRL	Technology Readiness Level
UI/UX	User Interface/User Experience

Appendix A. Safety Scanning

Appendix A.1. Approach

The purpose of Safety Scanning is to scan an air transport operational concept regarding all aspects important for safety. The method is supported by a Safety Scanning Tool (SST), which is a self-assessment tool that 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.

The SST is built on a set of twenty-two 'Safety Fundamentals', which are basic design criteria for safe operations [46]. The Safety Fundamentals are organised into four groups (regulation framework, safety management, operational safety, safety architecture); see Figure A1. The tool guides the user systematically through these Safety Fundamentals by asking for each fundamental between one and five related questions, with 108 questions in total. Example questions are (with the term 'Subject' referring to the operational concept being scanned): 'Does the implementation of the Subject result in changes in procedures?', 'Is the documentation of the Subject clearly understandable and traceable?', 'Does the implementation of the Subject result in changes in staffing requirements?'. 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. These results are analysed by a safety analyst, who draws conclusions on the findings.



(a) Safety Regulation perspective *(b) Safety Management perspective*

(c) Operational Safety perspective *(d) Safety Architecture perspective*

Figure A1. Four perspectives of Safety Scanning with the Safety Fundamentals in yellow.

Appendix A.2. Illustrative Results for ISA Use Case

This section presents the results of application of the SST to ISA. The application was done by one subject matter expert, assisted by a moderator. The duration was less than 2 h, including explanations and a coffee break. The questions were answered in the order as they came in the tool, i.e., regulation framework, safety management, operational safety, and safety architecture.

An overview of the application results of the SST to ISA is given by the bar charts in Figure A2; each bar represents a Safety Fundamental. If the bar for a given fundamental ends in the red area, this topic will require particular attention during the later phases of safety assessment and concept development. If the bar ends in the blue area, this topic will require average attention.

For the regulation framework, there are no significant outliers. Some attention is required on the topic of 'Needs for regulations': while preliminary guidance for AI approval exists [11], associated regulations still need to be formalised.

For safety management, the topic of 'Planning of safety achievement' requires some attention: some aspects of safety achievement are not addressed yet in ISA; for instance, a safety management function has not been identified yet. The implementation of ISA would also affect the 'Safety policy' since AI is new to ANSPs and raises questions for controllers and management. Regarding 'Safety planning', it was noted that due to AI being new, its application is not 'business as usual'; a gradual approach would be needed to develop processes and procedures.

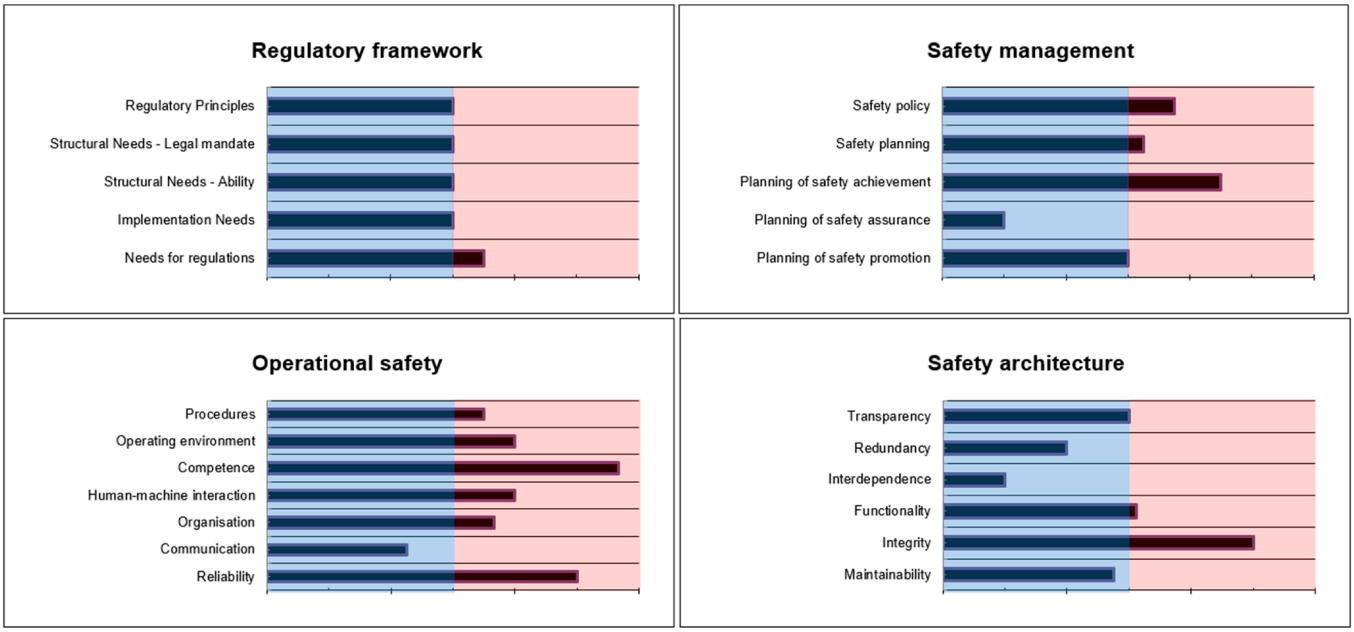


Figure A2. Overview of results of application of Safety Scanning to ISA.

For operational safety, attention is required for nearly every topic:

- The implementation of ISA will result in changes in procedures; however, these are well defined in the ConOps and Manual.
- There are operating environment conditions that may affect ISA, such as declared fuel emergency, hijack, or possibly military activity. Such conditions have not been considered, analysed and addressed yet in ISA. Other such conditions may need to be identified.

- Competence: ISA will change the required competencies of controllers, who will need a degree of AI literacy. Staff need to be shown AI errors during training to see what they look like when it goes wrong.
- Human–machine interaction: Items are added to the HMI of ISA, such as changes in the strip bay. However, it was noted that two simulations have shown a beneficial effect on workload.
- Organisation: ISA will result in organisational changes; e.g., certain tasks can be transferred from the controller to AI functions. Data science expertise may also be necessary within the organisation.
- There may be an impact on the reliability of the current operational environment or existing good practices. The detection of erroneous system behaviour needs to be explored. There is a need to know more about how AI can fail, and about the detectability of such failure. There is also a danger of complacency, e.g., controllers not cross-checking the AI's solution (e.g., because it is usually right).

For safety architecture, attention is required regarding integrity: the possibility of ISA producing harmful outputs, such as AI errors, is a consideration, since that could affect landing and departing aircraft. This will be monitored in real time and will be subject to post-operational analysis. Possibly, there is an effect on functionality as well: trade-offs between safety and other air transport operational objectives have been discussed and no significant trade-offs have been identified. Security has not been evaluated yet.

In conclusion, the application of the Safety Scanning Tool to ISA appeared to be very effective. With very low effort (two people and less than 2 h, plus 1 h for analysis and conclusions), it gave a clear and concise overview of aspects requiring attention during further concept development. This was not limited to safety, but addressed responsibility, accountability, resilience and Human Factors as well.

Appendix B. User Requirements Analysis

Appendix B.1. Approach

User requirements analysis involves designers working with future operators at an early stage of system design, with the aim of maturing a concept idea into a working system likely to be found useful and adopted by its intended operators. Two straightforward user requirements approaches are Focus Groups and Scenario-Based Testing [44]. Both approaches involve exploration of the concept and how it could work and be an improvement on today's system. Focus Groups are a less structured approach, sounding out end users about what works well today and what needs improvement (including so-called 'pain points' or problematic areas), and what they would like to see and be able to do in a future system. Scenario-Based Testing takes the design conversation one step further by exploring the future design in the context of specific scenarios with objectives, resources and constraints. This can be done theoretically (e.g., using a whiteboard, etc.) or with a static or interactive low-fidelity prototype. The output is a more mature design concept and a set of provisional design requirements. This feeds forward into design and more realistic Human-in-the-Loop (HITL) simulations.

Appendix B.2. Illustrative Results for ISA Use Case

Early Focus Groups with a number of tower controllers and supervisors helped to determine requirements for the role allocation between ATCOs and ISA, and ISA's overall functionality (Table A1).

Prototyping was then used to develop the concept and initial requirements, focusing on the usability of the interface given the sometimes 'high tempo' nature of tasks in an air traffic control tower during high-traffic periods, as well as potential constraints such

as poor weather and reduced visibility. Prototyping involved the use of static ISA HMI graphic displays presented to the controllers during normal traffic arrival and departure scenarios, as well as consideration of non-nominal scenarios, including high traffic, poor weather and emergencies. This early prototyping led to requirements on such aspects as speed of interaction, length of text messages, colour coding on the ISA display, etc. Overall, it led to usability properties such that the ISA tool could be smoothly integrated into tower operations, supporting the HF and efficiency KPAs, as well as safety via low error causation (due to the usable HMI).

Table A1. Illustrative requirements from Focus Groups in ISA design.

Agent	Topic	Requirements
ATCO	Goals and high-level tasks	(a) Decision-making for runway usage based on AI-generated sequences. (b) Monitoring and adjusting sequences in response to real-time conditions and system alerts. (c) Handling exceptions and emergencies with priority over automated suggestions.
	AI system interaction	(a) Receive sequence suggestions from ISA. (b) Provide real-time feedback to ISA for continuous learning and adjustment. (c) Utilise HMI to update aircraft status and manage the strip board.
ISA	Scope	(a) Design the system for single-runway airports to enhance efficiency and safety. (b) Focus on both arrival and departure sequencing, taking into account real-time data and operational constraints. (c) Design to always suggest the next 3 aircraft to use the runway (maximum).
	Primary responsibilities	(a) Compute optimal sequences for arrivals and departures using real-time data and machine learning algorithms. (b) Continuously integrate traffic surveillance, meteorological data, and runway status to update sequences. (c) Provide real-time alerts and recommendations through the HMI.
	Human interaction	(a) Receive feedback from controllers to refine and adjust the sequencing algorithms. (b) Provide intuitive and actionable information to controllers for decision support.

Appendix C. Task Analysis

Appendix C.1. Approach

In order to carry out formal analysis of human–AI teaming, a description of how the human operator interacts with the AI-based system is required. This is normally achieved by Human Factor techniques known collectively as task analysis [47], with Hierarchical Task Analysis being the best-known technique. The four most used techniques for charting human–automation interaction are as follows:

- Hierarchical Task Analysis (HTA)—A hierarchical view of the task from its high-level goal to its (skill- and rule-based) operations. A useful over-arching task analysis technique, giving a 'blueprint' of the task.
- Operations Sequence Diagram (OSD)/Timeline Analysis—A linear breakdown of events and operator responses in chronological sequence, showing 'who does what, when and why'. OSD is useful for dynamic or evolving scenarios, especially with multiple crew members and time-pressured scenarios.
- Tabular Task Analysis (TTA)—Highlighting detailed interactions between an operator and their user interface; useful for identifying errors where interaction with equipment, displays and alarms is critical.
- Link Analysis—Noting the sequence and frequency of task-based links between workstations (e.g., actions by crew on a warship bridge), or use of displays in a cockpit, or communication links between members of a distributed team. Link Analysis highlights the most densely used links, communications and teamwork complexity, safety-critical links and potential single points of failure.

The first two scope the whole task and all interactions, with HTA giving a top-down description most useful for developing roles, procedures and training, and OSD being aimed at seeing who does what in an event-driven timeline. The other two techniques can be used to carry out more fine-grained analysis, e.g., on the exact detail of human–computer interactions (TTA), and how tasks are distributed in a control room or amongst different team members who are not co-located.

Task analysis provides the blueprint of human–AI interaction and can then be used by other techniques, including HITL, HAZOP, FMEA, and Heuristic Evaluation, to derive design and performance insights, user requirements and safeguards.

Appendix C.2. Illustrative Results for ISA Use Case

In preparation for a HAZOP, an Operations Sequence Diagram (OSD) was developed. It shows what is happening with the real environment (aircraft, etc.), the AI (ISA) and the ATCOs, in sequential time-steps. It also documents what the ATCO likely believes to be happening, and what the AI computes is happening, as a way to consider potential misalignment between human and AI elements. The OSD was based on discussions with data scientists, ATCOs and the ISA interface designer, and was documented in Excel. An example of the first two columns in the OSD task analysis (for time T0 and time T0 + 10 s) is shown below in Table A2.

Table A2. Extract of Operations Sequence Diagram (input to HAZOP).

Element	State at Time T0	State at Time T0 + 10 s
Actual system state	Aircraft data suggest that an aircraft (BAW412) is going faster than expected and the sequence needs to be reshuffled	Resequencing begins and ISA signals the two involved aircraft (BAW412; KLM321)
Goal	Resequencing of the aircraft's order to relieve ATCO's workload	Making the ATCO aware that a resequencing is happening
Human	ATCO is not aware of the possible changes at the moment	The ATCO becomes aware that the ISA is resequencing
Info sources (non-AI)	Data from aircraft	Data from aircraft
Operator-believed system state	ATCO knows that the assistant will provide support if needed	ATCO knows that the assistant has been activated
AI-believed system state	ATCO is unaware of the aircraft state	ATCO is made aware of the possible changed
AI solution	-	Divert attention to the two aircraft
AI HMI	The AI HMI remains the same at the moment displaying the sequence	Visibility of system status expressed through a string that clearly states "resequencing…". The sequence's cell blinks in flashy magenta colour. On the radar, the speed data of BAW412 is highlighted in magenta.
AI Rationale (XAI)	-	While resequencing is ongoing the explanation is being constructed
H-AI Dialogue	ISA's resequencing will start automatically. The ATCO supervises and makes decisions based on these suggestions.	ISA signals to the ATCO that resequencing is ongoing and ATCO waits.
Authority gradient	ATCO is in control, supervising the state of the system and the sequence	ATCO is in control, supervising the state of the system and the sequence
Decision/Action	-	ATCO is processing the information and waiting for the end of the resequencing
HF Impact: trust, SA; startle/surprise; workload; engagement; competence	No negative impact on Human Factors at this stage in this scenario. The AI is simply monitoring the aircraft, and the ATCO is as yet unaware.	ATCO gains situational awareness about the ongoing event

Appendix D. Hazard and Operability Analysis (HAZOP)

Appendix D.1. Approach

The aim of HAZOP is to discover potential hazards, operability problems and potential deviations from intended operating conditions. It also establishes approximate qualitative likelihoods and consequences 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. A HAZOP study reviews the individual nodes by considering deviations from the intended behaviour, prompted by guidewords, and the consequences of these deviations.

The five elements of a HAZOP study are as follows [48]:

1. A team of multi-disciplinary 'experts', including chairperson, secretary, system designer, engineer, operator/controller, and Human Factor 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*, *None*, or *Not done* 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 process, 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.

HAZOP can be used for analysis of both technical faults and human errors; it covers human operators in the loop. A preliminary HAZOP can be applied to conceptual process descriptions early in the design stage, while a full HAZOP can be done later in the design process.

Appendix D.2. Illustrative Results for ISA Use Case

A HAZOP of ISA was carried out between the first and second validation experiments in the high-fidelity simulator. Six people were involved in the HAZOP, which took place during a single day in the ATC organisation's premises in Madrid:

- A HAZOP chairman (experienced in HAZOP and Human Factors);
- A HAZOP secretary (experienced in Human Factors);
- The ISA Product Owner (also a Tower Supervisor);
- Three tower ATCOs who had participated in the first main simulation and had therefore experienced the use of ISA in simulated tower operations.

Orientation material was sent to the participants prior to the HAZOP, explaining the aims and functioning of HAZOP. At the outset it was stated that the session was confidential, such that no names would be attributed to the findings, and that 'Chatham House Rules' applied; i.e., no one discloses who said what after the meeting.

In terms of the ISA system architecture, only the human–ISA 'interaction layer' was considered, namely what ISA presents on the ATCO HMI, and how the ATCO reacts. If data scientists had been involved, it would also have been possible to consider hazards related to the 'AI technical layer', including data collection, ingestion, computation and presentation stages underpinning core ISA functions such as ETA calculation, sequence calculation and explanation generation. Although the technical layer was not part of the exercise, the HAZOP did consider the possibility of ISA producing erroneous outputs, and whether the ATCO would detect them.

Three scenarios were considered, with most time focused on the first scenario, as it would be the most frequent.

i. Normal use—ISA suggests re-sequencing; ATCO must decide to accept or reject.
ii. Shift Handover—ISA in operation during handover between one ATCO and another.
iii. Go-Around—Major resequencing required due to an aircraft declaring a go-around (aborted landing); ATCO must work with ISA or work alone to re-sequence all aircraft.

For each step in the OSD, the group considered the HAZOP guidewords, and identified hazards, causes, and consequences, as well as existing safeguards that might counter the hazard and new potential safeguards that could be developed to make the system more robust. An extract of the HAZOP table for Scenario 1 is shown below in Table A3.

Table A3. Extract of HAZOP table (transposed for readability).

Element	Case 1	Case 2
Step	[T0] Aircraft data suggest that an aircraft (BAW412) is going faster than expected and the sequence needs to be reshuffled.	
Agent	ISA	ISA
Guideword	*Not done*	*Other than*
HMI	No blinking, No changes	HMI activates
Hazard	Failure to detect catch-up	False alarm of catch-up
Cause(s)	Software malfunctions, System unavailable, Failure of data capture (radar), Algorithm failure, Mode C off	Software malfunctions, Algorithm failure (aircraft performance models inadequate), Data inadequate or biased
Consequence(s)	Unsafe landing sequence; Loss of trust in the tool; ATCO is stressed if it detects late.	Resequencing will occur; ATCO and Pilot loss of trust in the system; Increase in ATC/Pilot communication and workload; Overconfidence in the system, complacency: fail to detect false alarm; Dangerous sequences may occur; Lose valuable time to detect the error in the sequence.
Existing Safeguard(s)	ATCO detects speed-up; Flight crew detects speed-up during internal landing checks.	ATCO can take control of the sequence.

Table A3. *Cont.*

Element	Case 1	Case 2
Recommendation(s)	Need to consider pilot interactions (difference between indicated speed and ground speed); Controllers need to be ready to monitor and ready to react to the system; The controller should not rely on the system 100%; Insert validation scenarios where ISA does not work and check if ATCO can detect AI error.	Consider implementing a short grace period prior to activation of resequencing; Accept or reject function; Safety analysis to determine potential failure modes and false alarms; Training to spot a false alarm and recover from a false alarm. ATCO should not over-trust the system.

The first case considers the guideword *Not done*, interpreted in the HAZOP as ISA fails to recognise the speed-up situation and so does not offer a re-sequencing solution. In this case the ATCO will do their job and perform the resequencing, though perhaps a little late if they have come to rely on ISA, but it is not seen as a safety consequence. The second case considers the guideword *Other than*, interpreted by the HAZOP team as ISA issuing a false alert. This was considered by the HAZOP team to be more serious, as a false alarm would have more impact on trust (both ATCO and pilot), and could result in a more dangerous sequence, perhaps resulting in a go-around. One of the recommendations is to insert such a scenario into either a further validation exercise, or into training at TRL 7+, to see if ATCOs detect such false alarms or not.

Additional HAZOP guideword-led insights included *More Than* (e.g., an escalating traffic picture resulting in a potential cascade of new sequences or overload effect with ISA), *As Well As* (e.g., medical emergency or fuel shortage declared by flight crew), *Not Done* (flight crew not abiding by new sequence), *Less Than* (e.g., fatigue or complacency resulting in reduced performance/situation awareness), and *Too Late* (e.g., network issues affecting ISA performance).

Such insights led to considerations of ATCO training (e.g., knowing what ISA would and would not be aware of, as well as how it can fail), and the HMI (e.g., limiting ISA re-sequences to a maximum of three 'strips', to avoid cascade). ATCOs themselves suggested only intermittent use of ISA, to maintain their skills and avoid becoming over-reliant and complacent about ISA.

At the end of the HAZOP, the ATCOs and Product Owner felt they had 'tested' ISA's vulnerabilities, and despite the various potential failures identified, that they could compensate for the failures and maintain safety. The HAZOP therefore gave them confidence concerning ISA. They also saw ISA as a support to their role, rather than replacing them in any way.

Appendix E. Heuristic Evaluation

Appendix E.1. Approach

A Heuristic Evaluation approach considers the design in the context of guidelines on Human Factors and usability. Since human–AI teaming (HAT) goes beyond traditional Human Factor guidance, various efforts are ongoing to develop new human–AI teaming requirements for AI-based systems. The most important of these in European aviation is the EASA guidance on HAT systems at EASA automation levels 1A to 2B [11]. These guidelines and a host of other existing and new requirements for HAT systems were amalgamated

into a review system [35], which was subsequently encapsulated in a web app: HAIQU (Human–AI Teaming Questionnaire) [34].

The HAIQU app aims to make standards and regulations accessible and user-friendly for design teams looking to integrate AI capabilities into safety critical applications. HAIQU contains 180 requirements in eight Human Factor areas (Human-Centred Design, roles and responsibilities, Sense-Making, communication, teamworking, Error and Failure Management, Competencies and Training, and Organisational Readiness). The app is sensitive to different design maturity levels (TRLs) and EASA's AI autonomy levels, consistent with EASA's classification of human–AI teaming arrangements (1A through 2B). An overview of HAIQU is shown in Figure A3.

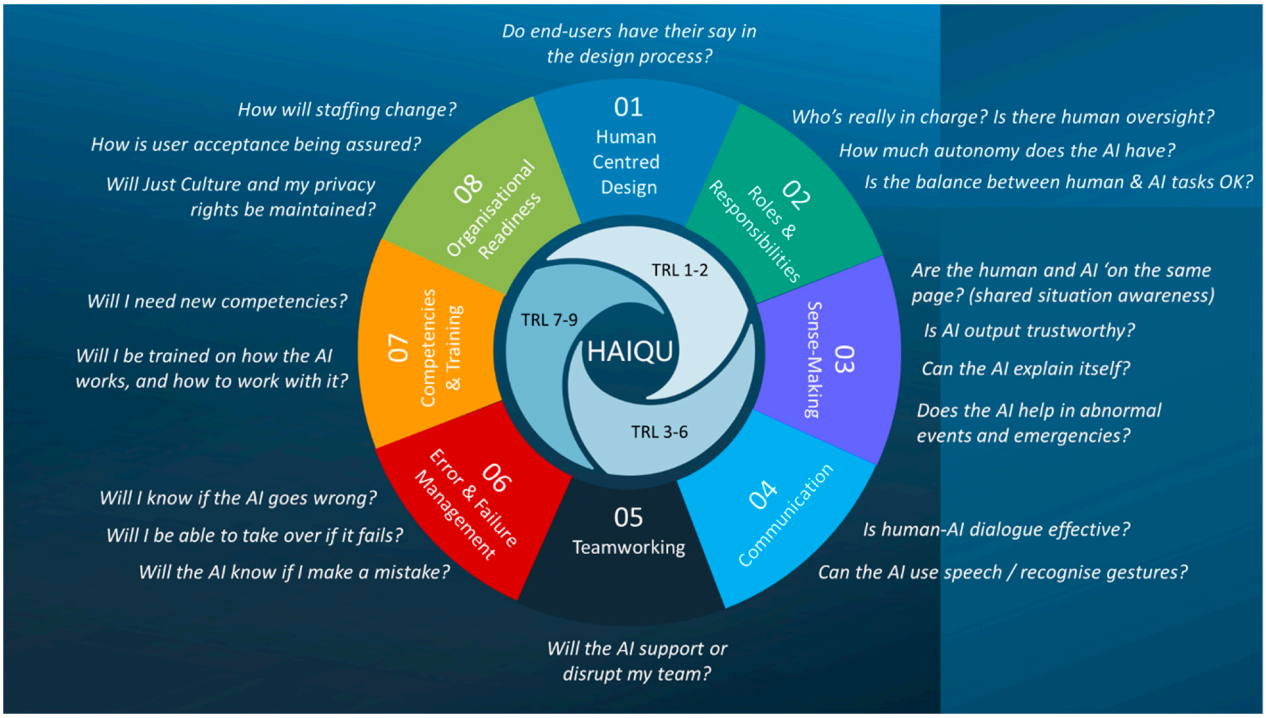


Figure A3. Overview of HAIQU Heuristic Evaluation method [34].

Appendix E.2. Illustrative Results for ISA Use Case

Given ISA's characteristics (EASA Level 2A; TRL 5–6), 95 of the available requirements were applied to the ISA design, following the first validation (D1). The ISA design team responded to these requirements during a one-day session with HF professionals. Figure A4 gives a summary picture of how ISA fared against the HAIQU questions in six of the areas, since for the remaining two areas (Competencies and Training, and Organisational Readiness) it was too early to ask such questions, as they are more appropriate to TRLs 7–9. The red areas in the figure do not necessarily mean that ISA will be unusable, error-prone, or even unsafe; rather, they reflect where these requirements were not seen as essential for the ISA tool, which is constantly supervised by the ATCOs using it. Nevertheless, the 'No' answers would need to be reviewed post-TRL 6 to determine whether there are any risks arising. In several cases, as shown in Table A4, the ISA design team agreed that some of the unsatisfied requirements could be further addressed by augmenting the design of ISA and its explainability capability.

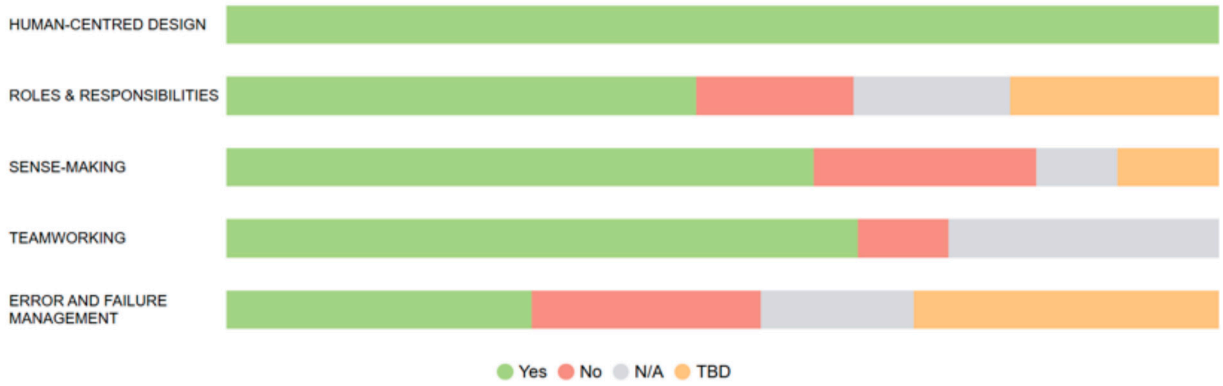


Figure A4. Summary picture of how ISA fared against the HAIQU questions.

Table A4. Extract from HAIQU analysis of ISA.

HAIQU HF Requirement	ISA Design Team Response	Req. Satisfied?	Future Design
Are licensed end-users participating in design exercises such as Focus Groups, Scenario-Based Testing, prototyping and simulation (e.g., ranging from desktop simulation to full scope simulation)? [Human-Centred Design]	Focus Groups and development of scenarios involved licensed ATCOs and a tower ATC supervisor. We had a validation session with licensed ATCOs in the simulator.	Yes	N/A
Does the human have responsibility for overriding the AI, or taking on its role if it fails? [Roles & Responsibilities]	The human always has the final say. If the ATCO notices that the AI is failing, (s)he can override the AI (change the sequence) by issuing orders to aircraft, or can even turn the AI off.	Yes	N/A
Can the human view both data that were used and data that were ignored by the AI, e.g., anomalies or outliers? [Sense-Making]	No, because the deeper explainability levels (CLT 5 & 6) that contain this were not pursued.	No	It could be a good idea to show counterfactuals and anomalies in explainability level CLT4.
Can the AI detect and notify the human if (s)he misunderstands the AI, based on the end-user's actions? [Sense-Making]	If the human does something strange and potentially unsafe, the AI would react by blinking, and display an exclamation mark next to the electronic strip.	Yes	N/A
Does the AI possess the ability to detect human errors or misjudgement and notify them or directly correct them? [Sense-Making]	The AI can only notify them via blinking. If humans make mistakes, the AI will constantly blink.	Yes	N/A
Is the information/decision offered accompanied by uncertainty estimates upon request? [Sense-Making]	No uncertainty estimates at the moment.	No	These could be added into explainability levels CLT 2 & 3.
If the AI is making trade-offs, are these made visible to the human? [Sense-Making]	Considered too cognitively taxing for an ATCO during ops. However, they might be shown at higher levels of explainability.	No	These could be added to future design of CLT levels

Table A4. *Cont.*

HAIQU HF Requirement	ISA Design Team Response	Req. Satisfied?	Future Design
Is the AI's situation representation made accessible to the end user, via visualisation and/or dialogue? [Sense-Making]	Yes, the user always knows the status of the system. There is text showing when ISA is calculating the sequence, but that is also made apparent by the highlighted electronic strips.	Yes	N/A

Appendix F. Human-in-the-Loop Simulation

Appendix F.1. Approach

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 and 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. Although HITL can be resource-intensive, it provides effective insights to designers and operators alike, who can see how it works in realistic scenarios, and serves as an effective validation of the operational concept. HITL can also uncover problems the other techniques may have missed, and lead to more effective design ideas.

Appendix F.2. Illustrative Results for ISA Use Case

In the spring of 2025, ISA was the subject of a high-fidelity real-time simulation involving eight sessions with individual tower controllers (five male, three female; age 23–48 years [mean = 37]; operational experience 3–17 years [mean = 8.7]) [31]. The simulator used was a full-scope training simulator representing the ATC Tower in Alicante, Spain's busiest single-runway airport. ISA was available via a laptop for the tower controller.

Each participant was trained in using ISA and then had the option of using it during the long scenario, which included a period of high traffic (40 aircraft arrivals and departures per hour). Following the session, the controller was debriefed via a semi-structured interview, answering questions on ISA's utility, accuracy and overall performance. The interview focused on key HF issues according to the EASA guidance on human–AI teaming [11]. Examples of the validation objectives and validation results are shown in Table A5.

Additionally, a number of user-interaction improvements were identified by the controllers:

- Make sequence changes more visually prominent;
- Add audio signals for critical warnings;
- Shorten mouseover explanations or replace text with symbols/icons;
- Reduce default detail in advanced explanation levels (CLT3/CLT4);
- Prioritise concise explanations during high-traffic scenarios;
- Reduce lag in sequence updates and recalculations;
- Provide explanations earlier to allow more reaction time;
- Clarify the meaning of blinking/exclamation mark alerts;
- Adjust conservatism in ISA to better balance safety and efficiency.

In summary, ISA was validated effectively by the HITL approach and was well received by ATCOs. Its added value in terms of safety and its explainability features were

highlighted as the most appreciated and operationally useful components. The ATCOs found ISA particularly useful as an additional safety net, due to the visual alerts for potential conflicts/unsafe situations. However, ISA as currently designed is not so useful for improving efficiency, as each controller has their own working style (some more conservative, some less so). To improve efficiency, ISA would need to be able to adapt to (and learn from) each controller's way of working.

To conclude, the HITL results bode well for the safety KPA, since it is seen as 'adding safety value', though detection of ISA errors (as recommended by the HAZOP) has yet to be tested. For the Human Factors KPA, the results are mostly positive, although some desired changes were identified. For the efficiency KPA, ISA is currently neutral. To make it positive would require a significant shift in ISA's capability and AI design basis.

Table A5. Examples of results from HITL simulations for ISA [31].

Validation Objective (Related to [11])	Validation Result
HF-05: ISA must be able to recognise a potential suboptimal sequence generated by the user's interaction and suggest a better alternative to the ATCO	Some ATCOs felt ISA adapted well to situations where pilots did not comply with instructions, while others did not provide specific feedback. The ability of ISA to recognise suboptimal sequences was seen as beneficial, but there were instances where ATCOs felt it could be more efficient.
HF-28: ISA must tolerate and adjust to user manual inputs, by maintaining an ongoing sequence recommendation capability.	ISA was able to adjust to the user's inputs and change the sequence accordingly. All ATCOs agreed that ISA was able to adapt appropriately to their decisions.
EXP-11: ISA HMI must show explanations related to sequence generation and sequence changes.	ATCOs appreciated the clarity of the explanations provided by ISA's HMI. The feedback was generally positive. The clarity of explanations was well-received, and ATCOs found the HMI intuitive and helpful.
EXP-15: Explanations about sequence changes must be available to the user at minimum delay and permit progressive levels of detail keyed to user needs.	Timing was generally appropriate, but some ATCOs suggested more time for reaction. The timing of explanations was mostly appropriate, but there were suggestions for slight adjustments to allow more reaction time.

Appendix G. Accountability, Responsibility and Liability Analysis

Appendix G.1. Approach

The purpose of an accountability, responsibility and liability analysis is to identify scenarios and to evaluate associated risks for issues regarding accountability, responsibility and liability of stakeholders in operational concepts that include (AI-based) advanced automation. More generally, this area concerns ensuring that proper governance is in place inside the organisation intending to use AI in its operational systems. In line with the holistic assessment cycle (Section 3.3), it includes the following steps.

1. *Scope and 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 liability analysis. The objectives concern a description of the types of results that need to be achieved. This can, for instance, be the identification of conditions where stakeholders may be responsible/accountable/liable, it can be the assessment of changes in responsibility/accountability/liability, or it can be the assessment of associated risks (outcomes and likelihoods) in these three areas.

2. *Describe sociotechnical system and accountability.* In this step, the sociotechnical system is described, including the accountability structure of the organisation, i.e., who is accountable for what to whom. It includes the identification of the specific provisions the organisation has put in place to be able to demonstrate its accountability, responsibility and liability obligations, commensurate with the risks associated with the new AI-based system. This may, for instance, concern senior management and their role during system development, validation, and testing, as well as training, maintenance, and monitoring the AI-based system performance once deployed.
3. *Identify critical scenarios.* The purpose of this step is 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, and 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, accountability 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 (e.g., safety, security), where severities of operational outcomes and likelihoods of attaining such severity levels have been assessed. For each of these cases, the accountability and responsibility of directly involved operators as well as of stakeholders are assessed for topics listed in Table A6. Key questions in such stress tests concern which stakeholders are involved, directly and indirectly, what their responsibilities are in the scenarios, and which stakeholders in the organisation can be held accountable. Based on the assessed accountabilities and 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 accountability, responsibility and liability assessment provide a structured oversight over responsibilities and possible liability risks for stakeholders in advanced automated operations. It may be concluded that some responsibilities and/or accountabilities are not well defined at this (albeit early) stage, and that additional operational procedures or organisational changes are advised. It may be concluded that liability risks are potentially too high and that changes to the use of the advanced automation or the organisational embedding may be advisable.

Table A6. Topics in an accountability, responsibility and liability analysis.

Topic		Future Design
Policy & Objectives	Clearly accountable roles	How are the roles defined in the corporate organigram, and are all roles filled? This includes designated managers or responsible officers for design, development, testing, deployment, operation and maintenance of AI-based systems.

Table A6. *Cont.*

Topic		Future Design
Policy & Objectives	Maintaining legal/ regulatory oversight	How is the organisation kept up to date with regulatory and legal changes concerning AI usage?
	Data provenance, quality and legality	How is data provenance determined, evaluated for quality, and assessed in relation to legal guidelines, and how is this process documented?
	Privacy protection	How is ATCO and other end-user privacy protected in terms of any data collected or inferable from use of the AI-based system?
Risk Management	Risk monitoring throughout life cycle	What risk analysis activities are carried out through the design, validation and deployment stages?
	Learning from operational experience	How is operational feedback solicited and analysed, what lessons were learned, and what changes were made as a result?
	Risk management	What type and level of independent analysis is used to confirm that risks and other impacts from the AI-based system are adequately identified and managed/mitigated?
	Monitoring of AI's parameters	How is the AI monitored for change or drift away from its original parameters?
	Vulnerable group protection	Are any vulnerable groups identified and, if so, protected from adverse impacts of usage of the AI-based system?
Training & Communication	End-user training	How are end users trained to use the AI, including recognition of AI error and how to recover from it, their liability in case of an adverse event, and how is their training updated as the AI evolves?
	Stakeholder engagement on AI governance	Which non-governmental organisations, public sector, policymakers, or multi-stakeholder initiatives does the organisation collaborate with in AI governance and staff wellbeing best practices?
	Protocols in case of an incident involving the AI-based system	What procedures are in place in the event of an incident, e.g., ATCO relief/debrief following the incident, storage/analysis of AI logs, service continuity with/without the AI, implementation of just culture and investigation procedures, etc.
Culture	Just culture and liability	What does the organisation's just culture policy say about AI usage, and are staff protections in case of honest mistakes still effective?

Table A6. *Cont.*

Topic		Future Design
Culture	Safety Culture	How is a strong safety culture maintained (such that ATCOs still 'own' safety rather than relying on the AI), including a healthy reporting culture where the AI-based system is concerned?
	AI alignment with organisational values	How is it ensured (e.g., via AI transparency and interpretability, as well as ATCO feedback) that the AI-based system is operating in line with organisational values?

Appendix G.2. Illustrative Results for ISA Use Case

Appendix G.2.1. Accountability and Responsibility

For the ISA study, a director from the associated air traffic organisation co-developing the AI-based tool was interviewed for half a day by a Human Factors and safety expert. The prepared interview questions were developed in part from [36], along with other questions derived from experience concerning how air traffic organisations manage safety both proactively and in the aftermath of high-profile incidents and accidents. A total of 40 questions were asked, dealing mainly with accountability and responsibility. An overview of the question set and the areas it addresses is given in Table A6.

A key hypothetical future scenario considered in relation to ISA was that of, some time after deployment, a safety-related incident (a 'close call' runway incursion) occurring that involved the use of the ISA tool, and the event being subsequently reported in local and national social media. Table A7 gives examples from the interview on accountability and responsibility. It should be noted that the interview was prospective and experimental, given that the ANSP does not currently employ AI-based systems. The aim was to test the questionnaire process and elicit provisional answers. The table shows an extract of the interview via answers for twelve of the forty questions in the full questionnaire, which took approximately two hours to administer.

The structured questionnaire approach, interviewing a company director who was also fully aware of ISA, enabled insights into how the organisation would exercise governance of the integration of AI-based systems into their operations. The scenario of an incident having occurred involving the possibility of both AI error and ATCO overreliance on the AI helped the director consider how they would demonstrate that they had put in place the requisite responsibilities and had executed their accountabilities to a sufficient degree. The result was a rich picture of their existing and provisional future governance mechanisms for AI-based systems, as well as some areas they had not yet thought of. The director found the exercise very useful in gaining 'forward vision' into how their current governance and safety management approaches might need to evolve to include integrating AI-based systems. For accountability and responsibility in particular, it was noted that the organisation already had a number of strong policies and processes in place, and others that would accommodate AI considerations if ISA were to be developed beyond TRL 6. Asked separately at the end of the interview to identify the most pressing need for research in this area, the director suggested ways to prevent ATCO 'skill fade' (meaning that the ATCOs lose a key skill or expertise such that if the AI fails they cannot efficiently compensate by reverting to former practices) once AI systems are in place.

In terms of the answers' import for the two KPAs, they were judged to be positive for both accountability and responsibility, though more would need to be done if ISA were to be

further developed and deployed. If in the near future the approach were to be utilised 'for real', various sources of evidence would be elicited, e.g., organigrams and accountability descriptions, SMS documentation related to AI, formal notes of meetings/workshops with end users, HF documentation, crisis plans, competency descriptions, AI-related KPI tracking records, just culture policy, risk assessment results, simulation testing schedules, etc., to substantiate and further inform the answers given.

Table A7. Extract of questions and responses in interview on accountability and responsibility.

Accountability Questions (Extract)	Organisation Responses
How is it ensured that legal and regulatory requirements of AI-based systems are understood and managed at Executive board level.	Annually, we all sign our accountabilities, so we are reminded of them, and if there are updates, e.g., new AI systems, we are made aware of it and who is accountable for what. There will be something in the SMS, on what is AI and how can it affect us, translated into accountabilities (who is accountable for what). AI is game-changing, but in the end the overall process of integrating it into operations can be quite similar to other new technologies.
What efforts have been made to keep the AI system transparent to end users (e.g., via explainability and interpretability)?	Involvement of end users from the beginning and a focus on training, more on the operational and safety part, not so much the technical AI part. What is the purpose of the AI, what can you do and not do, etc.
Are AI performance metrics and error rates monitored and analysed to ensure no performance degradation is occurring, nor any emergent problematic AI behaviour?	100%! There will be constant monitoring and supervision of performance KPIs that will allow us to react promptly.
How are emergent risks identified, tracked, assessed and mitigated?	We have a safety plan for each unit and one for HQ, with a mixture of reactive and proactive measures. Every time there is a new threat we add it and add new parameters to monitor and track threat evolution.
What measures are taken when new AI model training data are integrated into the AI system, i.e., to ensure continued safety and performance robustness?	We would have a checklist or benchmark that has to be passed every time there is a new data-set input. This would involve simulation to ensure the new data does not lead to data corruption or anomalies.
How is feedback from both end users (ATCOs) and third-party users (i.e., pilots, airlines, Network Manager) solicited and analysed for learning purposes?	Once a year we collect feedback from ATCOs, and we send all third parties a survey to rate our service and give feedback on anything they want. Perhaps can increase frequency at the beginning of introduction of AI. We also have many meetings with pilots, ramp agents, and get a lot of feedback from these local meetings.

Table A7. *Cont.*

Responsibility Questions (Extract)	Organisation Responses
Are AI responsibilities and reporting lines clearly laid out in the organigram, up to and including the Executive level?	Yes.
What new skill sets and competencies were taken on board to deal with AI-based systems (e.g., data science)?	Data science expertise is taken on for sure.
Who will take the lead in case of an ATC safety-related incident involving the AI-based system?	Safety director.
What procedures and protocols are in place in case of an incident involving the AI?	Crisis Plan will be invoked whenever there is an incident where there might be an impact on the reputation of the company. ATCO will be relieved, suspension of AI while system is checked, AI log storage, etc. Will be fast restoring the operation but probably at lower capacity and possibly under degraded conditions.
Do you align AI initiatives with organisational strategy and values?	Yes, for sure. Cannot have AI that goes against the pillars and principles of the organisation. No short-cuts.
How is data privacy of end users ensured?	There will be a strong company policy on this aspect.

Appendix G.2.2. Liability Issue Considerations

The legal and liability aspects of future aviation AI systems in operation are at present unknown; therefore, the following is the result of a speculative analysis identifying salient legal and liability aspects likely to emerge. The primary issue (as raised above in the discussion of responsibility and accountability) is what would happen in the case of an accident occurring involving human–AI teaming, i.e., a human (in this case an ATCO) utilising an AI-based system such as ISA. There are four obvious cases:

(i) The AI committed a serious error that the ATCO could not override.
(ii) The ATCO was not using the AI and committed a serious error.
(iii) The AI made a recommendation that was wrong, but the ATCO decided to 'trust' the AI.
(iv) The AI made a safe(r) recommendation and the ATCO decided to ignore it, compromising safety.

It is the last two that are of most interest, since jurisprudence for (i) and (ii) should be more straightforward. At a surface level, for case (iii), since ISA is Level 2A, the ATCO is still fully in command and so could be considered liable. Similarly, for case (iv), again the ATCO has made a conscious decision, and so could be liable.

However, reality could be more nuanced. If the AI is routinely very trustworthy and works well with a high reliability, it is only human to begin to rely on it, particularly, as is likely the case, if higher traffic levels are assumed in operations due to the improved HAT efficacy. In this case, the charge of 'complacency' on behalf of the ATCO is too simplistic and would be opposed by the *just culture* maxim that if another controller would likely have behaved exactly the same, then it is not fair to blame the controller, because the system is inducing a particular behaviour on their part.

ATM is an industry that spearheads just culture, and the ATM industry keeps itself and its constituent national members apprised of such matters. ATM is wary of controllers being prosecuted when doing their job to the best of their abilities, both to protect them, and to ensure that the high degree of safety culture embedded in the industry does not suffer. Safety culture is emblematic in European ATM organisations [49], yet could be adversely affected (for pilots as well as ATCOs) by legal issues that could arise during the early introduction of AI into operations, which is the most likely time incidents and accidents could happen [21].

Additional considerations relate to the user interface design, including explainability and training. If the interface design is 'clumsy' (difficult to use) or seen as adding extra 'noise' into the system (thereby detracting from situation awareness, or overly focusing it on particular aspects), then this must become part of the liability consideration. Similarly, if the explainability is not clear or does not fit all situations, or is indeed sometimes misleading, then this will impact human behaviour and judgement. This is compounded if training does not include being able to detect when the AI is in error or should be ignored/overridden. Such training would need to occur in realistic simulations in order to be effective, and repeated regularly (refresher training), especially as the AI-based tool is updated. ATCO experience and seniority would also play a part, since more experienced ATCOs would have more to base their judgement on when it comes to deciding to ignore/override the AI tool. The problem would be that as the tool is used more regularly, new ATCOs would be unlikely to develop such experience in the first place, and so their 'capture' by the tool is more likely (sometimes referred to as 'blind reliance'). Incidentally, this is one reason why the ATCOs during the HAZOP stated that they would sometimes turn the AI off to avoid skills degradation. 'Skill-fade' was also the director's top concern with integrating an AI-based tool into operations.

A specific design option that has liability implications is tailoring the ISA to individual ATCOs' preferences in terms of landing rate, etc., since some ATCOs are more conservative than others. If an accident happened where the ATCO had 'tuned' ISA to be less conservative (less margin for error, but higher productivity), this could well weigh into liability considerations both for the ATCO concerned and for the organisation allowing the practice, even if safeguarded by regular competency checks when using the tool.

The above has attempted to 'unpack' some of the key liability aspects likely to arise once AI-based systems are operational. The best course of action, however, rather than 'wait and see', would be to run legal test cases in a 'legal sandbox', to more clearly refine arguments and considerations before an actual accident occurs [21].

References

1. European Union. *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: Brussels, Belgium, 2024.
2. European Union. *Regulation (EU) 2018/1139 on Common Rules in the Field of Civil Aviation*; European Union: Brussels, Belgium, 2025.
3. High-Level Expert Group on AI. *Ethics Guidelines for Trustworthy AI*; European Commission: Brussels, Belgium, 2019.
4. National Research Council. *Autonomy Research for Civil Aviation: Toward a New Era of Flight*; The National Academies Press: Washington, DC, USA, 2014. [CrossRef] [PubMed]
5. NASEM. *Machine Learning for Safety-Critical Applications: Opportunities, Challenges, and a Research Agenda*; National Academies of Sciences, Engineering, and Medicine: Washington, DC, USA, 2025. [CrossRef] [PubMed]
6. NASEM. *Human-AI Teaming: State of the Art and Research Needs*; National Academies of Sciences, Engineering, and Medicine: Washington, DC, USA, 2022. [CrossRef] [PubMed]
7. Bainbridge, L. Ironies of automation. *Automatica* **1983**, *19*, 775–779. [CrossRef]
8. Endsley, M.R. Ironies of artificial intelligence. *Ergonomics* **2023**, *66*, 1656–1668. [CrossRef] [PubMed]
9. NASEM. *Test and Evaluation Challenges in Artificial Intelligence-Enabled Systems for the Department of the Air Force*; National Academies of Sciences, Engineering, and Medicine: Washington, DC, USA, 2023. [CrossRef] [PubMed]

10. EASA. *Artifical Intelligence Roadmap 2.0: Human-Centric Approach to AI in Aviation*; European Union Aviation Safety Agency: Cologne, Germany, 2023.
11. EASA. *EASA Concept Paper: Guidance for Level 1&2 Machine Learning Applications, Issue 02*; European Union Aviation Safety Agency: Cologne, Germany, 2024.
12. Díaz-Rodríguez, N.; Del Ser, J.; Coeckelbergh, M.; López de Prado, M.; Herrera-Viedma, E.; Herrera, F. Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Inf. Fusion* **2023**, *99*, 101896. [CrossRef]
13. EUROCAE. *Artificial Intelligence in Aeronautical Systems: Statement of Concerns*; ER-022; EUROCAE: Saint-Denis, France, 2021.
14. HUCAN. *Holistic Approach to Approval and Certification of Automated Systems*; D4.4, Edition 01.00; SESAR Joint Undertaking: Brussels, Belgium, 2025.
15. HUCAN. HUCAN: Holistic Unified Certification Approach for Novel Systems Based on Advanced Automation. Available online: https://research.dblue.it/hucan/ (accessed on 30 January 2026).
16. HAIKU. HAIKU: Human AI Teaming Knowledge and Understanding for Aviation Safety. Available online: https://haikuproject.eu/ (accessed on 30 January 2026).
17. Hollnagel, E.; Woods, D.D.; Leveson, N. *Resilience Engineering: Concepts and Precepts*; Ashgate: Aldershot, UK, 2006.
18. Bergström, J.; van Winsen, R.; Henriqson, E. On the rationale of resilience in the domain of safety: A literature review. *Reliab. Eng. Syst. Saf.* **2015**, *141*, 131–141. [CrossRef]
19. HFES. *Human Readiness Level Scale in the System Development Process*; ANSI/HFES 400-2021; Human Factors and Ergonomics Society: Washington, DC, USA, 2021.
20. Thompson, D.F. Designing Responsibility: The Problem of Many Hands in Complex Organizations. In *Designing in Ethics*; van den Hoven, J., Miller, S., Pogge, T., Eds.; Oxford University Press: Oxford, UK, 2017; pp. 32–56. [CrossRef]
21. Kirwan, B. The Impact of Artificial Intelligence on Future Aviation Safety Culture. *Future Transp.* **2024**, *4*, 349–379. [CrossRef]
22. SESAR Joint Undertaking. *SESAR Environment Assessment Process*, 5th ed.; SESAR Joint Undertaking: Brussels, Belgium, 2024.
23. Vagia, M.; Transeth, A.A.; Fjerdingen, S.A. A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? *Appl. Ergon.* **2016**, *53*, 190–202. [CrossRef] [PubMed]
24. SESAR Joint Undertaking. *European ATM Master Plan: 2025 Edition*; SESAR Joint Undertaking: Brussels, Belgium, 2024.
25. SESAR Joint Undertaking. *European ATM Master Plan: Digitalising Europe's Aviation Infrastructure, 2020 Edition*; SESAR Joint Undertaking: Brussels, Belgium, 2020.
26. Mankins, J.C. Technology readiness assessments: A retrospective. *Acta Astronaut.* **2009**, *65*, 1216–1223. [CrossRef]
27. SESAR Joint Undertaking. *Maturity Gate Guidance*, 2nd ed.; SESAR Joint Undertaking: Brussels, Belgium, 2025.
28. Blom, H.A.P.; Stroeve, S.H.; De Jong, H.H. Safety risk assessment by Monte Carlo simulation of complex safety critical operations. In *Developments in Risk-Based Approaches to Safety: Proceedings of the Fourteenth Safety-Critical Systems Symposium, Bristol, UK, 7–9 February 2006*; Redmill, F., Anderson, T., Eds.; Springer: London, UK, 2006.
29. Stroeve, S.H.; Blom, H.A.P.; Bakker, G.J. Systemic accident risk assessment in air traffic by Monte Carlo simulation. *Saf. Sci.* **2009**, *47*, 238–249. [CrossRef]
30. JARVIS. JARVIS: Just A Rather Very Intelligent System. Available online: https://research.dblue.it/jarvis/ (accessed on 30 January 2026).
31. Westin, C. *Second Validation Report (VAL2) and Demonstrator (DEM2), D6.4*; HAIKU Project: Rome, Italy, 2025.
32. HAIKU. *High Fidelity Prototypes, D4.5, Version 2.0*; HAIKU Project: Rome, Italy, 2025.
33. Venditti, R.; Minaskan, N.; Biliri, E.; Villegas, M.; Kirwan, B.; Westin, C.; Basjuka, J.; Pozzi, S. Construal Level Theory (CLT) for Designing Explanation Interfaces in Operational Contexts. In *Human Interaction and Emerging Technologies (IHIET-AI 2025): Artificial Intelligence and Future Applications*; Ahram, T., Lopez Arquillos, A., Gandarias, J., Morales Casas, A., Eds.; AHFE International: Malaga, Spain, 2025. [CrossRef]
34. Venditti, R.; Pozzi, S.; Frau, G.; Salam, R.; Imbert, J.; Duchevet, A.; Kirwan, B. HAIQU—A Human Factors Requirements App for Human-AI Teams in Aviation. In Proceedings of the Contemporary Ergonomics and Human Factors, Burton upon Trent, UK, 28–30 April 2025.
35. Kirwan, B. Human Factors Requirements for Human-AI Teaming in Aviation. *Future Transp.* **2025**, *5*, 42. [CrossRef]
36. NIST. *Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023.
37. Stroeve, S.H.; Som, P.; van Doorn, B.A.; Bakker, G.J. Strengthening air traffic safety management by moving from outcome-based towards risk-based evaluation of runway incursions. *Reliab. Eng. Syst. Saf.* **2016**, *147*, 93–108. [CrossRef]
38. Stroeve, S.H.; Van Doorn, B.A.; Bakker, G.J. Safety assessment of a future taxi into position and hold operation by agent-based dynamic risk modelling. *J. Aerosp. Oper.* **2012**, *1*, 107–127. [CrossRef]

39. Fota, N.; Everdij, M.; Stroeve, S.H.; Krakenes, T.; Herrera, I.A.; Quinones, J.; Contarino, T.; Manzo, A. Using dynamic risk modelling in Single European Sky Air Traffic Management Research (SESAR). In Proceedings of the European Safety and Reliability Conference ESREL, Wroclaw, Poland, 14–18 September 2014.
40. ETSI. *Securing Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems*; European Telecommunications Standards Institute: Sophia Antipolis Cedex, France, 2025.
41. EASA. *Ethics for Artificial Intelligence in Aviation: Aviation Professionals Survey Results 2024/2025*; European Union Aviation Safety Agency: Cologne, Germany, 2025. Available online: https://www.easa.europa.eu/en/document-library/general-publications/ethics-artificial-intelligence-aviation (accessed on 30 January 2026).
42. Stroeve, S. Stochastic dynamic agent-based modelling for evaluation of ACAS and DAA systems. In Proceedings of the 2025 Integrated Communications, Navigation and Surveillance Conference (ICNS), Brussels, Belgium, 8–10 April 2025. [CrossRef]
43. Stroeve, S.; Kirwan, B.; Turan, O.; Kurt, R.E.; van Doorn, B.; Save, L.; Jonk, P.; Navas de Maya, B.; Kilner, A.; Verhoeven, R.; et al. SHIELD Human Factors Taxonomy and Database for Learning from Aviation and Maritime Safety Occurrences. *Safety* **2023**, *9*, 14. [CrossRef]
44. SAFEMODE Project. Human Assurance Toolkit. Available online: https://safemodeproject.eu/EhuridHumanAssuranceToolkit.aspx (accessed on 5 February 2026).
45. Leveson, N.G.; Thomas, J.P. *STPA Handbook*; MIT: Boston, MA, USA, 2018. Available online: https://psas.scripts.mit.edu/home/get_file.php?name=STPA_Handbook.pdf (accessed on 30 January 2026).
46. Everdij, M.H.C.; Korteweg, H.; Penny, J.; Straeter, O.; Longhurst, T. *Development of a Set of Questions for the Safety Scanning Tool, NLR-CR-2009-628*; Netherlands Aerospace Centre NLR: Amsterdam, The Netherlands, 2010.
47. Kirwan, B.; Ainsworth, L. *A Guide to Task Analysis*; CRC Press: London, UK, 1992. [CrossRef]
48. Storey, N. *Safety-Critical Computer Systems*; Addison-Wesley: Harlow, UK, 1996.
49. Kirwan, B.; Shorrock, S.; Reader, T. *The Future of Safety Culture in European Air Traffic Management: A White Paper*; EUROCONTROL: Brussels, Belgium, 2021. Available online: https://skybrary.aero/sites/default/files/bookshelf/5993.pdf (accessed on 30 January 2026).