Haiku
Human AI teaming Knowledge and Understanding for aviation safety

Deliverable N. 7.2

Development of safety, HF and security approaches for Human Intelligent Assistance Systems

Authors: Paola Lanzi (DBL), Nikolas Giampaolo (DBL), Elisa Spiller (DBL)

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Abstract

One of the main goals of HAIKU is to develop a human-centric design approach, incorporating societal, value-based, ethical insights into AI design. In the project, attention is focused on the advance of a specific type of AI-power solution, i.e., Intelligent Assistants (IAs). In particular, the Use Cases (UCs) considered for validation include applications that transversally intertwine different aviation scenarios, covering respectively airport management, ATM and flight operations.

To pursue this objective of developing a human-centric design approach, the project embraces a holistic approach to Safety, HF and Security (SHS) assurance. Beyond the well-consolidated case-based approach proposed by the European Operational Concept Validation Methodology (E-OCVM) and further elaborated by the European research community, the HAIKU validation framework includes additional layers of analysis which consider the legal and ethical aspects of AI-enabled systems development and deployment. The overall 'case' for a new human-AI system will therefore depend on a combination of multiple sub-cases concerning safety, security, Human Factors (HF), ethics and legal aspects (including liability and regulatory compliance).

In this holistic framework, the different case-based approaches are tailored according to the specificities of AI-based systems and integrated. The reason is quite straightforward. The current aviation system is generally very operable, safe and secure, and liability is well understood in case of accidents. However, AI is a game-changer, and it cannot be assumed that existing Safety, HF and Security SHS methods will be sufficient to ensure that a new system is ready for use in live aviation operations. Instead, while AI is intended to improve at least some of these system performance areas, e.g. operability and safety, this cannot be taken for granted, and needs to be verified via various validation methods including human-in-the-loop trials. Additionally, for each prototype being developed, a combined (SHS) case needs to be carried out. Moreover, in the case of liability, should an AI's advice or action (or inaction) inadvertently contribute to an accident, new principles relating to liability need to be proposed, building on existing legal and regulatory frameworks.

The present document - D7.2 of the HAIKU project "Development of safety, HF and security approaches for Human Intelligent Assistance Systems" - presents a new framework along with methods for the assessment of the different dimensions taken into account. This framework will be iteratively applied to the HAIKU UCs as they evolve, and the results of the assessments will be reported in the annual editions of D7.3. Based on the results of this validation process, the integrated framework and the specific acceptable means of compliance will be reviewed and consolidated in the final version of D7.2, due at M36.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Information table

Deliverable Number	7.2
Deliverable Title	Development of safety, HF and security approaches for Human Intelligent Assistance Systems
Version	1.1
Status	Final
Responsible Partner	DBL
Contributors	Paola Lanzi (DBL), Nikolas Giampaolo (DBL), Elisa Spiller (DBL)
Contractual Date of Delivery	31.08.2023
Actual Date of Delivery	31.08.2023
Dissemination Level	PP

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Document History

Version	Date	Status	Author	Description
0.1	17/05/2023	Draft	Nikolas Giampaolo Paola Lanzi Elisa Spiller	
0.2	27/06/2023	Draft	Nikolas Giampaolo Paola Lanzi Elisa Spiller	
0.3	20/07/2023	Draft	Nikolas Giampaolo Paola Lanzi Elisa Spiller	
1.0	28/07/2023	Draft for internal review	Nikolas Giampaolo Paola Lanzi Elisa Spiller	
1.1	31/08/2023	Final version consolidated further to internal review by ECTL, EMBRT, DBL	Nikolas Giampaolo Paola Lanzi Elisa Spiller	

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

List of Acronyms

Acronym	Definition
AI	Artificial Intelligence
AI-RMF	Artificial Intelligence Risk Management Framework
ALTAI	Assessment List for Trustworthy AI
AMC	Acceptable Means of Compliance
ANS(s)	Air Navigation Service(s)
ANSP(s)	Air Navigation Service Provider(s)
Arg.	Argumentation
ARMS	Aviation Risk Management Solutions
ATC	Air Traffic Control
ATCO	Air Traffic Control Officer
ATCO-TO(s)	ATCO Training Organisation(s)
ATM	Air Traffic Management
ATO	Approved Training Organizations
CAMO(s)	Continuing Airworthiness Management Organization(s)
CIA	Confidentiality, Integrity, and Availability model
CPAIS	Collaborations Between People and AI Systems expert group
CSF	CyberSecurity Framework
DDoS	Distributed Denial-of-Service
DE	DEtect
DL	Deep Learning
DoS	Denial-of-Service

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

EAAI HLG	European Aviation High-Level Group on AI
EASA	European Aviation Safety Agency
EC	European Commission
ENISA	European Union Agency for Cybersecurity
E-OCVM	European Operational Concept Validation Methodology
ERC	Event Risk Classification
EU	European Union
EUROCAE	European Organization for Civil Aviation Equipment
FAA	Federal Aviation Authority
FSTD	Flight Simulation Training Device
GM	Guidance Material
GV	GoVern
HAIKU	Human AI teaming Knowledge and Understanding for aviation safety
HAT	Human AI Teaming
HAZOP	Human Hazard and Operability Study
HF	Human Factors
AI HLEG	High-Level Expert Group on Artificial Intelligence
HMI	Human-Machine Interface
HP	Human Performance
HPAP	Human Performance Assessment Process
IA(s)	Intelligent Assistant(s)
IATA	International Air Transport Association
ICAO	International Civil Aviation Organisation

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

ICT	Information and Communication Technology
ID	IDentify
IEC	International Electrotechnical Commission
IEEE SA	Institute of Electrical and Electronics Engineers Standards Association
ISO	International Organization for Standardization
IT	Information Technology
ITS	Intelligent Transportation Systems
JARUS	Joint Authorities for Rulemaking on Unmanned Systems
KPA(s)	Key Performance Area(s)
LbD	Legal by Design
LKB	Logic- and Knowledge-Based
LOAT	Level Of Automation Taxonomy
MAC(s)	Message Authentication Code(s)
ML	Machine Learning
MOC	Anticipated Mean Of Compliance
MRO(s)	Maintenance Repair Organisation(s)
NARS	Negative Attitude toward Robots Scale
NASA	National Aeronautics and Space Administration
NIST	National Institute of Standards and Technology (US)
NM	Network Manager
OT	Operative Technology
PIC(s)	Pilot(s) In Command
PR	PRotect

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

RC	ReCover
RS	ReSpond
SA	Situation Awareness
SAE	Society of Automotive Engineers
SAFE-AI	Situation Awareness Framework for Explainable AI
SAGAT	Situation Awareness Global Assessment Technique
SecRAM	SECurity Risk Assessment Methodology
SESAR	Single European Sky ATM Research
SHS	Safety, Human factors, Security
SHS-L	Safety, Human factors, Security, Liability
SIRA	Safety Issue Risk Assessment
SM ICG	Safety Management International Collaboration Group
SMS	Safety Management Systems
SOAR	State-Of-the-Art Review
SORA	Specific Operations Risk Assessment
SRA	Security Risk Assessment
STPA	Systems Theoretic Process Analysis
SUS	System Usability Scale
TAM	Technology Acceptance Model
TLX	Task Load Index
UC(s)	Use Case(s)
UEQ	User Experience Questionnaire
XAI	eXplainable AI

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Table of contents

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Table of Figures

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

1. Introduction

1.1. Scope of the document

This deliverable presents the results of Task 7.2 “Acceptable Means of Compliance for Intelligent Assistant”, as produced in the first 12 months of the HAIKU project.

In line with the purpose of the task, **it presents an integrated framework for the assessment of collaborative Human IA systems, that cover aspects of safety, Human Factors (HF), security, compliance and liability.** The definition of the framework is based on the review of the best practices for safety, HF, security (SHS) assurance processes in aviation and of the available compliance and liability assessment methods. The framework is complemented by a set of methods specifically selected to be applied in HAIKU for safety, HF, security, liability and compliance assessment and assurance.

The document is related to:

- D7.1 (from Task 7.1 “State of the art and regulatory landscape”) that presents the output of the current ethical and legal framework and a State-of-the-Art Review (SOAR) in regulations and consensus-based industry standards for the introduction of Artificial Intelligence (AI) in civil aviation. D7.1 also includes a set of tables to be used for the proactive compliance assessment of the Use Cases (UCs) during the project.
- D7.3 (from Task 7.3 “Safety, security and HF analysis”) in which the here-presented framework is applied for conducting a preliminary assessment of the safety, HF, security and liability (SHS-L) aspects of the different UCs. The results of the assessment have the purpose to feed the design of the UCs, and also to test the suitability and effectiveness of the framework presented in D7.2.
- D7.4 (from Task 7.4 “Legal Case and liability by design”) that will aggregate and summarise the overall set of results achieved during the project, with respect to compliance and liability risk analysis and mitigation.

This initial issue of the document – delivered at M12 (August 2023) – will be managed as an iterative live document during the entire project primarily. For this reason its dissemination level is limited to project partners. The final, consolidated, version of the report will be produced by M36 (August, 2025) and will have a public dissemination level.

1.2. Structure of the Document

This deliverable is divided into 9 parts: 7 sections (including the present introduction) and 2 Annexes (including references).

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

In the main **Sections**, numbered from 1 to 7, the reader will find **a general overview of the HAIKU validation framework and the methodologies adopted to improve the design of the UCs of the HAIKU project**.

Accordingly, the document is structured as follow:

1. Introduction, presenting the scope of the document and its structure;
2. The HAIKU design and validation framework (2);
3. Safety Methods and Assessments Frameworks for IAs (3);
4. Security Methods and Assessments Frameworks for IAs (4);
5. Human Factors Methods and Assessments Frameworks for IAs (5);
6. Methods to assess the legal and regulatory aspects of Human IA Systems (6).

Afterwards, the reader will find two **Annexes:** a list of references [Annex A - References and selected bibliography], and the assessment grids associated with practical data collection tools for the methods proposed in the main body of the document [Annex B - Assessment grids].

2. The HAIKU design and validation framework

One of the main goals of HAIKU is to develop a human-centric design approach, incorporating societal, value-based, ethical insights into the design and validation of AI-enabled Intelligent Assistants (IAs). These are a specific type of AI-based solutions, designed to collaborate with the operator and

HAIKU focuses on them in order to explore the topic of collaborative sociotechnical arrangements where human agents rely on and collaborate with new 'digital colleagues'.

In this document we propose a conceptual and operational framework to support the design and validation of such systems. **The framework addresses the multiple dimensions of human interaction with AI-based IAs and adopts a harmonised case-based approach to explore their impact on five specific key performance areas (KPAs), namely Human Factors (HF), Safety, Security, Liability and Legal Compliance.**

This approach goes beyond the well-consolidated case-based approach proposed by the European Operational Concept Validation Methodology (E-OCVM), and the HAIKU validation framework includes two additional layers of design and validation, which specifically consider the legal issues related to AI development and deployment and the compliance with current regulation. The reasons for this choice are twofold. First of all, the Consortium gives consideration to the upcoming compliance issues related to AI development and deployment. This is in line with the holistic human-centred approach to AI promoted by the European Commission (EC), aimed at the trustworthy development of this family of technologies (EC, 2020). Secondly, the Consortium aims to contribute to aviation culture by including regulation and liability as additional KPAs for the development and deployment of innovative technological solutions.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Overall, the HAIKU design and validation framework aims to assist in the consideration of HF, safety, security, liability and compliance aspects in the design and validation process of AI-based IAs, and to promote the gradual detection and mitigation of emerging issues concerning these topics. On a medium- and long-term basis, this approach proactively ensures the development of technologies that are compliant by design, safe, secure and acceptable in terms of HF and liability implications, improving the acceptability and marketability of the related solutions.

2.1. Compliance with related initiatives and approaches

Since the beginning of the HAIKU project, the field of AI in aviation has witnessed significant developments, and highlighted the need for an integrated and multilayered approach to the design and validation of AI-based systems.

The European Union (EU) initiative for trustworthy AI has made important progress and EU Aviation Safety Agency (EASA) has also delivered an updated version of its AI Roadmap (EASA, 2023). Meanwhile, global leading organisations for standardisation (e.g., International Organization for Standardization - ISO; International Electrotechnical Commission - IEC; Institute of Electrical and Electronics Engineers Standards Association - IEEE SA) are improving and enriching their standards to better address the certification of AI-based systems. Moreover, in 2023, the National Institute of Standards and Technology (NIST) delivered its own AI Risk Management Framework (AI RMF 1.0), providing further crucial insights for the development and deployment of AI (NIST, 2023).

The above-mentioned list includes initiatives aimed at regulating the future of AI for different scope and purposes. Each of these new frameworks add further inputs to better approach the challenges posed by this innovation. In order to ensure the consistency of the design and validation framework proposed by HAIKU with the overall set of initiatives currently ongoing at European and global level, four main sources were considered as reference, namely:

- the EU framework for a trustworthy and human centric AI
- the EASA roadmap for AI trustworthiness in aviation
- the international standards for AI
- the NIST AI Risk Management Framework

The key elements of these approaches and frameworks are reported in next sections.

2.1.1 EU framework for a trustworthy and human-centric AI

The first thread of development focuses on the EU Trustworthy Human-centric AI initiative. As explained in HAIKU D7.1 (HAIKU, 2023), **this initiative seeks to bolster and guide the advancement of AI technology within European industries, including aviation**. Unlike other sector-specific approaches, this endeavour adopts a general and agnostic perspective, transversally valid in different domains. Nonetheless, it offers a set of fundamental principles and a range of AI tools, products, and services that can be repurposed and prove highly advantageous in the aviation context. By adhering

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

to these principles, the European Trustworthy Human-centric AI initiative strives to ensure the ethical and responsible deployment of AI solutions in aviation operations.

In June 2018, the EC established the High-Level Expert Group on AI (AI HLEG) with the aim of supporting the implementation of the European AI strategy. The AI HLEG's responsibilities included providing recommendations on future policy development, addressing ethical, legal, societal, and socio-economic issues associated with AI. In April 2019, the AI HLEG proposed seven key requirements for trustworthy AI, which were published in its report on Ethics Guidelines on AI (AI HLEG, 2020).

In its Ethics Guidelines, the AI HLEG identified the AI ethics principles mirroring and relying on the main fundamental rights families:

- respect for human dignity,
- freedom of the individual,
- respect for democracy, justice and the rule of law,
- equality, non-discrimination and solidarity (including the rights of persons at risk of exclusion) and
- citizens' rights.

On these premises, the Guidelines outlined four fundamental principles:

- respect for human autonomy,
- prevention of harm,
- fairness and
- explicability.

To transpose these principles into concrete features and directly applicable rules, the AI HLEG also suggested a series of general requirements aimed at outlining the minimal compliance level of AI systems with the mentioned ethical expectations.

Basically, **the AI HLEG, as well as the EC, opted for a systematic understanding of the different ethical and socio-technical issues, including both individual and societal aspects.** Briefly, the list includes:

- **human agency and oversight**, including fundamental rights, human agency and human oversight;
- **technical robustness and safety**, including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility;
- **privacy and data governance**, including respect for privacy, quality and integrity of data, and access to data;
- **transparency**, including traceability, explainability and communication;
- **diversity, non-discrimination and fairness**, including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation;
- **societal and environmental wellbeing**, including sustainability and environmental friendliness, social impact, society and democracy;

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- **accountability**, including auditability, minimisation and reporting of negative impact, trade-offs and redress.

The AI HLEG's work was instrumental in shaping the development of the proposed AI Regulation. The proposed AI Regulation, officially known as the Regulation on a European approach for Artificial Intelligence (COM/2021/206 final), was introduced by the European Commission on April 21, 2021. It aims to establish a comprehensive regulatory framework for AI in the EU.

The AI Regulation builds upon the principles and recommendations put forth by the AI HLEG. It incorporates many of the group's proposals, such as the requirements for high-risk AI systems, transparency, human oversight, data governance and accountability. The Regulation seeks to balance the potential benefits of AI with the need to address potential risks, ensuring the protection of individuals' rights and the promotion of trustworthy AI.

The proposed AI Regulation focuses on establishing mandatory requirements for ensuring the trustworthiness of high-risk AI systems and outlines conformity assessment procedures that providers[1] must follow before introducing AI systems into the Union market, including those embedded in other products or services. In addition to these requirements, the regulation emphasises the need for high-quality data, comprehensive documentation, traceability, transparency, human oversight, and robustness in order to mitigate potential risks to fundamental rights and safety that are not covered by existing legal frameworks.

Regarding the assessment of high-risk AI systems, the pending proposal stipulates that an internal check should ensure full compliance with all the requirements of the future Regulation and recommends that providers adopt post-market monitoring systems as part of good risk management practice. These systems enable efficient identification and resolution of any emerging risks associated with AI-based systems after they have been introduced to the market. In addition, the draft Regulation encourages providers of non-high-risk AI systems to voluntarily adopt codes of conduct that align with the mandatory requirements applicable to high-risk AI-based systems (further details in D7.1).

The AI HLEG provided guidance and expertise on AI ethics, principles, and policy, which heavily influenced the development of the proposed AI Regulation. The Regulation, in turn, is a legal framework that aims to govern the use and deployment of AI technologies within the EU.

[1] According to the AI Act proposal (COM/2021/206 final), provider is intended as «a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge», as per Article 3(1bis).

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

2.1.2 The EASA Roadmap for AI trustworthiness in aviation

The Trustworthy AI initiative, as well as the AI Act, have a general scope and aim to provide a common framework for the regulation of AI in Europe. The EC, however, underlines the importance of adapting these general rules to the specific needs of safety-critical domains, including aviation.

According to Reg. (EU) 2018/1139, this mandate falls within the scope of competencies of EASA. This is the reason why the Agency launched its sector-based AI Roadmap. In May 2023, the second release of the document was published, which amends and integrates the previous one, dated February 2020. Over the last few years, this programme document was also complemented by more operative guidelines provided by two EASA Concept Papers, respectively reporting the "First usable guidance for Level 1 machine learning applications" (December 2021) and "First usable guidance for Level 1 and 2 machine learning applications" (February 2023). This second document updates the previous one.

For the HAIKU design and validation framework both these document series deserve particular attention. They indeed provide **official guidance for applicants introducing AI-based technologies into systems intended for the use of safety-related applications in all domains covered by the EASA competencies** (EASA, 2023, p. 4). Here you find a brief introduction to the framework and the main novelties elicited by the Agency. Further information is available in the full version of the documents.

First of all, even though the definition of AI according to the AI Act proposal is broad and comprehensive[2] EASA furtherly defined the scope of its AI, particularly focusing on machine learning (ML) (including deep learning (DL)), logic- and knowledge-based (LKB) approaches, hybrid AI and statistical approaches.

Moreover, EASA created and improved its framework for AI trustworthiness with the purpose of enabling readiness for use of AI in aviation. As illustrated by the figure 1 below, the Agency concept is based on the general EC Ethical Guidelines, whose objectives, principles and requirements are transposed into the aviation domains by means of a process articulated in four building blocks, which concern respectively: AI trustworthiness analysis, AI assurance, HF for AI and AI safety risk mitigation.

[2] AI Act proposal (COM/2021/206 final), Article 3(1): «'artificial intelligence systems' (AI systems) means software that is developed with one or more of the techniques and approaches listed in Annex I [e.g., machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning; logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines,(symbolic) reasoning and expert systems; statistical approaches, Bayesian estimation, search and optimization methods] and can, for any given set of human-defined objectives, generate outputs such as contents, predictions, recommendations, or decisions influencing the environments they interact with».

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

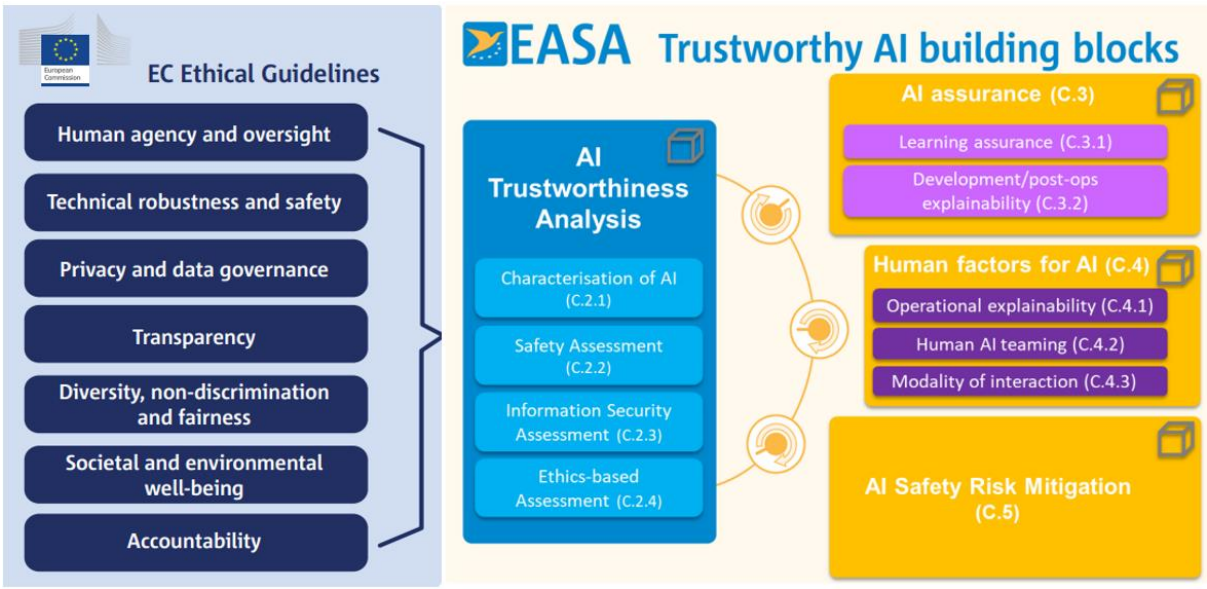

Figure 1. EASA AI trustworthiness building blocks

Building block 1: AI trustworthiness analysis

The first fundamental step of this process is the AI trustworthiness analysis which serves as a gateway to the three other technical building blocks. This analysis is composed of four activities: the characterization of AI applications and safety, security and ethics-based assessment.

The Characterization of AI applications is based on the EASA classification of AI applications. According to Agency guidance, AI classification is built on high-level tasks and AI-based systems definition and functional analysis of the application at issue. As shown by figure 2 below, in the new version of the Roadmap, this classification scheme has been refined, further specifying the contents of the levels according to a more exhaustive approach to human-AI teaming (HAT) interactions.

Level 1 AI: assistance to human	Level 2 AI: human-AI teaming	Level 3 AI: advanced automation
• Level 1A: Human augmentation • Level 1B: Human cognitive assistance in decision-making and action selection	• Level 2A: Human and AI-based system cooperation • Level 2B: Human and AI-based system collaboration	• Level 3A: The AI-based system performs decisions and actions that are overridable by the human. • Level 3B: The AI-based system performs non-overridable decisions and actions (e.g. to support safety upon loss of human oversight).

Figure 2.EASA AI trustworthiness building blocks

The AI levels are meant to be a classification of AI-based systems in light of their usage and interaction with

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

human end-users and not a mere automation scheme. This is why, especially when considering Level 2 and 3 applications, discriminating between them relies on the notion of authority, intended as «*the ability to make decisions and take actions without the need of approval from others*» (EASA, 2023, p. 24). To support the classification, EASA thus furtherly defined three different scenarios, respectively distinguishing among (1) full authority for the end users, (2) partial authority for the end users and (3) authority for the end users upon alerting.

In light of the above, each AI-based system has to be preliminarily classified. Indeed, proportionality and modulation of the AI guidance are primarily driven by the level of AI, and thus by the output of the characterization of the AI application. This step has a crucial value in the HAIKU validation framework, since the set of criteria adopted for SHS-L assessment methodologies may vary according to the level of each IA.

The next steps of the AI trustworthiness analysis encompass the **safety and security assessment**, according to the objectives and the anticipated means of compliance (MOCs) provided by EASA in its second concept paper (EASA, 2023, p. 20-40), and the **ethics-based assessment**, according to the customised version of the AI HLEG' Assessment List for Trustworthy AI (ALTAI) for aviation proposed by EASA in that same document (EASA, 2023, p. 214).

Building block 2: AI assurance

The second block proposes system-centric guidance to address the development of AI-based systems, in accordance with the new learning assurance concept. Learning assurance can be defined as follows:

> *All of those planned and systematic actions used to substantiate, at an adequate level of confidence, that errors in a data-driven learning process have been identified and corrected such that the system satisfies the applicable requirements at a specified level of performance, and provides sufficient generalisation and robustness guarantees (EASA, 2023, p. 13).*

More details are available in the concept paper, including the specific objectives and anticipated MOCs provided by EASA for learning assurance (EASA, 2023, p. 46-77).

For the purposes of the HAIKU validation framework, the new definition of **explainability**[3] deserves particular attention. In this context, EASA defined AI explainability as the:

[3] EASA defined AI explainability as the «capability to provide the human with understandable, reliable, and relevant information with the appropriate level of details and with appropriate timing on how an AI/ML application produces its results». Note that, whereas ' explainability ' refers to the capability, 'explanation' refers to the information as an instantiation of the explainability (EASA, 2023, p. 14).

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

capability to provide the human with understandable, reliable, and relevant information with the appropriate level of details and with appropriate timing on how an AI/ML application produces its results[4] (EASA, 2023, p. 14).

In the aviation domain a number of stakeholders need and require explanations for different purposes. The nature and quality of the explanations, therefore, may be affected by the target audience, as well as by different contextual aspects, such as the time to get the explanation or the format of the information obtained and the layout of Human-Machine Interface (HMI). This is why the guidance provided by EASA distinguishes between two types of explainability driven by the profile of the users and their needs. On the one hand, there is the information required to make a ML model understandable **(Development & Post-operation Explainability)**; on the other, understandable information for the operational user on how the system came to its results **(Operational Explainability)** (EASA, 2023, p. 14).

According to the specific motivations the different stakeholders may have on explanations, EASA identified three categories of actors, namely: 1) those involved in the developing of AI applications; 2) those involved in working operationally with AI applications and, eventually, 3) those involved in analysing what an AI application has done during the operations. This classification of actors leads to the definition of two types of explainability, one related to the development cycle and the post-operational phase and the other focused on operations per se.

According to these needs, EASA provided a set of specific objectives and the anticipated means of compliance (MOC) for the development and post-operations AI explainability and AI data recording capability (EASA, 2023, p. 72-77).

For the purposes of HAIKU, another way to consider explainability is via the intended recipient of the explanations. There is the system developer and maintainer who need to verify the technical explainability underpinning the AI system's functioning. Next there is the operational user, e.g flight crew, ATCO or airport manager, who need to trust the advice they are being given or the actions the AI system is taking. Last there are organisational personnel who want to know if the system is optimal, and these could for example be in an operational organisation's safety or training department. These three groups are essentially asking several questions:

- Is the system working properly/accurately? [system verification]
- Is it providing useful guidance/decisions/actions in the moment? [user validation]
- Is it adding value (safety, operational performance) to the organisation's activities, and can it be improved? [organisational validation]

[4] Note that, whereas ' explainability ' refers to the capability, 'explanation' refers to the information as an instantiation of the explainability (EASA, 2023, p. 14).

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Building block 3: HF for AI-based application

The third building block concerns **HF for AI-based applications**.

Focusing mainly on the need for operational explainability, EASA remarks that the introduction of AI is expected to modify the paradigm of interaction between human end-users and systems, and this will affect the task function allocation and distribution by progressively giving more authority to the AI-based applications. If not done in a human-centric way, this may lead to a reduction of end-users' awareness of the logic behind automated decision-making and, consequently, to a reduction or failure in establishing trust.

The HF section of the EASA document (EASA, 2023, p. 88, spec. § 4.2) devotes particular attention to the concept of HAT[5] to ensure adequate cooperation or collaboration between human end-users and AI-based systems to achieve certain goals. AI-based systems will become teammates for the human end-users. Therefore, HAT aims to address the critical challenges posed by the transition from human-human teams to human-AI-based teams. In particular, EASA focuses on the notions of cooperation and collaboration. More specifically, EASA makes explicit that:

> ***Cooperation*** *is a process in which the AI-based system works to help the end user accomplish their own objective and goal. The AI-based system will work according to a predefined task allocation pattern with informative feedback on the decision and/or action implementation. Cooperation does not imply a shared vision between the end user and the AI-based system. Communication is a paramount capability for cooperation (EASA, 2023, p. 16).*

and

> ***Cooperation*** *is a process in which the AI-based system works to help the end user accomplish their own objective and goal. The AI-based system will work according to a predefined task allocation pattern with informative feedback on the decision and/or action implementation. Cooperation does not imply a shared vision between the end user and the AI-based system. Communication is not a paramount capability for cooperation (EASA, 2023, p. 16).*

The expected AI-based systems capabilities for cooperation and collaboration scenarios are different. Therefore, the design of applications should target different goals requiring different types of

[5] EASA still has not provided its official definition of Human AI Teaming (HAT). Implicitly, the lemma refers to new forms of interactions between human agents and AI, also in light of the traditional iterations observed in human-human teams (EASA, 2023, p. 88). Generally, for the purposed of EASA, this new concept currently refers to the cooperation and collaboration between the end user and the AI-based system to achieve goals. This HAT concept, depending on the maturity of the AI-based system, involves a shared understanding of goals, roles and processes (decision-making/problem solving) between the members (EASA, 2023, p. 16 and p. 88-91).

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

interactions (EASA, 2023, p. 16). In particular, for efficient collaboration, AI-based systems (EASA, 2023, p. 89) should be designed to:

- enable and facilitate the sharing of elements of situational awareness;
- identify abnormal situations and perform diagnostics;
- evaluate the relevance of the solution proposed by the end-user;
- negotiation/argumentation; and
- adaptiveness.

In light of the above, EASA provides preliminary insights about the modality of interaction and style of interfaces, design criteria for gesture or non-verbal language, and design criteria for the management of multi-modal interaction. The EASA also introduces guidance for error management. The complete set of proposed objectives and anticipated MOCs about operational explainability, HF and HAT is available in the extended version of the second concept paper delivered by EASA (EASA, 2023, p. 84-98).

Building block 4: AI safety risk mitigation

The fourth building block for AI trustworthiness in aviation is **AI safety risk mitigation** (AI-SRM) which is aimed to mitigate the risks due to the partial satisfaction of explainability and learning assurance requirements. The intent of such mitigations is to minimise as far as practicable the probability of the AI/ML constituents as well as the related systems producing unintended or unexplainable outputs.

According to EASA guidance, this goal could be achieved by several means. The second concept paper, in particular, suggests as best suitable options the real-time monitoring of the output of the AI/ML constituent and passivation of the AI-based system with recovery through a traditional backup system (e.g., safety net) or, in a wider horizon, by considering the notion of 'licensing' for an AI-based agent (EASA, 2023, p. 99). Moreover, it should be essential that **organisations** that aim to introduce AI-based systems in their operations would introduce the due adaptations to their protocols and operative contexts, also in light of the objectives and MOCs introduced by EASA. In particular, they should be able demonstrate they can meet the objectives defined for each AI trustworthiness building block and maintain an adequate compliance level over the whole technology lifecycle (EASA, 2023, p. 101).

It is worth noting that, unlike in the past, the **life cycle process of AI-based systems** has a larger scope compared to traditional systems development and implementation. Developers, producers, providers as well as end-user organisations are responsible for data collection and data governance and continuous safety management process, including in the operational phase of the product life cycle. Moreover, safety management must include additional requirements for human end-users training phases, taking into consideration skilling, de-skilling, and upskilling needs from the early stage of technology development and deployment (EASA, 2023, p. 19).

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

2.1.3 International Standards for AI

As anticipated in D7.1 (HAIKU, 2023), the development of standards for AI is currently under progress, driven by the technical community's efforts to address the challenges associated with AI risk assessment. Leading international organisations, such as ISO, IEC, the Society of Automotive Engineers (SAE), the European Organization for Civil Aviation Equipment (EUROCAE), and the IEEE SA, are actively involved in formulating standards for AI risk management and governance. The objective is to establish guidelines and frameworks that can be universally applied to ensure the responsible and ethical implementation of AI technologies.

These organisations - namely ISO, IEC, and IEEE SA - recognize the significance of establishing a common set of principles and practices to guide organisations in managing AI risks. By defining comprehensive frameworks for risk management and governance, these standards aim to minimise potential harms, enhance accountability, and promote transparency in AI systems.

The US NIST also plays a crucial role in shaping the AI landscape. NIST is actively engaged in creating a series of documents and organising workshops to establish a robust risk management framework and corresponding standards for trustworthy AI. Their focus is on enabling organisations to effectively evaluate, assess, and mitigate risks associated with AI technologies while upholding principles of fairness, transparency, and privacy.

The following standards and recommended practices play a significant role in establishing a framework for responsible and trustworthy AI systems, emphasising ethical considerations, accountability, and the mitigation of potential risks:

- **ISO/IEC 23894 on Artificial Intelligence and Risk Management**: This standard provides guidelines on managing risks during the development and application of AI techniques and systems. It assists organisations in integrating risk management into their AI-related activities and functions, addressing specific issues such as transparency, explainability, privacy, fairness, safety, security, and human rights.
- **ISO/IEC 42001 on Artificial Intelligence — Management System**: This standard aims to establish requirements for implementing and maintaining an AI management system. It provides guidelines for deploying controls to measure the effectiveness and efficiency of AI processes, helping organisations develop or use AI responsibly and meet regulatory requirements.
- **ISO/IEC 38507 on Governance implications of the use of artificial intelligence by organisations**: This standard offers guidance for governing bodies of organisations to ensure effective and acceptable use of AI. It emphasises the establishment of policies, accountability, authority, and compliance management related to AI systems.
- **IEEE P2863 on Recommended Practice for Organisational Governance of Artificial Intelligence**: This recommended practice outlines governance criteria and process steps

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

for implementing AI within organisations, focusing on safety, transparency, accountability, responsibility, and bias reduction.

- **IEEE 7000-2021 on Model Process for Addressing Ethical Concerns During System Design**: This standard provides processes for incorporating ethical values throughout the concept exploration and development stages. It facilitates transparent communication with stakeholders and enables traceability of ethical values in system design.
- **IEEE 7010-2020 on Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being**: This recommended practice offers contextual measures of well-being and guidance for assessing the impact of autonomous and intelligent systems on human well-being throughout the life cycle.
- **NIST Special Publication 1270 on A proposal for Identifying and Managing Bias in Artificial Intelligence**: This report proposes a strategy for managing AI bias and identifies prominent biases that can contribute to societal harms. It suggests an approach encompassing pre-design, design and development, and deployment stages to address bias effectively.
- **NISTIR 8332: Trust and Artificial Intelligence**: This report highlights the importance of user trust in AI systems and provides an overview of the challenges associated with trust in AI. It emphasises the perception of technical trustworthiness and its impact on user trust.
- **SAE G34 / EUROCAE WG-114** Machine Learning standard. This standard aims to provide a comprehensive approach to the development, validation, and deployment of ML algorithms in safety-critical applications, ensuring their reliability, robustness, and compliance with relevant regulatory requirements. The standard addresses key aspects such as data collection, model training, verification, and ongoing monitoring, emphasising the need for transparency, interpretability, and rigorous assessment of ML models in safety-critical environments.

2.1.4 The NIST AI Risk Management Framework

The Artificial Intelligence Risk Management Framework (AI RMF 1.0) is a framework very recently developed by the US NIST under the initiative known as NIST AI 100-1 (NIST, 2023). The NIST AI 100-1 initiative itself was launched in response to the growing importance of AI and the need for addressing its associated risks. It considers that AI systems are deployed in ever more safety-critical and consequential situations, and that AI researchers and developers will increasingly confront safety, security, ethical, and legal challenges. In such a situation, understanding and managing the risks of AI systems will help enhance trustworthiness, and cultivate public trust.

The framework aims to provide guidance and a systematic approach for managing risks associated with the deployment and operation of artificial intelligence (AI) systems.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

The Framework is divided into two parts. Part 1 discusses how organisations can frame the risks related to AI and describes the intended audience. Next, in Part 2, AI risks and trustworthiness are analysed, outlining the characteristics of trustworthy AI systems, which include valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy enhanced, and fair with any harmful biases managed.

This framework provides a comprehensive understanding of the characteristics that contribute to trustworthy AI and offers guidance on how to address them effectively. The AI RMF Core is designed to facilitate dialogue, understanding, and activities related to managing AI risks and promoting the development of trustworthy AI systems. Figure 3 illustrates the Core, which consists of four functions: GOVERN, MAP, MEASURE, and MANAGE.

Each function is further divided into categories and subcategories, with specific actions and outcomes associated with them. It is important to note that these actions are not intended as a checklist nor a strict sequence of steps.

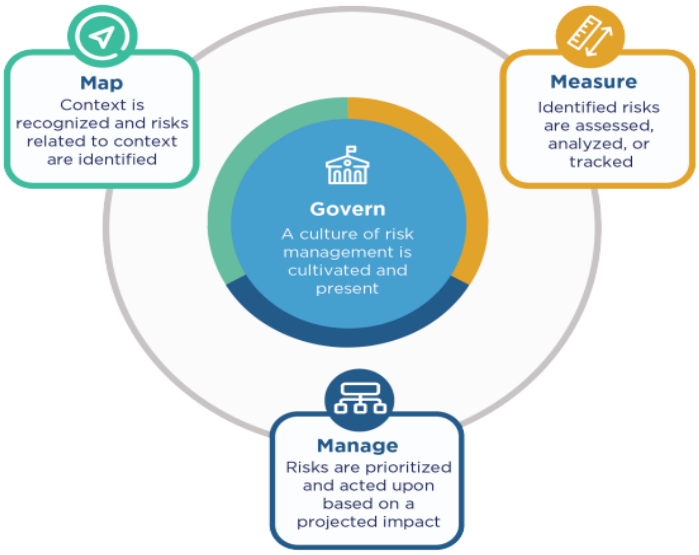

Figure 3.The NIST AI Risk Management Framework's Core

The GOVERN function is infused throughout the AI risk management process and enables the other functions. Governance is a continual and intrinsic requirement for effective AI risk management over the lifespan of an AI system and an organisation's hierarchy.

The MAP function establishes the context to frame risks related to an AI system and enhances an organisation's ability to identify risks and broader contributing factors. It is used to gather information to anticipate, assess, and address potential sources of negative risks, to mitigate uncertainty and enhance the integrity of the decision process. It enables proactive prevention of negative risks and the development of trustworthy AI systems by improving understanding of contexts, identifying positive

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

and negative impacts, and anticipating risks beyond the intended use of AI systems.

The MEASURE function employs quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyse, assess, benchmark, and monitor AI risks and related impacts, informing the MANAGE function. It uses knowledge relevant to AI risks identified in the MAP function to analyse, assess, benchmark, and monitor AI risks and related impacts, including tracking metrics for trustworthy characteristics, social impact, and human-AI configurations. It provides a traceable basis to inform management decisions when trade-offs among trustworthy characteristics arise.

The MANAGE function implements the identified risk management processes to maintain and improve the trustworthiness of AI systems. It aims to maintain and enhance the trustworthiness of AI systems by applying appropriate strategies to mitigate risks. This function involves ongoing monitoring, adaptation, and improvement of risk management practices.

Each function of the AI RMF Core is not a checklist of ordered steps but a flexible framework.

Framework users will enhance their purpose-driven culture focused on risk understanding and management by executing the GOVERN function continually as knowledge, cultures, and needs or expectations from AI actors evolve over time. All metrics and measurement methodologies developed and used in the AI RMF Core should adhere to scientific, legal, and ethical norms and be carried out in an open and transparent process. Framework users may need to develop new types of measurement, both qualitative and quantitative, to provide unique and meaningful information to the assessment of AI risks.

Trustworthy AI systems are characterised by their validity and reliability, safety, security, resilience, accountability, transparency, explainability, interpretability, privacy enhancement, and fair treatment with managed harmful biases**.** Achieving trustworthiness in AI necessitates striking a balance among these characteristics, taking into account the specific context in which the AI system will be employed. While all these characteristics are attributes of socio-technical systems, accountability and transparency also encompass the internal processes and external factors related to an AI system (figure 6). Neglecting any of these characteristics can significantly increase the likelihood and severity of negative consequences.

It is important to recognize that the trustworthiness characteristics outlined in this document are interrelated and mutually influential. Systems that are highly secure but unfair, accurate but lacking transparency and interpretability, or inaccurate despite being secure, privacy-enhanced, and transparent are all deemed undesirable. **An effective approach to risk management involves carefully balancing the trade-offs among these trustworthiness characteristics.**

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

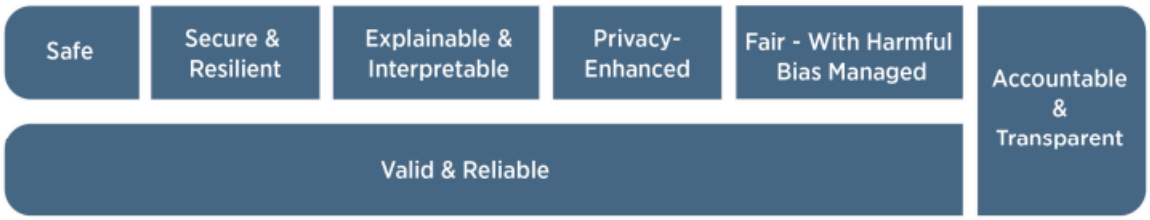

Figure 4. Characteristics of trustworthy AI systems.

2.2 HAIKU: towards an integrated assessment framework

A comparative analysis of the existing frameworks and approaches presented in previous sections shows a substantial alignment of principles and requirements, but also some divergences in the definition of the KPAs to be included in the framework.

Interestingly, the requirements defined by the AI HLEG for trustworthy AI (AI HLEG, 2020) and reconsidered by EASA (EASA, 2023, p. 40-45) align with the characteristics and requirements presented in the AI RMF (AI RMF 1.0) developed by the NIST.

In particular:

- **Human agency and oversight** (AI HLEG) aligns with the characteristics of accountability and fair treatment with managed harmful biases in the AI RMF 1.0. It emphasises the need for human control, decision-making, and ensuring that AI systems are designed to avoid unfair biases and discrimination.
- **Technical robustness and safety** (AI HLEG) corresponds to the characteristics of validity, reliability, safety, security, and resilience in the AI RMF 1.0. It emphasises the need for AI systems to be developed and deployed in a manner that ensures their robustness, safety, and security.
- **Privacy and data governance** (AI HLEG) aligns with the characteristic of privacy enhancement in the AI RMF 1.0. It highlights the importance of protecting personal data and ensuring that AI systems enhance privacy rights.
- **Transparency/Accountability** (AI HLEG) aligns with the characteristics of transparency and explainability in the AI RMF 1.0. It emphasises the need for AI systems to operate in a transparent manner, providing clear explanations of their capabilities and outcomes.
- **Diversity, non-discrimination, and fairness** (AI HLEG) aligns with the characteristic of fair treatment with managed harmful biases in the AI RMF 1.0. It emphasises the importance of addressing biases and promoting diversity, non-discrimination, and fairness in AI systems.

This mapping showcases the convergence of requirements across frameworks, illustrating the shared goal of establishing trustworthy AI systems characterised by validity and reliability, safety, security,

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

resilience, accountability, transparency, explainability, interpretability, privacy enhancement, and fair treatment with managed harmful biases.

At the same time, KPAs do not seem to be defined in the same harmonic way. A key example is the way HF are addressed and considered in the different frameworks. While safety and security are defined as homogeneously formulated as KPAs, HF in some cases (i.e., the EASA roadmap) is a KPA, while in others (i.e. the NIST AI RFM) it is not, being instead assimilated to aspects of explainability and interpretability that although relevant cover just a part of the relevant HF aspects.

Going further towards the development of the HAIKU design and validation framework, it seems that for the purposes of HAIKU the requirements outlined by the AI HLEG, EASA and NIST could be grouped into five main categories, i.e., safety, security, HF, liability and legal compliance.

The HAIKU project embraces these categories as fundamental pillars for the assessment of the UCs. By evaluating the UCs against these five key performance factors, a comprehensive analysis can be conducted, ensuring adherence to the requisite standards outlined by the frameworks in considerations.

In this regard, the HAIKU Consortium has assimilated the AI HLEG and EASA guidance updating its validation, now structured into the five key performance areas as proposed by figure 5 below.

Figure 5.The dimensions of the HAIKU's design and validation framework

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

More specifically, the scope of each KPA can be described as follow:

- **Safety**: This performance area focuses on ensuring the safety of AI-enabled systems and of their interactions with users and the environment. It involves implementing robust mechanisms to identify and mitigate potential risks and hazards associated with AI technologies. By prioritising safety, organisations can instil confidence and trust in the reliability and integrity of AI systems.
- **Security**: The security performance area emphasises the protection of AI systems against unauthorised access, data breaches, malicious attacks and malevolent usages. This entails identifying potential threats to AI systems and their associated data, and implementing security measures (such as encryption and access controls) to ensure that their confidentiality, integrity, and availability are maintained. By addressing security concerns, organisations can mitigate risks and ensure that the confidentiality, integrity, and availability of AI systems and their associated data are maintained.
- **Human Factors**: This area aims to enhance human performance by leveraging AI technologies. It involves using AI systems to augment human capabilities, foster H-AI cooperation, improve decision-making processes, and streamline tasks. By designing AI systems that are intuitive, user-friendly, and adaptable to user needs, organisations can empower individuals to leverage the full potential of AI in achieving their objectives.
  The area also encompasses considerations such as transparency, accountability, fairness, and the avoidance of biased outcomes.
- **Legal compliance**: The legal compliance performance area emphasises compliance with applicable laws, regulations, and legal frameworks governing AI systems. It involves understanding and adhering to legal obligations related to privacy, data protection, intellectual property rights, and other relevant legal considerations. By ensuring legal compliance, organisations can better fulfil their accountability duties, mitigate legal risks, improve public trust, and operate within the boundaries of the law.
- **Liability**: The performance area concerning liability aims to assess the risk exposure of producers, providers, user organisations and human end-users when a new AI-based system is introduced and the existing protocols and standards have to be modified. The results obtained will help to address these issues step by step, introducing suitable mitigations and improving the concept design.

The table below matches the KPAs included in the HAIKU design and validation framework with the AI trustworthiness objectives, showing the contribution that each of these KPAs offer for the achievement of EC and EASA objectives for AI trustworthiness in aviation. From the table it is evident that:

- the KPAs of the HAIKU harmonised design and validation framework completely cover all the AI trustworthiness objectives. This means that the application of the framework aims at ensuring that the AI-enabled system being designed achieves these objectives.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- the association between AI trustworthiness objectives and KPAs is not 1 to 1, thus implying a need for considering and integrating multiple perspectives when addressing specific objectives. For example, the objective of human agency and oversights is related to safety, HF, liability, and legal compliance. This means that it will be investigated from the multiple perspectives of these KPAs, and the different results obtained shall be compared and integrated.

		KPAs of the HAIKU design and validation framework				
		Safety	**Security**	**HF**	**Liability**	**Legal compliance**
AI Trustworthiness objectives	**Human agency and oversight**	✓		✓	✓	✓
	Technological robustness and safety	✓	✓		✓	✓
	Privacy and data governance		✓			✓
	Transparency	✓		✓	✓	
	Diversity, non-discrimination and fairness			✓		✓
	Societal and environmental well-being		✓			
	Accountability				✓	

As in HAIKU a case-based approach is proposed to practically apply this framework, the next sections of the documents are dedicated to present the case-based approach and methods proposed. For each of the KPAs, we present the specific challenges taken into account by HAIKU, the analysis of relevant already existing and the methods, and the specific case-based processes and techniques proposed to be adopted. In most cases, the proposed processes and techniques are complemented by practical questionnaires/checklists to be used by and with the UC leaders in order to get a screening of safety, security, HF, liability and legal compliance issues and mitigations. A preliminary application of the framework and the results obtained are reported in D7.3.

3. Safety Methods and Assessments Frameworks for IAs

3.1. Emerging Safety Issues concerning AI in Aviation

The literature discussing the safety of AI in the aviation sector often uses ML as a general proxy for AI. However, it is important to note that the research in this field is still in its early stages. Therefore, we have also included relevant papers that specifically refer only to ML.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

The challenges in achieving safety in artificial intelligence (AI) can be broadly categorised into three groups: robustness, assurance, and specification (Rudner & Toner, 2021). **Robustness** ensures that the AI system operates safely within acceptable limits, even in unfamiliar or unpredictable settings. **Assurance** refers to the ability to analyse and understand the AI system easily by human operators, while **specification** is concerned with ensuring that its behaviour aligns with the system designer's intentions.

For a ML system to be **robust**, it must operate safely and predictably in a variety of conditions, including those it was not explicitly trained for. One way to reduce the chance of failure in such situations is to incorporate confidence levels in the system's predictions, allowing it to recognize when it is uncertain and take appropriate action, such as reverting to a safe fallback option or alerting a human operator.

However, challenging inputs can take various forms, including those that the system has not encountered before, requiring it to recognize its limitations and act safely. Research in this area aims to train ML models to estimate confidence levels, called predictive uncertainty estimates, which can alert a human operator when inputs significantly differ from those on which the system was trained. AI systems trained on historical data may struggle to perform well in new or dynamic situations. The evolution of **operational conditions** in AI refers to the challenge of adapting AI models to changing environments, scenarios, or circumstances over time. This issue necessitates research into techniques like transfer learning, domain adaptation, and reinforcement learning to ensure AI systems remain effective as conditions change.

Moreover, robustness involves reliability **and security**, ensuring that the system behaves as intended in a wide range of situations, including under adversarial attacks (Dafoe, 2018). Additionally, **correctability** is essential to enable the system to be optimally open to correction by human overseers. A challenge relates to ensuring that the system behaves as intended through the entire lifespan of AI systems, including their maintenance, upgradability, and governance. Maintenance involves ensuring AI systems continue to function effectively, adapt to changes, and remain secure. Upgradability focuses on the ability to update and improve AI models to keep up with evolving needs and technology. Governance addresses the ethical and regulatory considerations in deploying AI, including accountability, transparency, and compliance throughout the **AI system's life cycle. Verification and validation** are crucial steps in ensuring the reliability and correctness of AI systems. AI systems pose challenges for conventional verification and validation (V&V) methods. However, alternative methods like model checkers and static analysis tools from software engineering can be adapted (Menzies & Pecheur, 2005).

Ensuring the safety of a machine learning system requires human operators to have a clear understanding of how the system operates and whether its behaviour aligns with the intended design. The robust assurance techniques used with traditional computer systems are strongly challenged by - or not even applicable to - ML techniques like deep neural networks. Designing systems with easily

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

understood decisions, allowing human operators to ensure that the system functions as intended and receive an explanation in case of unexpected behaviour, is of greater importance. In particular, **interpretability** in AI involves understanding a model's internal workings, while **explainability** takes it a step further by providing clear, human-readable explanations for the model's decisions.

One of the major challenges in ML systems is the presence of **entangled uncertainties** (Rueß & Burton, 2022)**.** Uncertainty arises from various sources, including the inductive capability of ML algorithms for extracting models from data, the uncertainty surrounding the operating context, models of the operating context and the human user, and behavioural uncertainty due to the approximate nature of heuristic learning algorithms. Additionally, uncertainty can arise from probabilistic and non-deterministic components, safety hazards, and safety envelopes in uncertain operating contexts. The uncertainty concerning meaningful fallbacks to responsible human operators and self-learning systems' emergent behaviour over time further complicates the situation. Another source of uncertainty is stochastic search heuristics that may lead to incorrect recall even for inputs from the training data, as well as the unpredictable nature of generalising from given data points. Furthermore, uncertainty surrounding the faithfulness of training data representing operating contexts and the correctness and generalizability of training itself further compound the uncertainty. This could lead to the issue of **brittleness**. AI brittleness stems from vulnerability to slight input variations due to models lacking robustness and adaptability. Though excelling in specific tasks, AI struggles with deviations from training data, risking failures (McCarthy, 2007).

The presence of **biased data** within AI training datasets poses a pivotal concern. Training bias, an aspect of AI data bias, stems from inadvertent systemic inequalities ingrained in historical data collection practices. Historical imbalances in data aggregation methodologies can lead to the skewed representation of specific aviation scenarios or aircraft classifications. Concurrently, the inadvertent transference of human cognitive biases to data annotation and collection processes can inadvertently perpetuate prejudiced data patterns. The operational implementation of AI systems, tasked with pivotal roles such as flight navigation and maintenance diagnostics, could be fraught with inaccurate determinations. Notably, biased training data might culminate in erroneous collision avoidance decisions or misinterpretations of engine performance data, thus impinging on the fundamental underpinnings of aviation safety. Addressing **data quality** involves data cleansing, validation, preprocessing, and ensuring diverse and representative datasets to improve the overall effectiveness and fairness of AI models.

The manifestation of uncertainty poses significant challenges to the safety assurance of safety-critical systems in several ways (Dafoe, 2018). Firstly, the operational domain's scope and unpredictability make it difficult to define desirable system behaviour for each possible set of conditions, leading to insufficiencies in the resulting system specification. Secondly, the complex, unpredictable environment is measured using imperfect sensors that provide a noisy, incomplete view of the environment, leading to inaccuracies and noise in sensors and signal processing. Finally, the use of AI

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

and ML techniques to solve the problem of uncertainties in the inputs to the system introduces another class of uncertainties related to the perception and decision-making functions.

Complexity is another challenge that arises from the interaction between parts of the system leading to behaviour that could not be predicted by considering individual parts and their interactions alone (Rueß & Burton, 2022). Complexity can manifest itself within different levels of the system, such as the increasing complexity within the E/E architecture, which results from the increasing number of technical components, their heterogeneity and technical implementation, the use of components and software of unknown pedigree, and changes in the system after release due to software updates or the integration of additional services. Non-linearity, mode transitions, and tipping points are examples of complexity's impacts, where the system may respond unpredictably based on its current state or context.

In the context of ML systems, the term "**specification**" refers to the process of defining the system's objective in a manner that ensures its behaviour aligns with the human operator's intentions. Typically, a ML system follows a pre-specified algorithm to learn from data, enabling it to achieve a specific goal. The learning algorithm and objective are typically provided by a human system designer. Examples of objectives may include minimising prediction error or maximising a reward.

It is important to note that during the training process, a ML system will attempt to reach the given objective, regardless of how well it reflects the designer's intentions. Therefore, designers must take care to specify an objective that will lead to the desired behaviour. If the objective set by the designer is a poor proxy for the intended behaviour, the system may learn the wrong behaviour, resulting in a "misspecified" system. This outcome is particularly likely in settings where the specified objective cannot fully capture the complexities of the desired behaviour. Poor specification of a ML system's objective can lead to safety hazards if the misspecified system is deployed in a high-stakes environment and does not operate as intended.

Value specification is a critical component of ML system design. Value specification in AI involves guiding an AI system's behaviour to align with human values through methods such as human guidelines, ethical frameworks, transparency, user feedback, and collaboration with domain experts. The aim is to ensure that AI systems make decisions and produce outputs that are consistent with desired ethical and societal standards (Dafoe, 2018). Value specification is particularly important in overcoming reward corruption and measuring and minimising extreme side effects, which are key challenges in ensuring the safe operation of ML systems.

3.2. Review of Safety Assessment Frameworks

In this section, we will explore various safety frameworks related to AI in aviation. These frameworks include:

- the EUROCONTROL-EASA FLY AI Report;
- the Single European Sky ATM Research (SESAR) Safety Reference Material;

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- the EASA Opinion 01/2020 that establishes a high-level regulatory framework for U-space;
- the US Federal Aviation Authority (FAA) - FAA AC 120-92B guideline;
- the Aviation Risk Management Solutions (ARMS) Methodology for operational risk assessment in aviation organisations;

This section also includes the presentation of two techniques:

- the Human Hazard and Operability Study (HAZOP);
- the Systems Theoretic Process Analysis (STPA).

3.2.1. Frameworks

EUROCONTROL- EASA The FLY AI Report Demystifying and Accelerating AI in Aviation

The European Aviation/ATM AI High-Level Group (EAAI HLG), composed of key representatives from various aviation sectors, has developed the EUROCONTROL EASA FLY AI report (EUROCONTROL, 2020). This report aims to demystify and accelerate the adoption of AI in aviation, encompassing all aspects of Air Traffic Management (ATM), including U-Space and avionics.

The report includes recommendations to create a federated AI infrastructure, accelerate AI deployment in non-safety-critical areas, conduct more AI research and development for safety-critical aviation operations, foster an AI culture through training and upskilling, establish partnerships with Digital Innovation Hubs and AI specialists, and promote knowledge sharing and communication.

The report takes into consideration the development of the EUROCAE WG 114/SAE G34 group on artificial intelligence. This group aims to adapt certification and approval frameworks for AI-based applications, covering both on-board certified systems and ATM/ANS AI-based applications/services. The certification/approval process being discussed includes features such as learning assurance, formal methods, testing, explanation, licensing, in-service experience, and online learning assurance, based on the product's requirements and certification strategy.

AI development and deployment present new areas of focus, particularly in terms of safety strategy and addressing emerging safety issues. Certification/approval of AI-based solutions involves numerous complex factors, including automation/autonomy levels, software assurance, liability, HF, trust, ethics, cybersecurity, training, licensing, data quality management, and verification/validation processes. Ensuring reliable behaviour and mitigating potential failures of the AI-integrated ATM/ANS functional system pose significant challenges, requiring the understanding and explanation of AI reasoning.

To address these concerns, the report emphasises the need for new processes, procedures, and tools to verify algorithms and AI solutions, both on-premise and in the cloud. This includes proper training, integration, maintenance, and the prevention of unintended side effects. As the aviation industry progresses with AI integration, understanding, validating, and demonstrating AI reasoning become vital aspects for achieving safe and efficient AI implementations in complex aviation systems.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

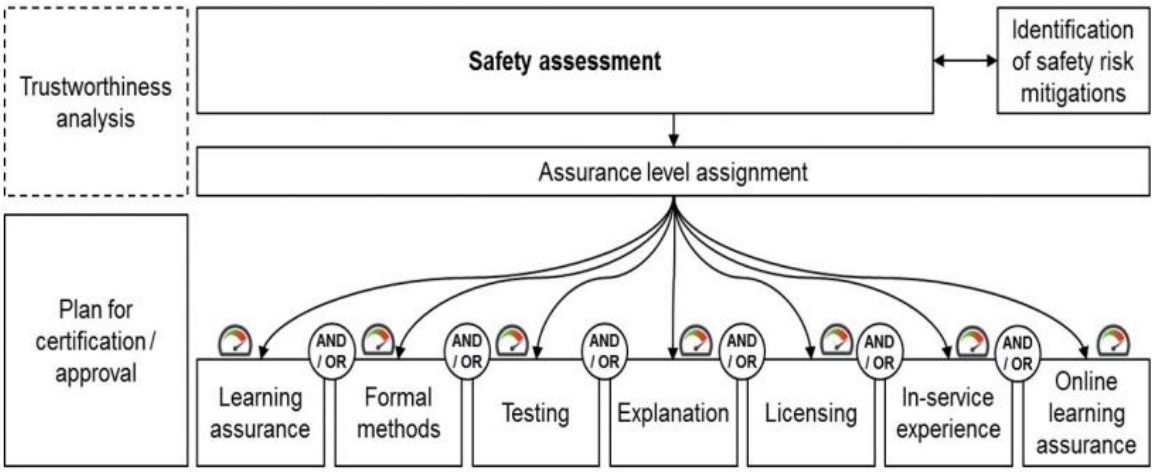

Figure 6.The EAAI HLG proposal for the future process for AI-based products

SESAR Safety Reference Material

SESAR Safety Reference Material is a comprehensive resource that plays a critical role in enhancing safety within the aviation industry (EUROCONTROL, 2018). Developed as part of the SESAR initiative, which aims to modernise and harmonise air traffic management in Europe, this reference material provides a standardised framework for addressing safety-related challenges and promoting a proactive safety culture.

The SESAR Safety Reference Material encompasses a wide range of documentation, guidelines, and best practices that cover various aspects of aviation safety. It serves as a central repository of knowledge, offering practical guidance and insights to aviation stakeholders, including air navigation service providers, airports, airlines, regulators, and industry experts.

One of the key objectives of the SESAR Safety Reference Material is to foster harmonisation and consistency in safety practices across Europe. The reference material addresses a wide range of safety topics, including risk assessment and management, safety performance monitoring, safety culture, HF, safety data analysis, incident reporting, and safety regulatory frameworks. Furthermore, the SESAR Safety Reference Material serves as a foundation for the development and implementation of safety management systems (SMS) across the European aviation community. It provides guidance on the establishment, maintenance, and continuous improvement of SMS, helping organisations to identify and mitigate risks, enhance safety performance, and comply with regulatory requirements.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

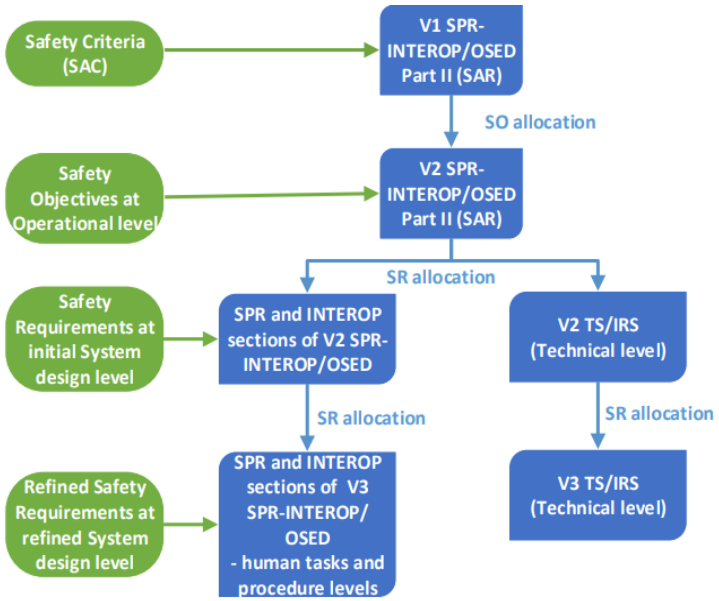

Figure 7.The relationship between the key SESAR formal deliverables and the Safety Requirements

The Safety Requirements are crucial design features that ensure the proper functioning of a functional system. The verification of these requirements is essential for meeting safety objectives and criteria. At the initial system design level, the Safety Requirements are derived from the allocation of safety objectives to the elements of the system. The SESAR SRM mandates that the initial Safety Requirements should cover the equipment, procedures, human and airspace elements of the system, including both success and failure approaches.

EASA Opinion 01/2020: High-level regulatory framework for the U-space

EASA's Opinion 01/2020 offers guidance and regulations for unmanned systems, which can be relevant to the use of AI in aviation due to the potential integration of AI-powered systems in unmanned aircraft. Unmanned aircraft and AI-powered systems have proven to be transformative technologies with applications spanning diverse domains. The objective of EASA Opinion 01/2020 is to establish and harmonise the necessary conditions for safe manned and unmanned aircraft operations in U-space airspace (EASA, 2020). The aim is to prevent aircraft collisions and mitigate air and ground risks. To achieve this, the U-space regulatory framework should provide clear and straightforward rules that enable safe aircraft operations across all areas and types of unmanned operations.

Due to insufficient data to conduct a comprehensive quantitative safety risk assessment, EASA will employ a general qualitative approach to assess the safety risks associated with the analysed options in this impact assessment. Since there is limited experience with the implementation of the proposed basic U-space services, the impact assessment is based on qualitative considerations. EASA has drawn inspiration from the approach taken by the Joint Authorities for Rulemaking on Unmanned Systems (JARUS) in developing the air risk model, which led to Annex C and Annex D of the Specific Operations Risk

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Assessment (SORA). The framework involves a structured process that considers various aspects of the operation, including the environment, the operational characteristics of the UAS, and the potential consequences of accidents. This model has been adopted by EASA as an Acceptable Mean of Compliance (AMC) to Commission Implementing Regulation (EU) 2019/947. In figure 8 an example of the SORA mitigation process to Air-Conflict is presented. By taking inspiration from the SORA framework, organisations can systematically analyse operational risks, tailor their implementations to specific contexts, and ensure safe and efficient integration.

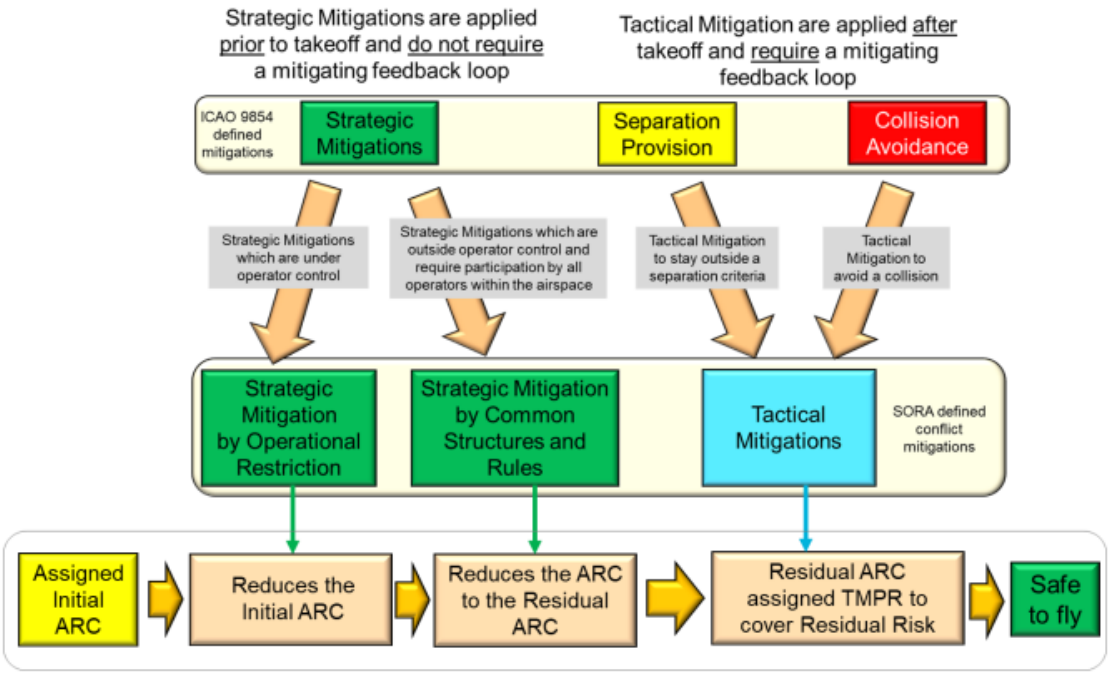

Figure 8. SORA Air-Conflict Mitigation Process

FAA AC 120-92B

Although the FAA Advisory Circular AC 120-92B does not pertain to AI in aviation, the FAA Advisory Circular AC 120-92B serves as a crucial document providing guidelines for the implementation of SMS within aviation organisations (Federal Aviation Administration, 2015). Its main objective is to enhance aviation safety through a structured approach to identifying and managing safety risks.

In particular, the advisory circular is a comprehensive approach to safety management, encompassing all sectors of the aviation industry, such as operators, maintenance organisations, and repair stations. This ensures consistent implementation of safety measures across the industry. Furthermore, the framework outlined in AC 120-92B allows organisations to adapt and tailor their safety management processes to suit their unique needs and circumstances. This flexibility facilitates the successful integration of SMS within diverse aviation environments.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

The document also introduces an important improvement known as the Safety Issue Risk Assessment (SIRA) method. This method addresses a limitation of the traditional severity x likelihood formula by incorporating the consideration of barriers, referred to as Risk Controls. The SIRA method incorporates four factors for risk assessment: the frequency/probability of the triggering event, the effectiveness of avoidance barriers, the effectiveness of recovery barriers, and the severity of potential accident outcomes.

The ARMS Methodology for Operational Risk Assessment in Aviation Organisations Developed by the ARMS Working Group, 2007-2010

The ARMS Methodology for Operational Risk Assessment in Aviation Organizations is a valuable and systematic tool developed through collaborative efforts between 2007 and 2010 (ARMS Working Group, 2010). This methodology enables aviation organisations to identify, assess, and mitigate operational risks in a structured manner. By considering various risk factors such as human performance, organisational processes, technology, and external influences, the ARMS Methodology provides a comprehensive framework for proactive risk management.

The ARMS methodology has been developed for Flight Safety risks. However, the working group believes that the methodology could easily be adapted for other types of risks. The ARMS methodology links with the following elements of the International Civil Aviation Organisation (ICAO) SMS framework: risk assessment (and mitigation), safety performance monitoring and measurement, and management of change.

The ARMS Methodology aligns with international standards and best practices in aviation safety, ensuring that organisations can meet global industry expectations. The versatility of the ARMS Methodology is another noteworthy aspect. It can be tailored to specific organisational requirements and is not limited to flying organisations alone. Maintenance Repair Organizations (MRO), Air Traffic Control (ATC), and airport operators can also benefit from its application. This adaptability makes it a valuable resource for various sectors within the aviation industry.

The methodology involves a multi-step process, starting with hazard identification through the collection and analysis of operational safety data. The Event Risk Classification (ERC) process allows for the quick estimation of the inherent risk in events, providing a risk class and numerical value for further analysis. The safety event data is stored in a database for future reference and analysis. Historical events are extrapolated to estimate the risk they posed at the time, taking into account the barriers that prevented them from becoming accidents.

Data analysis focuses on identifying safety issues that affect current operations. These issues undergo risk assessment using the SIRA technique. The SIRA calculates risk based on four factors: prevention, avoidance, recovery, and minimization of losses. This comprehensive approach includes risk controls (barriers) within the risk assessment, providing a holistic view of risk. The output of the SIRA is a risk value for each identified safety issue.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Regular analysis of the safety database helps detect adverse trends and monitor the effectiveness of previous risk reduction actions. Urgent issues identified through data analysis or prompted by events are addressed promptly without formal risk assessment. However, they should eventually undergo a formal SIRA to be measured and tracked in the risk register.

Systems Theoretic Process Analysis (STPA)

Systems Theoretic Process Analysis (STPA) is a robust safety analysis methodology that emerged as a response to the limitations of traditional hazard analysis techniques in complex socio-technical systems. STPA provides a holistic framework for identifying and mitigating safety hazards by delving into the underlying causal relationships and interactions within a system. It offers a proactive approach that focuses on understanding system behaviour, interactions, and dependencies, aiming to prevent hazards from materialising rather than merely responding to their consequences.

STPA is built on the foundation of systems theory, which recognizes that accidents are seldom the result of isolated incidents, but rather the consequence of complex interactions. STPA centres around the notion of control loops, sequences of actions and system responses that encompass various stages of control and feedback mechanisms. This concept helps unveil potential unsafe interactions between system elements. Causal factors play a central role in STPA. The methodology examines control actions, feedback loops, and information flows to uncover potential hazards rooted in complex interactions. The integration of safety constraints is a crucial aspect of STPA. Safety constraints define acceptable and unacceptable system behaviours, guiding the identification of potential hazards and the formulation of necessary safety requirements.

The process of STPA unfolds through a series of interlinked steps:

- **System definition**: The analysis begins with a comprehensive delineation of the system's boundaries, including the identification of components and interfaces. This provides a holistic perspective to identify potential interactions that could lead to hazards.
- **Control structure identification**: STPA involves recognizing control loops within the system, encompassing control actions, feedback mechanisms, and information flows. This unravels the intricate web of control relationships that govern system behaviour.
- **Hazard analysis**: Central to STPA is the identification of potential hazards. By scrutinising interactions within control loops, STPA identifies causal factors that could contribute to hazardous outcomes. These factors may include erroneous feedback or unforeseen control actions.
- **Causal analysis**: The analysis delves into how identified causal factors might lead to hazardous scenarios. This entails tracing the paths from causal factors to potential hazardous outcomes through the complex network of control loops.
- **Safety constraints formulation**: Safety constraints are introduced to guide the behaviour of the system within safe limits. These constraints operate to prevent unsafe interactions and ensure compliance with safety requirements.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- **Control actions and recommendations**: The culmination of the analysis results in the formulation of control actions and recommendations. These measures are designed to mitigate or prevent the identified hazards, which could involve modifying control loops or introducing safeguards.

HAZOP: Human Hazard and Operability Study

The Hazard and Operability (HAZOP) study is a structured and systematic examination of a planned or existing process or operation aimed at identifying and evaluating problems that may pose risks to personnel or equipment or hinder efficient operation. Initially developed to analyse chemical process systems, the HAZOP technique has since been extended to other types of systems, complex operations, and even software systems (Chemical Industry Safety & Health Council, 1977).

Conducted by a multidisciplinary team (HAZOP team) during a series of meetings, the HAZOP study employs qualitative techniques based on guidewords. Ideally, the study should be carried out as early as possible in the design phase to have a significant influence on the design. However, to conduct a HAZOP, a relatively complete design is required. As a compromise, the HAZOP is typically performed as a final check once the detailed design has been completed.

Moreover, a HAZOP study can also be conducted on an existing facility to identify necessary modifications that can reduce risk and improve operability. In this capacity, HAZOP studies are used at various stages of a project's life cycle, including the initial concept stage when design drawings are available.

By applying the HAZOP technique, industries can proactively assess and address potential hazards and operational issues, leading to improved safety, reliability, and overall performance of complex systems and processes. The collaborative and comprehensive nature of HAZOP studies ensures that a wide range of perspectives and expertise are considered, contributing to a thorough analysis and effective risk mitigation strategies. As a result, the HAZOP technique continues to be a cornerstone of modern safety management systems across various industries, playing a crucial role in safeguarding personnel, equipment, and the environment while facilitating efficient and trouble-free operations.

3.3. Proposed and Adapted Safety Assessment for IAs

Building upon ALTAI assessment (AI HLEG, 2020) questions (Annex B) and the three dimensions derived from the SESAR safety reference material (EUROCONTROL, 2018), the HAIKU proposal is to consider questions related to general safety, accuracy, reliability, and traceability. The questions deriving from the ALTAI assessment list can be categorised into three key areas stemming from the SESAR safety reference material:

- **Initial Design analysis under normal operations**: under normal operations, it is crucial to define risks, risk metrics, and risk levels specific to each UC of the AI system. Continuous measurement and assessment of risks, along with informing end-users and subjects about

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

potential risks, are essential steps to ensure safety. Additionally, the quality and representativeness of the data used to train the AI system should be ensured, and steps should be taken to monitor and document the system's accuracy.

- **Initial Design analysis considering abnormal conditions**: identifying potential threats to the AI system, such as design faults, technical faults, and environmental threats, is vital. Assessing the risk of malicious use, misuse, or inappropriate use of the AI system, as well as defining safety criticality levels, helps in addressing possible consequences. The dependency of critical AI system decisions on stable and reliable behaviour should be evaluated, aligning reliability and testing requirements accordingly. Furthermore, the impact of low accuracy and the system's ability to invalidate its training data or assumptions should be considered.
- **Initial Design analysis in faulted conditions**: in faulted conditions, planning fault tolerance and evaluating the technical robustness and safety of the AI system after changes are essential. Verification and validation methods, along with documentation practices like logging, contribute to assessing reliability and reproducibility. Failsafe fallback plans should be tested and implemented to handle errors, while procedures for handling low confidence score results should be in place. Continual learning, if utilised, requires considerations for potential negative consequences from the system learning novel or unusual methods.

These categories represent different aspects of the development and evaluation of an AI system's safety and reliability.

The assessment relies on a series of questions designed to help the UC owners pinpoint and identify the safety requirements.

Initial Design analysis under normal operations:

1. Did you define risks, risk metrics, and risk levels of the IA system in the specific UC?

2. Did you define clear risk mitigation strategies to address the identified safety risks?

3. Did you put in place measures to continuously assess the quality of the input data to the IA system?

4. Did you put in place a series of steps to monitor and document the IA system's accuracy?

5. Did you put in place measures to continuously assess the quality of the output(s) of the IA system?

Initial Design analysis considering abnormal conditions:

1. Did you identify the risk of possible misuse or inappropriate use of the IA system? If yes, did you identify the possible consequences?

2. Did you identify the potential impact of the IA system's failures or malfunctions on human safety?

3. Did you define safety critical levels of the possible consequences of faults or misuse of the IA system in terms of severity and likelihood?

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

4. Could a low level of accuracy of the IA system result in critical, adversarial, or damaging consequences?

5. Could the IA system cause critical, adversarial, or damaging consequences (e.g., pertaining to human safety) in case of low reliability and/or reproducibility?

6. Did you identify whether specific contexts or conditions need to be taken into account to ensure accuracy and reliability?

Initial Design analysis in faulted conditions:

1. Did you evaluate the robustness and reliability of the AI system under different operating conditions and potential failure scenarios?

2. Did you develop a mechanism to evaluate when the IA system has been changed to merit a new review of its technical robustness and safety?

3. Did you put in place tested failsafe fallback plans to address IA system errors of whatever origin and put governance procedures in place to trigger them?

4. Did you put in place a proper procedure for handling the cases where the IA system yields results with a low confidence score?

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

4. Security Methods and Assessments Frameworks for IAs

4.1. Emerging Security issues concerning AI in Aviation

Transport is a sector that is highly targeted by cyber threats, such as denial of service (DoS), data theft, malware, phishing, software manipulation, unauthorised access, destructive attacks, masquerading of identity, abuse of access privileges, social engineering, defacement, eavesdropping, misuse of assets, and hardware manipulation (EC, 2020).

The integration of Information and Communication Technology (ICT) tools into mechanical devices used in the aviation industry has raised cybersecurity concerns due to various factors. Sharing information among stakeholders, often using disparate systems, has led to legacy systems with diverse communication protocols. Balancing new tech integration with securing vulnerable legacy systems is challenging. The convergence of OT and IT systems, increased use of off-the-shelf products, and expanding connectivity amplify security issues. The rise of IoT, digital towers, and new aircraft concepts further complicates matters. Inherent vulnerabilities in aviation protocols contribute to risks. Additionally, a global shortage of cybersecurity experts and competent inspectors exacerbates the situation (Hawley, 2022). As the level of integration of these systems increases, the inherent vulnerabilities in the software that drive them also increase. Given the crucial role of cyber-technologies in the operational integrity of the aviation industry, the International Air Transport Association (IATA) has been relied upon to provide guidance to improve and update cybersecurity regulations, standards, and principles for the entire avionics system, including air-traffic controls, airlines, and airports. The business objectives for these improvements include enhancing ground, air, and space operations, as well as customer services such as ticket bookings, in-flight entertainment systems, flight check-in and -out, and security screening of passengers, among others.

When it comes to securing information in Intelligent Transportation Systems (ITS), the Confidentiality, Integrity, and Availability (CIA) model is often used as a framework. In addition to the CIA dimensions, authentication and identification are also important aspects to consider in ITS security classification (Hahn, Munir, & Behzadan, 2019).

- Confidentiality is a key concern in ITS as it forms a requirement for enabling secure communication between devices and parties without disclosing information to unauthorised parties. Confidentiality supports other processes, procedures, and security training to enable secure communication. Encryption mechanisms are commonly used to ensure confidentiality, and recent research has explored alternative methods like steganography and covert channels to conceal information from malicious actors.
- Ensuring data integrity is critical for the proper functioning of ITS components such as vehicles, infrastructure, and traffic controllers. Malicious attacks can alter messages between vehicles

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

and other ITS components, resulting in incorrect information being used for various calculations and decision-making.

- The availability of ITS components is also crucial to maintaining traveller safety, as threats and attacks can have critical consequences due to the real-time nature of many ITS operations.
- Authentication and identification are essential in ITS to ensure secure communication and data transfer. Message Authentication Codes (MACs) and challenge-response protocols are commonly used for verification but can introduce additional computational overhead.

Of all the attacks studied, the majority (71%) focused on stealing login details, such as administrative passwords, and malicious hacking to gain unauthorised access to the IT infrastructure. DoS attacks, which compromise availability, were the second most common attack (25%), followed by attacks that corrupt the integrity of files, either by intercepting them while in transit or at rest (4%) (Ukwandu, Ben-Farah, & Hindy, 2022).

Figure 9 shows the assessment of cyber-attacks by type, with malicious hacking activities topping the list at 26%. This type of attack aims to gain unauthorised access using password cracking techniques, such as brute force or dictionary attacks. Data breach and ransomware attacks were the second most common type of attack at 14% each, followed by attacks related to phishing and malware at 11% each. Cyber-incidents classified as human error, bot attacks, worms, and DDoS were the rarest, each accounting for 4% of all attacks (Ukwandu, Ben-Farah, & Hindy, 2022).

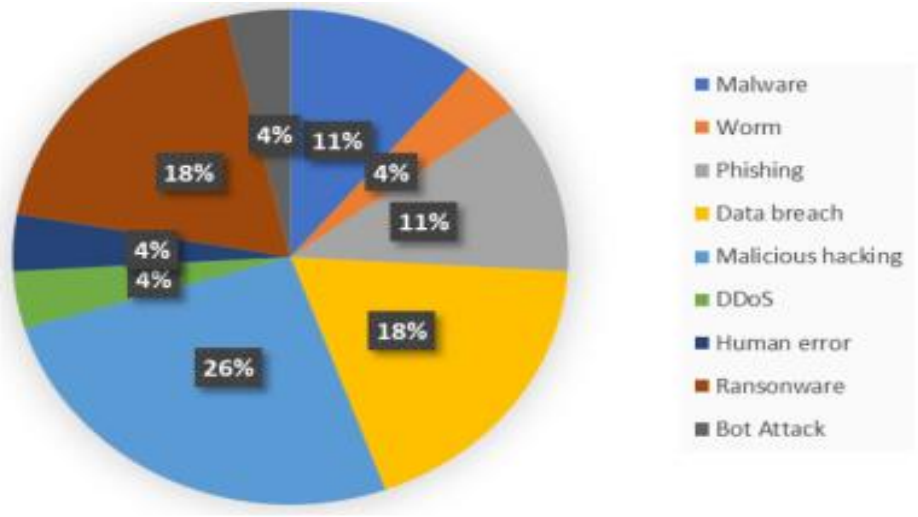

Figure 9.Cyber-Attacks by type. Source: Ukwandu,et al. (2022). Cyber-security challenges in the aviation industry: A review of current and future trends.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

The increasing levels of automation through the integration of operational systems have created new attack surfaces, necessitating the revision of existing cyber-security implementations, and the assessment of the ramifications of evolving threats (Koroniotis, Moustafa, & Sitnikova, 2019).

The figure 10 below presents a list of high-level categorizations of threats based on the European Union Agency for Cybersecurity (ENISA) threat taxonomy (ENISA, 2020).

Figure 10.Threat Taxonomy. Source ENISA. (2020). AI Cybersecurity Challenges.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

4.2. Review of Security Assessment Frameworks

In the context of ensuring the security of AI in the aviation sector, we will now delve into various standards, roadmaps and frameworks that have been developed to address the cybersecurity challenges. We will consider these frameworks:

- the ICAO Cybersecurity Strategy, which aims to enhance the cybersecurity posture of the aviation industry.
- the ENISA Securing Machine Learning Algorithms framework, which focuses on securing the integrity and confidentiality of AI algorithms.
- the AISecurity Framework, which provides comprehensive guidelines for ensuring the security of AI systems. the Securing Smart Airports framework developed by ENISA, which addresses the unique security challenges in airport environments.
- the NIST Cybersecurity Framework, offering a comprehensive approach to managing and mitigating cybersecurity risks. the EUROCONTROL ATM Cybersecurity Maturity Model Level 1 and the Security Risk Assessment methodology for SESAR 2020, both of which provide frameworks for assessing and enhancing cybersecurity in the aviation sector.

4.2.1 Standards and Regulations

The ISO/IEC 27000 series encompasses a set of international standards and guidelines developed by ISO and the IEC that pertain to information security management systems (ISMS). This series provides a comprehensive framework for organisations to establish, implement, maintain, and continually improve their information security posture. At its core is ISO/IEC 27001, which outlines the requirements for designing and operating an ISMS, encompassing processes such as risk assessment, security controls implementation, and ongoing monitoring. The complementary standards within the ISO/IEC 2700X series provide guidance on specific aspects of information security, ranging from risk management (ISO/IEC 27005) and controls implementation (ISO/IEC 27002) to security governance (ISO/IEC 27003) and incident response (ISO/IEC 27035). The ISO/IEC 2700X series is widely recognized for its role in assisting organisations across industries in safeguarding their sensitive information, ensuring data confidentiality, integrity, and availability, and fostering a culture of continuous security improvement.

Commission Delegated Regulation (EU) 2022/1645, issued on 14 July 2022, assumes a pivotal role in the enhancement of aviation safety protocols within the European Union. This regulation intricately outlines the application of requirements pertaining to the management of information security risks that possess the potential to impact the integrity of aviation safety, as mandated by Regulation (EU) 2018/1139 of the European Parliament and of the Council. Through its comprehensive provisions, this regulation establishes a stringent framework for addressing information security risks within aviation operations. By meticulously detailing protocols for risk identification, assessment, mitigation, and ongoing management, it ensures that information systems crucial to aviation safety remain fortified

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

against threats such as cyberattacks, unauthorised access, and data breaches. In doing so, Commission Delegated Regulation (EU) 2022/1645 emerges as an indispensable tool in bolstering the resilience of the aviation sector against evolving information security challenges.

Regulation (EU) 2023/203 delineates a comprehensive framework for organisations and competent authorities within the EU to effectively manage information security risks that could impact aviation safety. The regulation encompasses several distinct aspects. It outlines requirements to identify and address information security risks that have the potential to affect information and communication technology systems and data crucial for civil aviation purposes. The regulation mandates the detection of information security events and the identification of those that qualify as information security incidents with the potential to impact aviation safety. Furthermore, the regulation stipulates the necessity of robust response and recovery mechanisms to address and overcome such information security incidents.

This regulatory framework applies to a range of organisations, including maintenance organisations, continuing airworthiness management organisations (CAMOs), air operators, approved training organisations (ATOs), aircrew aero-medical centres, flight simulation training device (FSTD) operators, air traffic controller training organisations (ATCO TOs), and various service providers within the aviation domain.

Notably, the regulation extends its application to competent authorities, including the EASA, as well as to the competent authority responsible for aircraft maintenance licence issuance and oversight.

The inclusion of **"Acceptable Means of Compliance" (AMC) and "Guidance Material"** (GM) within the Articles of Regulations (EU) 2022/1645 and 2023/203 adds a crucial layer of practicality and clarity to the regulatory landscape. AMC, in the context of these regulations, delineates specific methodologies, processes, or practices that organisations can adopt to meet the regulatory standards effectively. It provides tangible steps and solutions, offering a clear path to compliance by addressing the "how" of implementing regulatory requirements.

Concurrently, GM serves as an invaluable companion to AMC, offering further insights, clarifications, and illustrative examples. GM expounds on the intent behind regulatory provisions, helping stakeholders understand the underlying principles and considerations. It provides additional context and best practices, assisting organisations in achieving a deeper comprehension of the regulations. This context, in turn, aids in tailoring AMC to specific operational scenarios, thereby optimising the application of regulatory requirements.

Collectively, AMC and GM work in tandem to ensure a balanced approach to compliance, offering both prescriptive and interpretive elements. This comprehensive approach not only simplifies the compliance process but also encourages a higher level of adherence by aligning operational practices with the overarching objectives of the regulations. AMC and GM thus enhance the efficacy of Regulations (EU) 2022/1645 and 2023/203 by providing pragmatic guidance and interpretive context,

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

fostering a harmonised and informed approach to information security risk management in aviation within the EU.

4.2.2 Strategies and roadmaps

International Civil Aviation Organization (ICAO) Cybersecurity Strategy

In 2018, the ICAO adopted its Cybersecurity Strategy as a response to the growing cybersecurity risks faced by the aviation industry (ICAO, 2019). The primary objective of this strategy is to provide guidance to ICAO in strengthening cybersecurity measures and ensuring the resilience of aviation systems and infrastructure.

The ICAO Cybersecurity Strategy is founded upon several key pillars. These include:

- International cooperation: Promoting collaboration and cooperation among nations, organisations, and industry stakeholders to collectively address cybersecurity challenges in civil aviation.
- Governance: Establishing effective governance structures and frameworks at national and international levels to facilitate the implementation of cybersecurity measures and ensure accountability.
- Effective legislation and regulations: Encouraging the development and enforcement of robust cybersecurity legislation and regulations that align with international standards and best practices.
- Cybersecurity policy: Formulating comprehensive cybersecurity policies that encompass risk management, incident response, information protection, and resilience planning.
- Information sharing: Facilitating the sharing of timely and relevant cybersecurity information and threat intelligence among stakeholders to enhance situational awareness and enable proactive defence measures.
- Incident management and emergency planning: Establishing protocols and frameworks for effective incident response, crisis management, and emergency planning in the face of cyber incidents.
- Capacity building, training, and cybersecurity culture: Promoting education, training, and awareness programs to enhance the cybersecurity skills and knowledge of aviation personnel and fostering a culture of cybersecurity throughout the industry.

The ICAO Cybersecurity Strategy aims to adopt a comprehensive and proactive approach to address cybersecurity risks in civil aviation. The strategy emphasises a risk-based approach to cybersecurity, focusing on threat intelligence, risk assessment, and risk management. It emphasises the need for collaboration and information-sharing among various stakeholders in the aviation industry, including regulatory bodies, airlines, airports, and service providers.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

ENISA Securing Machine Learning Algorithms

The ENISA report titled "Securing Machine Learning Algorithms" (ENISA, 2021) delves into the security risks associated with ML algorithms and provides a range of recommendations to ensure their robust security. The report emphasises the distinctive security challenges posed by ML algorithms, arising from their reliance on vast amounts of data and complex algorithms, which can be exploited by malicious entities to manipulate the system.

One of the key security risks identified in the report is data poisoning attacks, where attackers manipulate the training data to introduce biases or modify the algorithm's behaviour to achieve specific objectives. Additionally, model stealing attacks are highlighted, wherein attackers gain unauthorised access to and steal the ML model, which can then be utilised for making predictions or launching attacks against the system. Model evasion attacks, another security risk, involve manipulating the inputs to the ML model to evade detection or manipulate the output. Lastly, the report acknowledges the privacy risks associated with ML algorithms, as they often require access to sensitive data.

To effectively address these risks, the ENISA report puts forth several recommendations for securing ML algorithms. It suggests:

- the adoption of secure development practices and incorporating security considerations into the design of ML algorithms.
- the implementation of access controls to protect ML models and the associated data is also recommended.
- the use of secure communication protocols and encryption techniques to safeguard data in transit and at rest is emphasised.
- the regular monitoring and updating of ML models to detect and address security vulnerabilities is deemed essential.
- the implementation of privacy protection measures such as anonymization and data minimization techniques can help protect sensitive data.

One notable advantage of the report is its exclusive focus on ML, enabling the identification and mitigation of security risks specific to this domain. By addressing the unique challenges and offering practical recommendations, the report aims to enhance the security of ML algorithms and mitigate potential risks.

4.2.3 Frameworks

Artificial Intelligence Security Framework

The AI security framework (Jing, Wei, & Zhou, 2021) encompasses four dimensions: security goals, security capabilities, security technologies, and security management. These dimensions serve as guidelines for enterprises to construct an AI security protection system in a hierarchical manner. Setting

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

appropriate security goals serves as the foundational starting point for ensuring the security of AI applications. Security capabilities, on the other hand, act as the effective means to achieve these security goals, while security technologies and management provide support and practical implementation of these capabilities.

In the security assessment, the following components are described:

- Security Goals: A systematic analysis of security risks associated with AI and their underlying causes must be conducted. This analysis helps establish security requirements and goals for AI systems, considering six key aspects: application, function, data, decision-making, behaviour, and incidents.
- Security Capabilities: Considering the challenges involved in constructing security capabilities and optimising resource allocation, five AI security capabilities are proposed. These capabilities are designed to build upon each other, forming a progressive hierarchy. The capabilities include architecture security, which focuses on planning and designing secure AI applications. Passive defence involves deploying static security measures beyond the applications. Active defence strengthens AI security teams and promotes dynamic and adaptive security capabilities. Threat intelligence aids in acquiring and utilising AI security threat information to enhance security systems. Lastly, offence empowers enterprises to develop lawful offensive capabilities against malicious attackers targeting AI.
- Security Technologies: The core components of AI application construction, such as AI applications, algorithms, training data, and framework platforms, require robust security protection. This framework provides security technology means for protecting AI applications, algorithms, data, and platforms.
- Security Management: To ensure comprehensive security, enterprises should comply with national and industry-specific AI security laws, regulations, policies, ethical norms, and technical standards. This framework outlines the implementation requirements for enterprises regarding AI security organisation, personnel, and systems, taking into account the relevant guidelines and requirements.

According to the paper, these components do not represent an assessment process, they represent the components of the AI Security Framework. The paper significantly underscores the imperative of addressing the matter of artificial intelligence (AI) security with a heightened sense of urgency. The current absence of a coherent global strategy for ensuring the security of AI systems engenders two principal predicaments. Firstly, the deficiency of a comprehensive international framework impedes the formulation of robust regulatory protocols to govern AI security. Secondly, this absence acts as a deterrent to the unabated advancement of the AI industry on a global scale.

Securing Smart Airports ENISA

"Securing Smart Airports" is a publication issued by ENISA that offers comprehensive guidance on safeguarding smart airports against cyber threats (ENISA, 2016). The primary objective of the report is to

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

assist airport operators and relevant stakeholders in comprehending the risks and potential threats associated with smart airports and to devise robust cybersecurity strategies and measures.

The report furnishes a risk assessment framework intended to aid airport operators in identifying and evaluating potential risks and threats specific to their smart airport systems. Furthermore, it includes guidelines for formulating effective cybersecurity strategies encompassing the development of policies, procedures, and incident response plans.

Additionally, the report provides an exhaustive analysis of various security measures that can be implemented to mitigate risks and threats in the context of smart airports. These measures encompass network segmentation, access controls, encryption, intrusion detection and prevention systems, as well as security monitoring and logging mechanisms.

The security assessment outlined in the report underscores the vulnerability of smart airports to diverse cyber threats due to the intricate and interconnected nature of their systems, extensive employment of sensors and data analytics, and the involvement of numerous stakeholders. The identified threats encompass unauthorised system access and control, data breaches, malware and ransomware attacks, and DoS attacks.

To address these challenges, the report puts forth a series of recommendations to enhance the cybersecurity posture of smart airports. These recommendations include the development of a unified cybersecurity framework and guidelines, fostering collaboration and information sharing among stakeholders, and bolstering the security of legacy systems. Additionally, the report suggests conducting regular security assessments and testing of smart airport systems, incorporating security-by-design principles into the implementation of new technologies, and augmenting cybersecurity awareness and training initiatives for all stakeholders involved.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Attack scenario

Type of Attack	Asset affected
Network attack	• Baggage handling • ICS SCADA • Way-finding services
CRITICALITY	**LIKELIHOOD**
High with emphasis on operations but it could also escalate to safety.	Medium
CASCADING EFFECTS	**STAKEHOLDERS INVOLVED**
• Baggage handling systems • Computerised Maintenance Management Systems (CMMS) • Energy Management	• IT support services • Passengers • Baggage handling • Building and other maintenance
RECOVERY TIME AND EFFORTS	**GOOD PRACTICES**
System recovery and efforts depend on the time needed to identify the security flaw as well as isolate and block the attack. Due to the interconnection among systems and possible cascading effects this could require a significant effort from several of the stakeholders involved. Recovery time could be reduced by prioritising which services should be recovered first focusing on recovering the most relevant in the first instance.	• **GP13 – Integrate shutdown procedure / remote deactivation of capabilities for assets based on assets** • **GP 11 – Firewalls, network segmentation and defence in depth** • **GP 12 – Strong user authentication** • **GP 03 – Change default administrator credentials of devices** • **GP 01 – Intrusion Detection Systems (IDS)** • GP08 – Conduct vulnerability and penetration tests • GP06 – Operating systems updates and backups • GP 16 – Set up an information security management system and implement international standards • GP 19 – Establish an inventory of the information and information systems available • GP 22 – Conduct risk assessments • GP 23 – Create a risk register and monitor risk effectively • GP 24 – Perform continuous monitoring of information security • GP 25 – Manage risk according to international standards and a methodological approach • GP 35 – Provide specialised information security training • GP 38 – Develop a contingency plan • GP 42 – Provide incident response capabilities for airports • GP 43 – Train airport personnel in their incident response roles with respect to the information system • GP 45 – Track and document information system security incidents
CHALLENGES AND GAPS	
One of the key challenges in relation to SCADA is that cyber security good practices and countermeasures commonly applied to IT infrastructure have not been applied to ICS. Another challenge is related to the increasing interdependence and connection of SCADA with other airport systems. The degree of interconnections among systems increase the number of vectors attacks while opening up back doors to connected systems. This increase in complexity and functionality requires an enhanced approach to cyber security focusing on holistic assessments and planning (see Gap 8). ENISA has released several guidelines on ICS SCADA security[64].	

Figure 11.Example of an Attack Scenario. Source: ENISA (2016), Securing Smart Airports.

NIST Cybersecurity Framework

The NIST Cybersecurity Framework, although not specific to the aviation industry, offers comprehensive guidance on how organisations can effectively manage and mitigate cybersecurity risks. It comprises five core functions: Identify, Protect, Detect, Respond, and Recover (NIST, 2018).

- **IDENTIFY (ID)** – This function helps determine the existing cybersecurity risk for the organisation. Understanding its assets (such as data, hardware, software, systems, facilities, services, and people) and the related cybersecurity risks allows the organisation to prioritise efforts

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

in line with its risk management strategy and mission needs identified under GOVERN. This function also involves identifying necessary improvements for the organisation's policies, processes, procedures, and practices that support cybersecurity risk management, informing actions under all six functions.

- **PROTECT (PR)** – This function involves using safeguards to prevent or lessen cybersecurity risk. Once assets and risks are identified and prioritised, PROTECT aids in securing those assets to minimise the likelihood and impact of adverse cybersecurity events. This function covers outcomes such as awareness and training, data security, identity management, authentication, access control, platform security, and technology infrastructure resilience.
- **DETECT (DE)** – The DETECT function focuses on locating and analysing potential cybersecurity attacks and compromises. It enables the timely identification and analysis of anomalies, indicators of compromise, and other potentially negative cybersecurity events that might suggest ongoing cybersecurity attacks and incidents.
- **RESPOND (RS)** – The RESPOND function is about taking action in response to detected cybersecurity incidents. It aims at containing the impact of such incidents. Outcomes within this function encompass incident management, analysis, mitigation, reporting, and communication.
- **RECOVER (RC)** – The RECOVER function is dedicated to restoring assets and operations that have been affected by a cybersecurity incident. It aims to promptly reinstate normal operations, reduce the impact of cybersecurity incidents, and facilitate appropriate communication during recovery efforts.

In the updated version of the NIST Cybersecurity Framework (CSF) 2.0, released on August 8, 2023, a notable enhancement has been introduced: the inclusion of an additional, 6th, core function known as **GOVERN (GV).** This function involves establishing and overseeing the organisation's cybersecurity risk management strategy, expectations, and policy. The GOVERN function has a cross-cutting nature, offering insights to guide how an organisation will achieve and prioritise the outcomes of the other five functions within its mission and according to stakeholder expectations. Governance activities are crucial for integrating cybersecurity into the broader enterprise risk management strategy of an organisation. GOVERN encompasses understanding organisational context, setting up cybersecurity strategy and managing cybersecurity supply chain risk, defining roles, responsibilities, and authorities, establishing policies, processes, and procedures, as well as overseeing cybersecurity strategy.

These core functions within the framework work together to support organisations in managing cybersecurity risks by organising information, facilitating risk management decisions, addressing threats, and promoting continuous improvement. Additionally, the framework aligns with existing incident management methodologies and demonstrates the impact of cybersecurity investments. For instance, investments in planning and exercises can enhance the organisation's ability to respond and recover swiftly, thereby minimising disruptions to service delivery.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

The framework further utilises categories and subcategories to provide subdivisions and specific outcomes related to technical and management activities. These categories, such as "Asset Management" and "Detection Processes," aid in achieving the desired cybersecurity outcomes within each function. Informative References, drawn from widely recognized standards, guidelines, and practices, are also included to illustrate approaches for achieving the outcomes associated with each subcategory.

EUROCONTROL ATM Cybersecurity Maturity Model Level 1

The EUROCONTROL ATM Cybersecurity Maturity Model Level 1 is a framework that outlines a range of capabilities expected in an organisation with an effective approach to cybersecurity (EUROCONTROL, 2017). It describes activities and processes at different levels of maturity, allowing organisations to assess their cybersecurity practices and compare them against the described levels of each capability.

Originally developed for the Air Traffic Management (ATM) industry by the Network Manager (NM) and Air Navigation Services Providers (ANSPs), this model primarily applies to ATM but may also have relevance beyond this specific context. It combines elements from various existing standards and guidelines, tailored to the needs of the ATM industry. Notably, the model does not impose new requirements on ATM stakeholders but focuses on capabilities and processes, allowing organisations to determine their specific requirements.

One of the advantages of using this model is its ability to provide a consolidated snapshot of critical information that may not otherwise be available in a single document. The model is based on the NIST CSF and incorporates elements from ISO 27001. The NIST CSF was chosen due to its practicality and wide adoption. The CSF's tiers serve as a starting point, connected to a broader target-setting process that considers an organisation's business objectives, threat/risk environment, and requirements and controls. Additionally, the model emphasises the importance of leadership, governance, and HF in cybersecurity.

The model consists of two levels of detail. The high-level model, described here, is a simplified maturity assessment that includes 13 capabilities and predefined answers indicating the meaning of each maturity level (e.g., Level 3, Level 4). This allows for a relatively quick assessment by an individual with a comprehensive overview of the organisation. However, due to the comprehensive nature of the maturity model, assessments from various perspectives, including technical and operational personnel, are often necessary. Differences in perspectives can lead to valuable discussions and insights during the assessment process.

The selected capabilities in the model are deemed the most important and relevant for the ATM industry. A provided form assists in completing the assessment, including providing rationale and evidence. An organisation must fulfil all elements of one level before progressing to the next level. If any earlier elements are missing, the organisation must assess itself at the lowest level where it fulfils

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

all elements. Each capability also includes a set of probing questions that can be used for further exploration through open-ended questioning.

The model can be applied to suppliers by the suppliers themselves, their customers, or third parties. It can also be adapted as needed to fit the specific requirements of an organisation, serving as a flexible framework for implementation.

Security Risk Assessment methodology for SESAR 2020

Within the framework of the SESAR program, safeguarding the security of air traffic management (ATM) systems is of utmost importance. As an integral part of the SESAR 2020 initiative, an all-encompassing Security Risk Assessment (SRA) methodology has been devised to identify, evaluate, and mitigate potential security risks specific to the ATM domain (SESAR Joint Undertaking, 2017). This methodology serves as a structured approach, empowering stakeholders to proactively address and manage security threats and vulnerabilities (figure 13).

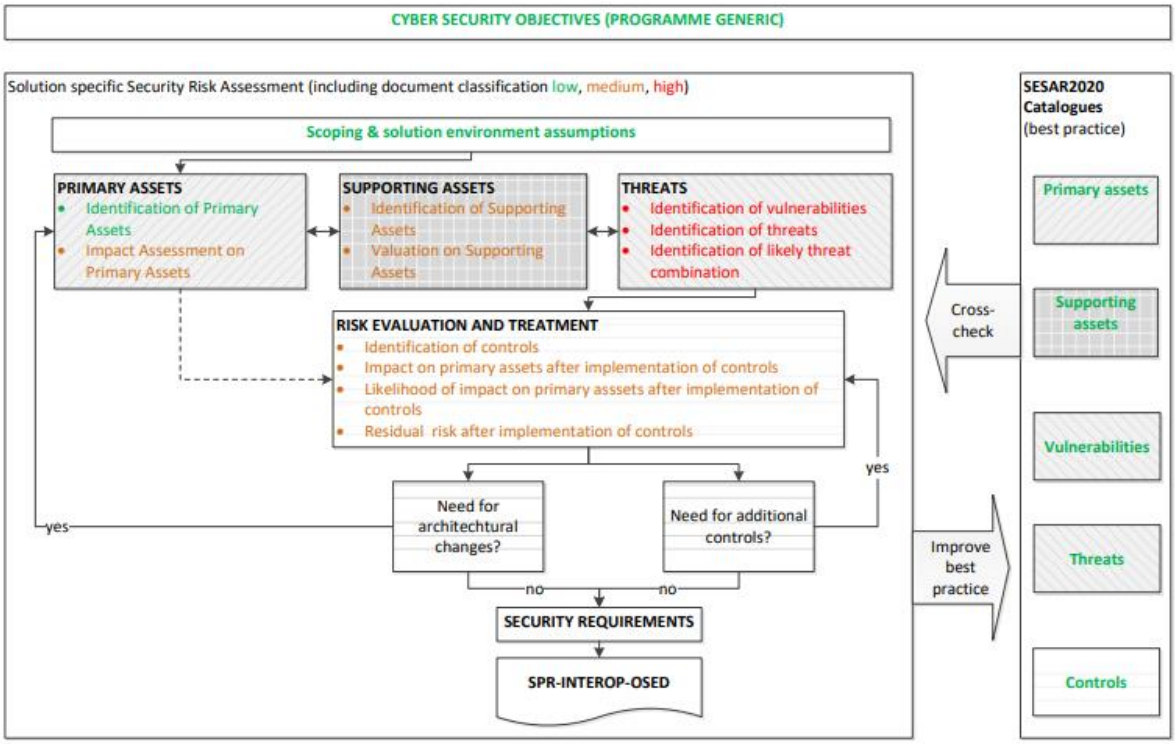

Figure 12.The SecRAM methodology.

The Security Risk Assessment Methodology (SecRAM) encompasses several essential steps, including:

- Defining the scope of the risk assessment: This involves describing the roles, equipment, and systems involved in the assessment, as well as identifying dependencies on other systems and infrastructure. Specialist operational or design knowledge of the system is required for this step.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- Identifying assets and evaluating potential impacts: Assets are the targets of security attacks, and assessing potential impacts involves evaluating the harm that would result from compromising each asset through an attack.
- Identifying vulnerabilities, threats, and likely threat combinations: This step entails identifying possible or credible threat sources and associated threat scenarios. Each threat is linked to vulnerabilities within the system that could be exploited by an attacker. The aim is to gain insights into all potential routes (threat scenarios) that a threat may employ to access an asset.
- Identifying a set of security controls: This involves selecting security controls associated with supporting assets and minimising the impact on primary assets. The impact on primary assets is evaluated after implementing the security controls. The initial risk evaluation may be conducted using controls already in operation and generic organisational controls to focus solely on identifying controls that mitigate risks not meeting the program's generic security objectives.
- Determining the likelihood of the impact on primary assets: This step involves assessing the likelihood of the identified impacts occurring on primary assets.
- Assessing the security risk: Based on the likelihood and potential impact, the security risk is assessed, taking into consideration the established Cyber Security Objectives.
- Determining the acceptability of the security risk: This involves comparing the assessed security risk against the acceptable level set by the Cyber Security Objectives. If the risk exceeds the acceptable level, further analysis is necessary to identify improvements to the current situation.

4.3. Proposed and Adapted Security Assessment for Intelligent Assistant

To assess the security risks associated with an AI system, we will map the ALTAI questions (Annex B) to the SecRAM methodology. In the context of security risk assessment, the ALTAI questions specifically focused on Resilience to Attack and Security, Privacy, and Data Governance can be used. These questions enable the evaluation of security risks associated with an AI system. It is important to note that a risk assessment must be carried out on the underlying system prior to addressing AI-specific issues in order to ground the study on a solid general basis. By considering these ALTAI questions and using the SecRAM model, developers and users of AI systems can effectively identify and address security vulnerabilities.

Identification of Primary and Supporting Assets:

1. Did you identify the primary and supporting assets that could be affected in the event of outages, attacks, misuse, or threats associated with the IA?

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Identification of Threats/Vulnerabilities/:

2. Did you define potential forms of attacks to which the IA system could be vulnerable?

3. Did you define the potential adversarial, critical or damaging effects in case of outages, attacks, misuse or threats associated with the IA?

4. Did you define how exposed the IA system is to cyber-attacks?

5. Did you consider the impact of the IA system on the right to privacy, the right to physical, mental, and/or moral integrity, and the right to data protection?

Identification of Controls:

6. Did you evaluate the IA system's resilience against adversarial attacks or manipulation attempts?

7. Did you consider robust authentication and access control mechanisms to ensure only authorised users can interact with the IA system?

8. Did you identify mechanisms to detect and mitigate potential privacy breaches or leaks of sensitive information by the IA system?

9. Did you identify monitoring mechanisms to detect and respond to security incidents or breaches involving the IA system?

10. Did you identify measures to ensure the integrity, robustness, and overall security of the IA system against potential attacks over its lifecycle?

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

5. HF Methods and Assessment Frameworks for IAs

5.1. Emerging HF Issues concerning Aviation

Numerous challenges remain unresolved in the interaction between human and AI systems. These challenges include the impact of AI systems on pilots' mental workload and situation awareness, as well as their levels of acceptance, trust, and reliance on such systems. Additionally, potential changes in human behaviour due to automation, the required skills, and the role of humans in emergencies are all challenges that require attention (Sheridan & Parasuraman, 2005). The level of supervisory, control and cooperation, between human and AI systems also needs clarification, along with the minimum time required for human to resume control when instructed by AI systems[6]. between human and AI systems also needs clarification, along with the minimum time required for human to resume control when instructed by AI systems. The majority of the reviewed literature predominantly draws insights from the application of automation in the automotive industry, yet these findings hold potential applicability within the aviation sector as well.

Problems Due to Changes in Tasks and Task Structure

AI is often used to alleviate labour-intensive and error-prone tasks, but it can also result in changes to the tasks that operators must perform. This increased complexity often requires operators to possess new skills, replacing simple tasks with complex cognitive ones that may appear deceptively easy. AI systems are typically highly complex. This complexity can lead to misunderstandings between the operator's mental model and the behaviour of the system. Consequently, mismatches can occur, leading to errors in the operation of the system. Organisations may place not enough emphasis on training, leading to errors caused by insufficient skill levels (Sheridan & Parasuraman, 2005).

Problems Due to Operators' Cognitive and Emotional Response to Changes

AI complexity can lead to errors and misunderstandings, as the mental model of the operator may not match the behaviour of the system. Furthermore, operators' cognitive and emotional responses can amplify problems. For example, as AI changes the operator's task from direct control to monitoring, the operator may be more prone to direct attention away from the monitoring task, leading to decreased feedback from the system (Sheridan & Parasuraman, 2005). The adaptation of AI in aviation can similarly result in inappropriate reliance and compliance, wherein operators might overly depend on or comply with AI systems even in situations where they perform inadequately or fail to fully leverage their capabilities. Poorly calibrated trust can lead to both overtrust and distrust, issues exacerbated by delayed responses to changes in AI performance. Operators could also excessively rely on AI advice, even in cases of omission or commission errors. Operators' cognitive and emotional

[6] See D3.3 Human-AI Teaming Validation Framework for further details

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

responses to AI can also compromise their health, especially if automation increases the demands of the work without increasing decision latitude.

Problems Due to Changes in Feedback

The provision of feedback holds paramount importance in the context of AI in aviation systems. Yet, a significant factor contributing to potential failures is the alteration of feedback experienced by operators due to the integration of AI. Poorly designed AI systems can disrupt the system's potential for performance enhancement by modifying or eliminating feedback mechanisms. To optimise human-AI interaction and operational performance, it's imperative to develop AI systems that furnish operators with information encompassing AI modes, system statuses, and upcoming actions.

Integrating AI with the aim of supporting human operators can lead to instances of AI surprises, which elevate the need for coordination. When AI can consistently and transparently execute assigned tasks, and operators are adequately trained to anticipate AI actions, the cognitive load can be reduced. The central issue revolves around effective cooperation and observability, rather than asserting authority or autonomy. Successful collaboration between humans and AI hinges on shared representations, where the operator's mental model aligns with the machine's functional and causal behaviour, both of which correspond to the operator's interface (Borst, Mulder & van Paassen, 2019).

Effective feedback-driven communication can be achieved by designing interfaces that demand minimal cognitive processing from users yet can be rapidly grasped with a quick glance, enabling swift action.

5.2. Review of HF Assessment Frameworks

In the aviation context, assessing HF in relation to AI is crucial for ensuring safe and effective operations. To gain insights into this aspect, we will explore several methods and frameworks that provide valuable guidance for evaluating human performance in the context of AI. These frameworks include:

- the Situation Awareness Framework for Explainable AI (SAFE-AI), which focuses on understanding and enhancing situation awareness when utilising AI systems. Additionally, we will examine;
- the Technology Acceptance Model (TAM), which explores the factors influencing the acceptance and adoption of AI technologies by human operators. Furthermore, we will explore;
- the Human Performance Assessment Process developed by SESAR, which provides a comprehensive approach to evaluating the impact of human performance on the safe and efficient operation of AI systems in aviation.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

5.2.1. Frameworks

The Situation Awareness Framework for Explainable AI (SAFE-AI)

Drawing upon the existing body of knowledge in HF research, the authors propose the SAFE-AI (Sanneman & Shah, 2022). This framework, consisting of three distinct levels, serves as a valuable tool for the development and assessment of explanations concerning the behaviour of AI systems. These levels of explainable AI (XAI) are derived from the information requirements of human users, which can be determined using the established levels of situation awareness (SA) framework found in HF literature.

- **Level 1** in the proposed XAI framework pertains to explanations that focus on **perception**, encompassing information about the actions taken by an AI system and the decisions it has made. This level aims to address "what?" questions and provides insights into both the inputs and outputs of the AI system. In the realm of explainable machine learning, level 1 explanations may include details about the data fed into the system or the resulting classifications, regression analyses, or cluster information.
- **Level 2** within the XAI framework focuses on **comprehension**, aiming to provide explanations as to why an AI system acted in a specific manner or made certain decisions, as well as elucidating the implications of these actions in relation to the system's goals. Level 2 XAI addresses "why?" questions, and generally encompasses information about the system's underlying model.
- **Level 3** within the XAI framework deals with **projection**, providing explanations concerning the future actions of an AI system under normal circumstances or in alternative scenarios or contexts. Level 3 XAI addresses "what if?'' and "how?" questions. The goal is to explain how the system would respond if certain inputs or parameters were modified or if human users took specific actions. Level 3 XAI also incorporates **counterfactual** or simulated information to shed light on the system's future behaviour in the presence of changes to inputs or system parameters that may arise from human actions.

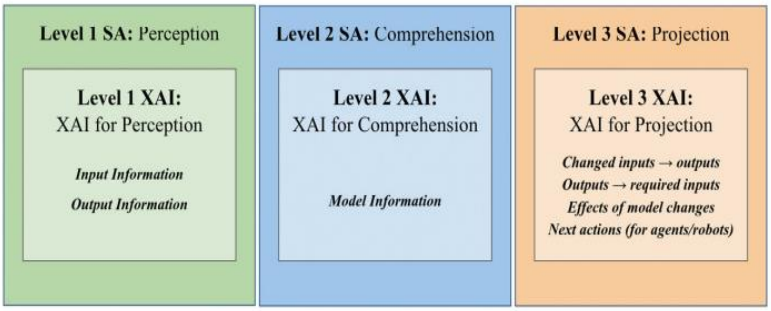

Figure 13.Situation Awareness Framework for Explainable AI.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

The Situation Awareness Framework for Explainable AI (SAFE-AI) and the Construal Level Theory (CLT) share intriguing parallels in their approaches to effective communication and tailored information.

Construal Level Theory (CLT)

Construal Level Theory (CLT) is a psychological framework that sheds light on how individuals perceive and process information based on the perceived psychological distance from a topic. Developed in social psychology, CLT suggests that people mentally construct and interpret information differently depending on factors such as time, space, relevance, and personal interest. CLT proposes that psychological distance can be classified into various levels, ranging from proximal (close in time, space, and relevance) to distal (distant in these aspects). As psychological distance increases, individuals tend to think more abstractly, focusing on core concepts and general ideas. Conversely, as psychological distance decreases, people adopt a more concrete mindset, emphasising specific details and immediate information. SAFE-AI's three levels of explainable AI align with CLT's six levels discussed by McDermott and Folds (2022). In their work, the authors apply Construal Level Theory (CLT) in the design of informational systems. They establish six CLT levels adaptable based on specific requirements (McDermott & Folds, 2022). Executive Summary conveys the core claim and outcome intent in an abstract manner using visuals and concise text. Mission Overview adds context and engagement rules while highlighting performance factors. Mission Summary provides a succinct sequence of actions and essential parameters for success. Mission Brief includes comprehensive background, contingencies, and relevant considerations. Mission Plan/Report, encompasses all plan elements, including detailed parameters and potential constraints. Mission Details/Logs, offers on-demand elaboration of Level 5 with further intricate data. This application of CLT offers adaptable information granularity, catering to users' cognitive preferences and optimizing information transfer.

Both frameworks, SAFE-AI and McDermott and Folds' CLT application (2022), emphasise adjusting information to the audience's understanding, with CLT's psychological distance guiding the adaptation of detail and SAFE-AI's levels catering to diverse user needs, ultimately enhancing the communication of AI system behaviour and decisions.

Technology Acceptance Model (TAM)

The **TAM** has been extensively employed in the aviation industry to assess the acceptance and adoption of AI-enabled systems by pilots and other aviation personnel. This model takes into account various factors that influence users' intentions to use a particular technology, such as their perceived usefulness and ease of use.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

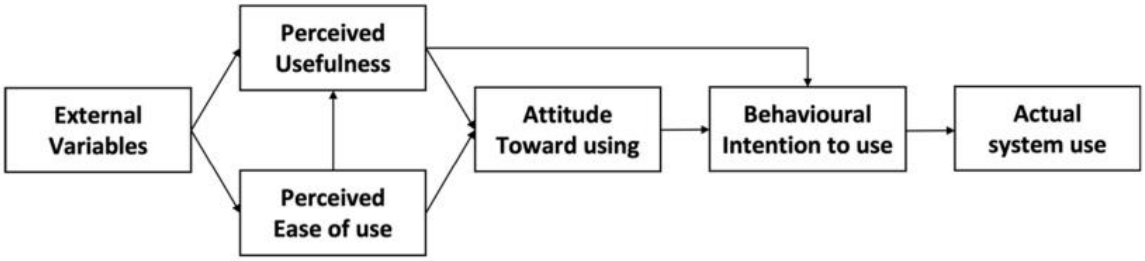

Figure 14.Technology Acceptance Model.

Although TAM originated in the United States, its principles and concepts have gained rapid and widespread recognition, extending its application to various global contexts (Basak, Gumussoy, & Calisir, 2015) (Punnoose, 2012).

A systematic analysis of 23 studies revealed that only five out of the 56 external factors examined were widely acknowledged for their role in fostering technology acceptance among user groups. These factors encompass managerial, operational, organisational, strategic, and IT infrastructure considerations (Emad, El-Bakry, & Asem, 2016).

HP Assessment Process SESAR

The SESAR Human Performance Assessment Process (HPAP) is a systematic and adaptable approach aimed at identifying and addressing potential human performance issues within ATM operations (EUROCONTROL, 2020). The process encompasses the evaluation of individual controllers, teams, or entire organisations, allowing for flexibility in its application across various contexts.

The HPAP process takes a systemic perspective on human performance, considering both individual and environmental factors that can influence performance outcomes. Factors considered include workload, stress, fatigue, training, communication, equipment design, and organisational culture. By adopting this comprehensive approach, the HPAP process facilitates the identification of potential issues that may be overlooked by narrower assessments.

The process commences with a task analysis, which involves a meticulous examination of the specific tasks involved in ATM operations, assessing their cognitive, physical, and sensory requirements. This analysis aids in the identification of potential performance issues and areas in need of improvement.

The subsequent step entails a HF analysis, which takes into consideration both individual and environmental factors that can impact performance. This analysis examines variables such as workload, stress, fatigue, training, communication, equipment design, and organisational culture. By incorporating the HF analysis, potential issues that may not be adequately captured by solely focusing on task demands can be identified.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Based on the outcomes of the task and HF analyses, the HPAP team formulates recommendations aimed at optimising human performance and minimising risks. These recommendations may involve changes to procedures, training programs, equipment design, or organisational culture.

The final stage of the process involves implementing and monitoring the recommended changes. This entails making necessary modifications to the ATM system and continuously monitoring the effectiveness of these changes over time. The HPAP process follows an iterative approach, allowing for ongoing assessments and adjustments to ensure that human performance remains optimised.

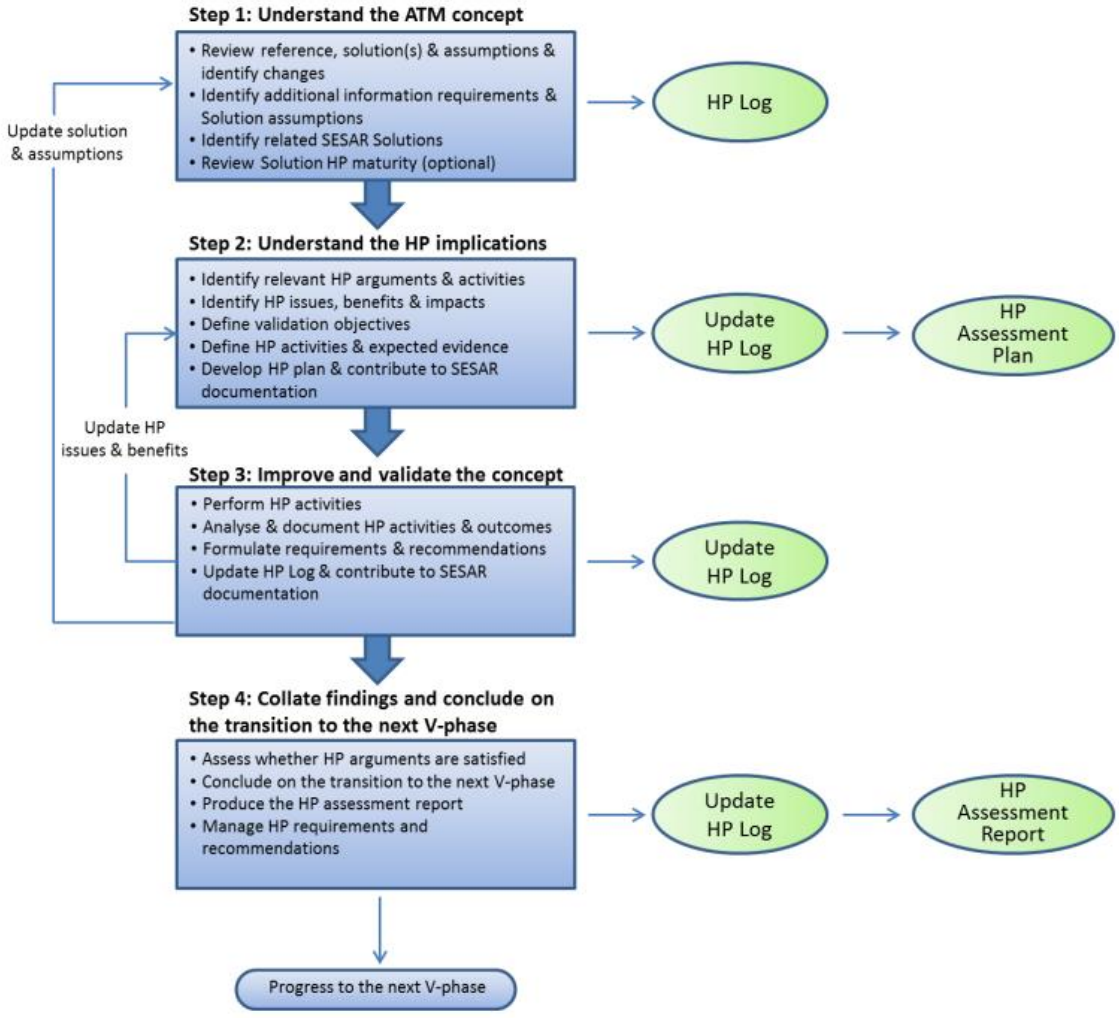

Figure 15.Steps of the HP assessment process.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

The Human Performance assessment is performed in SESAR as per the HP Reference Material. The main HP arguments addressed in the Human Performance assessment approach are (more detail in Annex B):

- Arg1: The role of human actors in the System is consistent with human capabilities and limitations
- Arg2: The contribution of the human within the system supports the expected System Performance
- Arg3: Team structures and team composition support the human actors in performing their tasks
- Arg4: Human Performance related transition factors are considered 6.2.2 Tools for HP assessment

Human-AI Collaboration Framework (CPAIS, 2019)

The Collaborations Between People and AI Systems (CPAIS) Expert Group within the Partnership on AI has devised a Human-AI Collaboration Framework (CPAIS, 2019). This framework consists of 36 questions that identify distinguishing characteristics of human-AI collaborations. By highlighting the nuances associated with specific AI technologies, along with their implications and potential societal impacts, the Framework can serve as a valuable catalyst for responsible design of products and tools, policy formulation, and research endeavours pertaining to AI systems that interact with humans. By emphasising the intricacies, including the specific implications and potential societal effects, of distinct AI technologies, the Framework can offer valuable guidance towards designing responsible products/tools, forming policies, or conducting research related to AI systems engaged in human interaction. It is important to note that the Framework does not aim to dictate definitive solutions to the questions it raises. Rather, these questions are designed to stimulate deeper insights into human-AI collaboration, contribute to the decision-making processes within the AI community regarding responsible AI development and implementation, and ultimately influence technological practices. The questions proposed by the framework will be presented in the next section.

5.2.2 Human Performance Assessment Tools

Ensuring safe human-AI interaction in aviation necessitates assessing psychosocial safety beyond physical aspects. Psychosocial safety is concerned with creating conditions that support positive mental health, reduce stress, and promote a sense of belonging, respect, and fairness. Validated questionnaires evaluate dimensions like safety, anthropomorphism, and likability. Negative Attitude toward Robots Scale (NARS) by Nomura et al. (2006) quantifies negative attitudes toward robots, including AI-powered agents. The BEHAVE-II questionnaire by Joosse et al. (2013) combines subjective and objective metrics. Established tools like System Usability Scale (SUS) and User Experience Questionnaire (UEQ) evaluate usability and experience. Situation Awareness Global Assessment

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Technique (SAGAT) by Endsley measures situation awareness through freeze-on-line probes. Subjective mental workload assessments include the well established NASA Task Load Index (TLX) which measures subjective workload in tasks, encompassing mental, physical, and temporal demands, relying on individuals' subjective perceptions. Furthermore, behavioural metrics offer direct insights into perceived safety, reflecting human behaviour and responses. In aviation, Human Reliability Analysis incorporates HF to address risk assessment

5.3. Proposed and Adapted HF Assessment for IAs

The development of effective guidelines for collaborations between individuals and AI systems necessitates a comprehensive understanding of the dynamics inherent to such collaborations, encompassing aspects of transparency, trust, responsibility for decision-making, and appropriate levels of autonomy. In light of these considerations, we consider adopting the Human-AI Collaboration Framework (CPAIS, 2019) developed by the Collaborations Between People and AI Systems Expert Group. Specifically, the 36 questions of the framework were adapted to the IA domain.

Category	Questions
I. Nature of Collaboration	
Stage of development or deployment	1. Is the IA fixed once deployed or evolving over time via model updates/continual interaction? 2. To what extent is there ongoing collaboration between the IA's developer(s) and the system? [No collaboration, limited collaboration, moderate collaboration, active collaboration] 3. Is the IA system currently used by people other than the original developers?
Goals	4. Are the goals of the human-AI collaboration clear or unclear? 5. What is the nature of the collaboration's goals? [Physical, knowledge/intellectual, emotional, and/or motivational in nature] 6. Is empathy a precondition for the human-AI interaction to function as intended? 7. Are the human and the IA system's goals aligned?
Interaction Pattern	8. Is the collaboration repeated over time or is it a one-time engagement? If over time, at what time-scale? 9. Is the interaction concurrent – with both human and IA contributing in parallel – or does it depend on taking turns?

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Degree of agency	10. Does the IA or human agent contribute more to the system's decision-making? Action-taking? 11. How much agency does the human have? The IAsystem? [None, limited, moderate, high, full]
II. Nature of Situation	
Location and context	12. Are other people or other IA systems involved as third-parties? 13. Are the human and IA agents co-located physically or virtually?
Awareness	14. Is the human likely aware that they are interacting with a IA system? 15. Does the human need to consent before interacting with the IA system?
Consequences	16. How significant are the consequences should the IA fail to perform as designed/expected? What are those consequences? [Low, moderate, high] 17. How significant are the benefits of the IA to the users should it perform as designed/expected? What are those benefits? [Low, moderate, high] 18. What are the potential consequences and benefits of the outcome of the collaboration? 19. What might be the broader impacts of the human-AI collaboration? 20. To what extent do typical users consider privacy and security when interacting with the IA agent? [Low, Moderate, High]
Assessment	21. Who is the main party or individual assessing the nature and effectiveness of the human-AI collaboration? 22. Are assessments of the human-AI collaboration's outcome subjective or objective?
Level of Trust	23. Are both the human and the IA trusting and trustworthy? AI trustworthiness can be defined broadly, driven by task competence, safety, authority, and authenticity, amongst other features (e.g., we know an AI comes from the same affiliation it claims to be from).
III. AI System Characteristics	
Interactivity	24. What is the mode of the interaction between the two agents? [Via screen, voice, wearables, virtual reality, or something else] 25. Could the nature of the data that the IA system operates over impact its interactivity?
Adaptability	26. Is the IA system passively providing information or proactively anticipating the next steps of the interaction?

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Performance	27. How predictable is the IA system? [Low, moderate, high] 28. Does the system often produce false-positives? False-negatives?
Explainability	29. Can the IA system communicate its confidence levels to a human? 30. How does the IA system communicate its decision-making process and inputs to that decision-making process to the human?
Personification	31. How human-like is the IA system? [Not very, moderately, or highly human-like] 32. How easily anthropomorphized is the IAt system?
IV. Human Characteristics	
Age	33. Is the person(s) collaborating with the IA system a child (under 18), an adult (18 - 65), or a senior (over 65)?
Differently-abled	34. Does the person collaborating with the IA have special needs or accommodations?
Culture	35. Are there cultural consistencies/norms for those collaborating with the IA system? 36. What level of previous technology interaction has the user(s) of the system had? [Low, moderate, high]

CPAIS aligns well with the SESAR Human Performance (HP) assessment approach. The SESAR Human Performance (HP) assessment approach recognizes several key arguments that are integral to evaluating and understanding human performance in the context of AI collaboration. The mapping between the HP arguments addressed in the SESAR Human Performance assessment approach and the CPAIS categories highlights the interconnected nature of these factors. The assessment approach recognizes that evaluating human performance in AI collaborations requires considering the nature of the collaboration, the specific situation in which it occurs, and the characteristics of the AI system involved. By mapping these arguments to the CPAIS categories, a comprehensive assessment framework can be established to understand and optimise human performance in the context of AI collaboration.

The first argument, Arg1, emphasises that the role of human actors in the system should align with their capabilities and limitations. This aspect is closely aligned with the CPAIS category of Nature of Collaboration (I), which explores the extent to which the collaboration between humans and AI systems is consistent with the human's abilities.

The second argument, Arg2, highlights the significance of the human contribution in supporting the expected performance of the overall system. This argument not only relates to the Nature of

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Collaboration (I) category but also intersects with the AI System Characteristics (III) category. It emphasises that the design and performance of the AI system should be aligned to facilitate and enhance the capabilities of the human actor within the collaboration.

Arg3 focuses on the importance of team structures and composition in supporting human actors in their tasks. This argument aligns with both the Nature of Collaboration (I) and Nature of Situation (II) categories of CPAIS. It recognizes that effective collaboration between humans and AI systems relies on well-designed team structures and appropriate allocation of tasks among team members.

The final argument, Arg4, addresses human performance-related transition factors, underscoring the need to consider changes in competence requirements, staffing levels, and training needs. This argument aligns with both the Nature of Collaboration (I) and Nature of Situation (II) categories in CPAIS. It emphasises that a comprehensive assessment of human performance in AI collaborations should consider the impact of transitions on the capabilities and requirements of human actors.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

6. Methods to assess liability and legal compliance aspects of Human IA Systems

6.1 Main issues of AI for liability risks and legal compliance in aviation

These pages aim to introduce and explain the new KPAs - compliance and liability - introduced by the HAIKU validation framework and their respective methodologies.

With the introduction of automation – including AI-based systems – task responsibilities are progressively delegated to technology, and liability for damages will tend to shift from human operators to the organisations that designed and developed the technology, defined its context and uses, and are responsible for its deployment, integration, and maintenance of technologies.

EASA (EASA, 2023) remarked that developers, manufacturers and organisations play a pivotal role before the new common challenges posed by the development and deployment of AI-based solutions in aviation. Indeed, the intrinsic features of these technologies shift the traditional development paradigms as well as the current regulatory framework, urging for ad hoc adaptations. Each stakeholder has to continuously assess the impact of these new technologies on its intended users as well as on its internal process. This continuous assessment process should be systemic including all the areas and people affected by the proposed innovation (EASA, 2023, p. 19).

It is crucial to note that the lifecycle process for AI applications has a larger scope than the one considered for traditional systems development. Not only does the safety management process go beyond the existing requirements and cover the whole technology lifecycle, since the early stages of the design, but the use of these new tools has to be coupled with additional requirements on users' training. In other words, AI application requirements will not cover only the development of new technologies but also the deployment and operational phases. Strategies, tasks and responsibilities that until now could have been considered clearly compartmentalised should be read within a unitary framework.

The HAIKU project embraces this approach and introduces in its design and validation framework KPAs and methodologies aimed to provide insights on the responsibilities and liability risk exposures of the actors involved, in order to mitigate the possible negative consequences step by step, over the development of the applications.

6.1.1 HAIKU liability framework

Before approaching the legal framework concerning liability, a terminological premise is needed. Legal scholars and practitioners use to distinguish the consequences of actions or omissions according to

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

different criteria. In this connection, the three keywords in the analysis of the HAIKU legal framework should be accountability, responsibility, and liability.

- **Accountability** within a relational context involves an individual or agency being held to answer for the performance expected by some significant "other". Accountability can furtherly be intended as a principle having a procedural dimension. From an operative perspective, accountability is framed on an individual basis, and basically involves: (1) organisational relationship among two or more subjects, defined by law or by factual conditions; (2) a general duty to care about a process or procedure; (3) a general duty to monitor the regular (i.e., correct, and safe) functioning of a process or procedure; (4) a general duty to report and explain the organisational and operative choices related to a process or procedure.
- **Responsibility** refers to the duty or obligation to carry out a defined task or operation. This duty can be framed on an individual or collective basis, and the subjects involved answer their contribution and its consequences. For the purposes of HAIKU, responsibility implicitly involves: (1) full personal and situational awareness; (2) adequate professional capacity to carry out the assigned task; (3) relational and contextual understanding of individual contributions and the performance of the procedure taken as a whole.
- **Liability** is defined as the condition of being subject to legal consequences deriving from an action or omission. For legal liability to occur, there need to be certain preconditions: (1) a harmful event (2) linked to the action of a person, (3) who was acting in a professional role/task, (4) with no possible justification for the unexpected action. There are also the moral grounds of legal liability that, according to the just culture, should always overlap with legal liability: the person should have moral blame (liability) only when the harm was caused by consciously or recklessly violating a duty/task.

These three different profiles usually coexist and, in some cases, they coincide and are referred to by the same actor. However, in some others, there is no perfect overlap. In these cases, we may have different actors subject to diversified legal regimes. In particular, those in accountable positions can answer for the action and/or omission of those who took part in the procedures they have to supervise (secondary or vicarious liability).

In case of an accident, liability can affect different categories of operators in different ways. Legal persons – such as air carriers, ANSPs, States, and insurance companies – can incur organisational and vicarious liability and can be obliged to repair material and economic damages. On the other hand, natural persons that materially perform different tasks – like ATCOs, PIC and other human operators – may be charged if their behaviour caused the negative occurrence.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

However, aviation mostly experienced peculiar criminal offences. Usually, incidents and casualties are due to accidental situations that the involved operators can difficulty predict or control. Intentional wrongdoings are minimal and quite remote.

These considerations suggest extending the scope of the analysis even to indirect criminal liability issues related to organisational and training gaps and deficits. Inadequate ex-ante and ex-post estimations of each operator's workload, as well as the lack of specific training sessions, may have detrimental consequences on the personal and professional capacities of the involved subject. And these organisational deficiencies can materially influence the state of mind of the actors performing their tasks.

6.1.2 The manufacturers and product liability in HAIKU

The state-of-the-art AI development in aviation calls for a debate on the role of developers, manufacturers and producers in developing and implementing new technological solutions. Indeed, according to product liability law, these actors are responsible for the declared and expected qualities of the developed tools. The main references for the regulatory requirements concerning the development of AI-based solutions are the EASA Concept Papers nn. 1 and 2 (including their future amendments, integrations and updates) as well as any consolidation and/or update of the current legal and regulatory framework (see D7.1).

It is essential to note that, according to the proposed amendment to the EU product liability regime, AI systems and AI-enabled goods may be plainly qualified as "products" (D7.1, p. 107). In this regard, not only hardware manufacturers but also software providers and providers of digital services that affect how the product works could be held liable (D7.1, p. 107).

In light of the above, a product liability hypothesis ("the manufacturer is strictly liable under product liability") may be usually confirmed if the following conditions are jointly satisfied:

- the technology counts as a **product**,
- the technology is **defective**,
- the technology **causes damage**, and
- the technology manufacturer qualifies as a producer.

Under general principles, three alternative conditions for a product's defectiveness are commonly distinguished in:

- **Design defects**, intended as a defect where the product corresponds to the intended design, but the design is flawed. Design defects are by nature a less predictable legal risk for manufacturers because the product functions as intended by the manufacturer, and yet may be regarded as defective.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

- **Manufacturing defects**, as a defect where the product is flawed in that it does not correspond to the intended design. Such defects are probably the least controversial, rather foreseeable, and most within the ambit of control of a producer.
- **Warning defects**, namely situations where a product with intrinsically dangerous qualities is not accompanied by warnings that permit the utilisation of the product in a manner that minimises or eliminates this unreasonably dangerous quality.

At the present stage of HAIKU it is warmly recommended to take into careful consideration the potential issues concerning design defects and warning defects. Indeed, issues concerning these two categories may emerge since the early stage of the design, and if gradually mitigated over the development process ensure satisfying complaint levels and better traceability and acceptability of the solutions at issues by adopting organisations and end-users.

6.1.3 Organisations and enterprise liability in HAIKU

According to the EASA guidelines, organisations that aim to introduce in their processes and/or participate in the development of AI-based solutions for aviation need to introduce the appropriate adaptations in order to ensure the adequate capability to meet the objectives defined within the AI trustworthiness building blocks, and to maintain the compliance of the organisation with the corresponding implementing rules (EASA, 2023, p. 101).

Considering the liability risks correlated to this transition, it is essential to note that generally organisations can be liable in two different situations, namely:

- **Vicarious liability** for negative occurrences attributable to its employees
- **Enterprise liability** for negative occurrences due to its own organisational choices and strategies [1]

These two forms of liability need scrupulous attention when organisations consider or decide to adopt and implement highly automated solutions, like AI-based applications. Indeed, by opting for these innovative strategies, the organisations (e.g., for the purpose of this report you should consider airlines and air carriers, ANPS, USSP, and airport managing companies) are responsible for all the organisational aspects of these innovations. Moreover, they may be also responsible for the behaviour of their employees in their interactions with the new tools. In this connection, they thus have to ensure adequate training for familiarising with and using these new technologies, ensuring the efficient and safe performance of usual tasks. They are further responsible for all the organisational aspects of these innovations, choosing only those products or solutions adequate for their operative purposes and tasks and reviewing all the internal policies and procedures impacted by the innovation. It is important to note that vicarious liability aspects are frequently correlated to the organisation's operative framework and the applicable procedures and practices. For this reason, also taking into account the

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

maturity level of HAIKU UCs, we decided to approach organisation liability risks from this broader perspective.

In this regard, from a more general perspective, an enterprise can be exposed to be liable for organisational liability if the following conditions are jointly met:

- there is an **injury** to a legally protected interest; and
- there is a **causal link** between the activities or processes of the enterprise and the injury (even when they cannot be traced to any individual wrongdoing), and
- the operational activities or processes are inadequate ('**organisational or systemic fault**').

Organisations' liability is one of the main grounds of enterprise liability, with vicarious liability, and may include other specific forms of liability (e.g., product liability).

In this connection, the automation of tasks basically involves the development and/or the deployment of a technology that can integrate or replace the human agency. Beyond efficiency considerations, the tools adopted must satisfy common security and safety standards and ensure that different humans involved in the procedures may be able to monitor the activities automatically performed by machines and promptly and efficiently intervene in the process when needed.

From a legal perspective, these concurrent requirements can be approached in light of four different criteria:

- **the quality of products**, intended as the appropriateness and suitability of the design of technology developed/adopted for the intended uses;
- **the quality of the procedures**, intended as the proper and adequate review, amendment and/or renewal of current standards and protocols in light of the changes introduced by the new solutions;
- **the quality of the implementation**, intended as all the active and proactive measures adopted or to be adopted for a secure and safe implementation of the solutions
- **the quality of the investment**, intended as the delivery of funds for the execution of the project and the secure, safe, and efficient use of the new solution over time.

The careful assessment of all these elements, indeed, contributes to ensuring activities or processes constantly meet the 'best organisational and technical standards', mitigating the liability risks exposure of the organisations involved. The measures may include (but are not limited to) initial and periodic training sessions, initial and periodic audits on the correct functioning of systems and procedures, and initial and cyclic assessment of the technological layout of the procedures even in light of the innovation that meanwhile occurred.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Looking at the guidelines provided by EASA (EASA, 2023), general requirements need to be updated and integrated when organisations aim to introduce AI technologies. More specifically, organisations not only have to review their process and adapt them to the specific features and functions of the new tools (P.ORG-01). They also have to continuously assess and manage the information security risks associated with the design and operation phases of AI applications (P.ORG-02) and they should adapt their continuous risk management process to accommodate the specificities of AI, including interactions with their end-users (employees) and all the relevant stakeholders (P.ORG-05). Analogous considerations are true for the update of training processes (P.ORG-06). Moreover, organisations should establish protocols and processes to continuously assess ethics-based aspects, also considering the establishment of AI ethics review boards (P.ORG-08).

6.1.4 HAIKU approach to liability assessment

Two methodologies will be adopted to assess the legal and regulatory aspects of the HAIKU UCs:

- the **Legal by Design approach**– here intended as a proactive approach to technological design, embedding legal principles and ethical values. The following section about this methodology should be read as a continuum of the D7.1 (delivered at M6), since here you find useful suggestions for profitable use of the regulatory requirements provided by the legal framework for AI in aviation.
- the **Legal Case methodology** – an approach that, drawing from the Safety and HF Cases of the E-OCVM, fosters a proactive approach to liability. This latter offers some insights into the rationale of the liability assessments of the UCs considered for validation contained in D7.3 (expected by M12).

6.2 Legal by Design

Generally, the approach “Legal by Design” [3] is an experimental approach that aims to use design principles and methodologies in the legal domain. A review of the literature on this topic provides different definitions of the lemma (Corrales Compagnucci, Haapio, Hagan, & Doherty, 2020; Danezis, et al., 2014; De Filippi & Wright, 2019; Ducato & Strowel, 2021; Hildebrandt, 2011; Hildebrandt, 2017; Hildebrandt & Tielemans, 2013; Lippe, Katz, & Jackson, 2015) (van den Hoven, Vermaas, & van de Poel, 2015).

For the purposes of HAIKU, we opted for the technology-based understanding of this notion, namely the aim to embed legal principles and ethical values into the technological design (Hildebrandt, 2017). Its purpose, indeed, is to ensure the protection of safety and human rights since the early stages of the design process, tailoring human needs and social expectations on the outline of technological and organisational solutions.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union’s Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

This section provides an introduction to the method, showing its purpose and how the Consortium will use it for the purposes of HAIKU.

6.2.1 Purpose and scope of the approach

The approach "Legal by Design" (hereafter also: LbD) is a generalist method that is not specifically envisioned for aviation. It is inspired by the pivotal question of whether we can use design principles and methodology to fill the gap between written law and technological design and functioning (Hildebrandt, 2019, p. 267-270).

The work already done by EASA with the introduction of the anticipated MOCs for AIenabled systems somehow already confirms the importance of this approach in aviation. The progress the HAIKU project can make in this direction is to test and assess the quality of the insights provided by the Agency in the development and deployment of IAs, with the purpose of contributing with the lesson learnt carrying the project forward.

In light of the above, the approach LbD requires that the legal conditions the legislators have agreed upon are translated into the technical requirements that inform the technological and organisational architecture of operative environments. These requirements should instigate technical specifications and default settings that – other than current systems – afford the protection of ethical values and human rights of the people involved, proactively mitigating the legal risks associated with the technologies development and deployment.

This methodology is primarily intended for use in a proactive way during the design phase of a new operational concept/system, the point is to be able to identify the technologies that may raise some ethical and human rights issues and approach them before the testing phase. This approach is expected to provide important benefits if used early on in the design phase when remedies can be implemented in a cost-effective way. The application of the proactive process is expected to be systematically and periodically applied during the design process in order to assess, at different levels of concept maturity, the legal issues of the IA being developed.

6.2.2 Specific application to HAIKU

In HAIKU, the LbD approach will be applied in the design process of each UC. The developers and the UCs owner shall take into account the technological, organisational and human-based aspects and align and/or adapt their concepts to the legal and regulatory standards outlined in D7.1, with the support of the checklist provided therein. The results obtained in this first round will feed the following liability assessments. The LbD approach has to be intended as an iterative process.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

6.3 The Legal Case

The Legal Case[7] is a methodology with an associated tool intended to support the integration of automated technologies into complex organisations, particularly in ATM. Its purpose is to address liability issues arising from the interaction between humans and automated tools, ensuring that these issues are clearly identified and dealt with at the right stage in the design, development, and deployment process.

This section provides an introduction to the method, showing its purpose, the way it is structured, and the process specifically applied in the reported project.

6.3.1 Purpose and scope of the method

The Legal Case (Contissa, et al., 2013) can be applied to any ATM concept involving automation, i.e., the use of automated technology, including those based on AI. By automated technology, we mean any "device or system that accomplishes (partially or fully) a function that was previously carried out (partially or fully) by a human operator". Two key elements are implicit in our characterization of automation:

Automation is not all-or-nothing. In most cases, automated systems do not fully replace human activity but rather change it, in a way that depends on what tasks are supported by automation, on the extent to which human performance is involved, and on the impact on that performance.

Automation is not tantamount to modernization or technological innovation as such. It covers only those cases where technology has an impact on human activities, and in particular on the interaction between humans and machines. For example, updating a computer with a more powerful system does not necessarily amount to increased automation, nor does an improvement in multi-radar tracking performance, which only implies a reduced radar-update time or more-accurate surveillance data. Our analysis is focused on the cooperation or co-agency between the human and machine when performing certain tasks and on the ensuing changes in the human operator's roles and responsibilities.

The Legal Case has been designed to be flexibly applied across all the phases of maturity in a system's life cycle. The methodology can be applied both proactively (from V1 to V3 of E-OCVM) and retroactively (from V4 on, of E-OCVM). Depending on the maturity phase of the technology, the Legal Case analysis will rely on different types of background information, can be used for different purposes, and will provide different sorts of output.

The Legal Case is primarily intended for use in a proactive way during the design phase of a new operational concept/system, the point is to be able to address possible legal issues arising in the future from

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

potential accidents or malfunctions. Indeed, the Legal Case is expected to provide important benefits if used early on in the design phase, when remedies can be implemented in a cost-effective way. The application of a proactive process is expected to be systematically and periodically applied during the design process in order to assess, at different levels of concept maturity, the legal issues of the ATM system being developed.

It is worth noticing that in none of these cases the Legal Case is intended to apportion liability [6] [7] and blame people or the organisation, conversely it is intended to enforce the safety culture of the organisation making all the actors involved aware of the liability risks associated with their roles, tasks and activities and proactively identify suitable mitigations.

6.3.2 The process

The Legal Case process consists of the following four steps:

- **Understand context and concept.** This step involves collecting and elaborating background information about the object of the study so as to understand its socio-technical and normative aspects. The information collected concerns the operational concept itself, the context of its deployment, and the legal and regulatory aspects. This step includes the identification of the level of automation of the concerned ATM system, its impact on roles, tasks and responsibilities and a set of UCs considered relevant for the following legal analysis. Where available, the solutions adopted according to the LbD approach may inform and feed this analysis.
- **Identify liability issues**. [8] [9] This step involves identifying the possible liabilities related to the object of the study and determining the associated liability risks.
- **Address the liability allocation.** This step involves analysing the acceptability of liability risks for all stakeholders, proposing possible mitigations that may improve liability allocation, and making design recommendations accordingly.
- **Collect findings and Systemic Analysis**. This step presents the results of the study, highlighting the liability issues associated with the object of study and the ways to deal with legal risks, as well as making further recommendations.

The application of this methodology requires the use of special tools, also known as argumentation maps. These means are based on the applicable legal requirements to each of the actors involved in the development and deployment of new technology, providing relevant insights about the legal regime applicable to producers, deploying organisations and end-users. More specifically, the maps provide the logical representation of the factual conditions that may confirm a liability hypothesis according to a cause-effect approach (i.e., if these factual conditions may be true, this actor is exposed to liability risks in using this technology). The methodology is tailored to the needs of the aviation domain and includes maps to assess the liability risks of the producers and manufactures of the

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

technology, air carriers and other deploying organisations and end-users like the ATCO, PIC and other front-raw operators. Comparing the results obtained for each subject, the Legal Case allows for identification by design and by default mitigations to improve the liability risk exposure of the subjects more impacted by the introduction of a new technology.

White rectangles represent actions, i.e., sub-steps within each step of the Legal Case. Black rectangles represent a flow of objects from one activity to another, that is, the flow of the information produced in each sub-step of the Legal Case. Bold arrows represent the main workflow. Light arrows represent other connections between objects and actions, that is, the information used as an input for each sub-step. The Levels of Automation Taxonomy (LOAT) (Save & Feuerberg, 2012) table, and the legal argumentation maps used in the process (Failures maps, and the complete set of Legal Analysis maps [10] [11]) are also inputs and appear as red triangles.

SFS assessments are external inputs and appear as white triangles, meaning that – in case those reports are not available - the Legal Case can be applied without using them. Actually, should the Legal Case be completed before the SHS assessments, it can also be considered an input for them.

6.3.3. Specific application to HAIKU

The Legal Case is designed to assess liability risks in the automation process. However, in HAIKU, this methodology will be applied for liability assessment of AI-based systems, according to UCs descriptions as defined in June 2023 (Month 10). The baseline for the application consists in the descriptions of the operative contexts and concepts of operations. The results of the first release of the assessments will be available in D7.3 (expected by M12).

A second iteration is planned at M24 and M36 (respectively expected in August 2024 and August 2025). The recommendations obtained by the intermediated deliveries will contribute to the future steps of the design process over the duration of the project.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

7. Conclusions and Recommendations

This document includes a comprehensive review of validation frameworks for safety, HF, security, liability, and regulatory compliance in the context of AI integration in aviation. For the purposes of HAIKU, these frameworks have been used to elaborate a dedicated design and validation framework to be used in the HAIKU project. The framework is complemented by a set of methods and preliminary questionnaires addressed to the UCs owners. The questions of this document cover the project KPAs and respectively drawn from:

- ALTAI and SESAR methodologies for the safety assessment
- ALTAI and SecRAM methodologies for the security assessment
- SESAR methodologies for HF assessment
- DeepBlue methodologies for legal compliance and liability assessments

The purpose of this questionnaire is to facilitate a compliant development of the IAs, providing some insights on SHS, legal compliance and liability since the early stage of the design process.

Overall, these methodologies collectively serve as the foundation for the forthcoming D7.3 Validation of the SHS case-based approach in case studies.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Annex A - References and selected bibliography

AI HLEG. (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment.

Alexy, R. (1994). *Teoria dei diritti fondamentali (Theorie der Grundrechte).* Bologna: Il Mulino.

ARMS Working Group. (2010). The ARMS Methodology for Operational Risk Assessment in Aviation Org.

Borst, C., Visser, R. M., Van Paassen, M. M., & Mulder, M. (2019). Exploring short-term training effects of ecological interfaces: A case study in air traffic control. IEEE Transactions on Human-Machine Systems, 49(6), 623-632.

Chemical Industries Association. Chemical Industry Safety & Health Council. (1977). A guide to hazard and operability studies. Chemical Industry Safety and Health Council of the Chemical Industries Association.

Civil Aviation Authority. (2020). Cyber Assessment Framework (CAF) for Aviation.

Contissa, G., Laukte, M., Sartor, G., Schebesta, H., Masutti, A., Lanzi, P., Tomasello, P. (2013). Liability and automation: Issues and challenges for socio-technical systems. *Journal of Aerospace Operations, 2*(1-2), 79-98.

Corrales Compagnucci, M., Haapio, H., Hagan, M., & Doherty, M. (2020). *Legal Design. Integrating Business, Design and Legal Thinking with Technology.* Northampton, MA, USA: Edward Elgar Publishing.

CPAIS. (2019). Human-AI Collaboration Framework.

Dafoe, A. (2018). AI governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford.*, 1-53.

Danezis, G., Domingo-Ferrer, J., Hanses, M., Hoepman, J.-H., La Métayer, D., Tirtea, R., & Schiffner, S. (2014). Privacy and Data Protection By Design - from policy to engineering. Enisa.

De Filippi, P., & Wright, A. (2019). *Blockchain and the Law. The Rule of Code.* Boston, MA, USA: Harvard University Press.

Delgado Bellamy, D., Chance, G., & Caleb-Solly, P. (2021). Safety assessment review of a dressing assistance robot. *Frontiers in Robotics and AI*.

Ducato, R., & Strowel, A. (2021). *Legal Design Perspectives. Teoretical and Practical Insights from the Field.* Milan, Italy: Ledizioni.

EASA. (2020). Opinion 01/2020: High-level regulatory framework for the U-space.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

EASA. (2023, May). Artificial Intelligence Roadmap 2.0. Human-centric approach to AI in aviation. Cologne, Germany.

EASA. (2023, February). Concept paper: First usable guidance for Level 1 and 2 machine learning application. A deliverable of the EASA AI Roadmap. Cologne, Germany.

Elmarady, A., & Rahouma, K. (2021). Studying cybersecurity in civil aviation, including developing and applying aviation cybersecurity risk assessment. . *IEEE access*, 143997-144016.

Emad, H., El-Bakry, H. M., & Asem, A. (2016). A modified technology acceptance model for health informatics. International Journal of Artificial Intelligence and Mechatronics, 4(4), 153-161.

ENISA. (2016). Securing Smart Airports.

ENISA. (2020). Artificial Intelligence Cybersecurity Challenges.

ENISA. (2021). Securing Machine Learning Algorithms.

EUROCONTROL. (2017). ATM cybersecurity maturity model – Level 1 - Ed 1.0 .

EUROCONTROL. (2018). SESAR Safety Reference Material .

EUROCONTROL. (2019). White Paper Human Factors Integration in ATM System Design.

EUROCONTROL. (2020). *SESAR Human Performance Assessment Process V1 to V3*.

EUROCONTROL. (2020). Fly AI Report - Demystifying and Accelerating AI in Aviation.

EC. (2020). Transport cybersecurity toolkit .

Federal Aviation Administration. (2015). AC 120-92B - Safety Management Systems for Aviation Service Providers.

Franssen, M. (2015). Design for Values and Operator Roles in Sociotechnical Systems. In J. van den Hoven, P. E. Vermaas, & I. van de Poel, *Handbook of Ethics, Values, and Technological Design.* Dordrecht: Springer.

Hahn, D., Munir, A., & Behzadan, V. (2019). Security and privacy issues in intelligent transportation systems: Classification and challenges. *IEEE Intelligent Transportation Systems Magazine,131(1)*, 181-196.

Haiku (2023). State of the art in safety, human factors, and security (SHS) assurance processes in aviation. HORIZON-CL5-2021-D6-01-13

Hart, H. (1994). *The Concept of Law* (3rd ed.). (J. Raz, & P. Bulloch, Eds.) Oxford, UK: Oxford University Press.

Hildebrandt, M. (2011). Legal protection by design: objections and refutations. *Legisprudence*.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Hildebrandt, M. (2017). Saved by Design? The Case of Legal Protection by Design. *NanoEthics*, 1-5.

Hildebrandt, M. (2019). 'Legal by Design' or 'Legal Protection by Design'? In M. Hildebrandt, *Law for Computer Scientists and Other Folks* (pp. 267-270). Oxford: Oxford University Press.

Hildebrandt, M., & Tielemans, L. (2013). Data protection by desing and technology neutral law. *Computer Law and Security Review*.

International Civil Aviation Organization (ICAO) . (2019). Cybersecurity Strategy.

Jahan, F., Sun, W., Niyaz, Q., & Alam, M. (2019). Security modeling of autonomous systems: A survey. . *ACM Computing Surveys (CSUR), 52(5)*, 1-34.

Jing, H., Wei, W., & Zhou, C. (2021). An Artificial Intelligence Security Framework. *Journal of Physics: Conference Series (Vol. 1948, No. 1, p. 012004).*

Jing, H., Wei, W., Zhou, C., & He, X. (2021, June). An Artificial Intelligence Security Framework. (n.d.). *Journal of Physics: Conference Series (Vol. 1948, No. 1, p. 012004).*

Joosse, M., Sardar, A., Lohse, M., & Evers, V. (2013). BEHAVE-II: The revised set of measures to assess users' attitudinal and behavioral responses to a social robot. International journal of social robotics, 5, 379-388.

Koroniotis, N., Moustafa, N., & Sitnikova, E. (2019). Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset. *Future Generation Computer Systems, 100*, 779-796.

Lasota, P., Fong, T., & Shah, J. (2017). A survey of methods for safe human-robot interaction. Foundations and Trends® in Robotics, 5(4), 261-349.

Lawson, C. (2010). Technology and the extension of human capabilities. *Journal for The Theory of Social Behaviour, 40*(2), 207-223.

Lippe, P., Katz, D. M., & Jackson, D. (2015). Legal by Design: A New Paradigm for Handling Compleixity in Banking Regulation and Elsewhere in Law. *Oregon Law Review, 93*(4).

Lykou, G., Anagnostopoulou, G., & Gritzalis, D. (2018). Smart airport cybersecurity: Threat mitigation and cyber resilience controls. *Sensors, 19(1)*.

McCarthy, J. (2007). From here to human-level AI. Artificial Intelligence, 171(18), 1174-1182.

McDermott, T., & Folds, D. (2022). Construal level theory in the design of informational systems. Frontiers in Physics, 10, 958450.

Menzies, T., & Pecheur, C. (2005). Verification and validation and artificial intelligence. Advances in computers, 65, 153-201.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

National Institute of Standards and Technology. (2018). Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1.

NIST (2023). Artificial Intelligence Risk Management Framework (AI RFM 1.0). 1-43.

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006, July). Altered attitudes of people toward robots: Investigation through the Negative Attitudes toward Robots Scale. In Proc. AAAI-06 workshop on human implications of human-robot interaction (Vol. 2006, pp. 29-35).

Rudner, T. G. (2021). . Key concepts in AI safety: an overview. *Computer Security Journal* .

Rudner, T., & Toner, H. (2021). Key concepts in AI safety: an overview. *Computer Security Journal* , 1-10.

Rueß, H., & Burton, S. (2022). Safe AI-How is this Possible? *arXiv preprint* , 1-22.

Sanneman, L., & Shah, J. (2022). The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. *International Journal of Human–Computer Interaction,38(18-20),*, 1772-1788.

Save, L., & Feuerberg, B. (2012). Designing Human-Automation Interaction: a new level of Automation Taxonomy. In D. B. De Waard, *Proc. Human Factors of Systems and Technology.*

SESAR Joint Undertaking. (2017). Security Risk Assessment Methodology for SESAR 2020.

Sheridan, T., & Parasuraman, R. (2005). Human-automation interaction. *Reviews of human factors and ergonomics, 1(1)*, 89-129.

Ukwandu, E., Ben-Farah, M., & Hindy, H. (2022). Cyber-security challenges in aviation industry: A review of current and future trends. . *Information, 13(3)*.

van den Hoven, J., Vermaas, P., & van de Poel, I. (2015). *Handbook of Ethics, Values, and Technological Design.* Dordrecht, The Netherlands: Springer.

Vermaas, P., Kroes, P., van de Poel, I., Franssen, M., & Houkes, W. (2011). Sociotechnical Systems. In P. E. Vermaas, P. Kroes, I. van de Poel, M. Franssen, & W. Houkes, *A Philosophy of Technology. From Technical Artefacts to Sociotechnical Systems* (pp. 67-81). Dordrecht: Springer.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Annex B - Assessment grids

ALTAI assessment questions considered for the Safety Assessment

GENERAL SAFETY				
No.	**Question**	**YES**	**NO**	**Why?**
1	Did you define risks, risk metrics and risk levels of the AI system in each specific UC?			
1.1.	Did you put in place a process to continuously measure and assess risks?			
2	Did you identify the possible threats to the AI system (design faults, technical faults, environmental threats) and the possible consequences?			
2.1	Did you assess the risk of possible malicious use, misuse or inappropriate use of the AI system?			
2.2	Did you define safety criticality levels (e.g. related to human integrity) of the possible consequences of faults or misuse of the AI system?			
3	Did you assess the dependency of a critical AI system's decisions on its stable and reliable behaviour?			
3.1	Did you align the reliability/testing requirements to the appropriate levels of stability and reliability?			
4	Did you plan fault tolerance via, e.g. a duplicated system or another parallel system (AI-based or 'conventional')?			
ACCURACY				
No.	**Question**	**YES**	**NO**	**Why?**

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

1	Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences?			
2	Did you put in place measures to ensure that the data (including training data) used to develop the AI system is up-to-date, of high quality, complete and representative of the environment the system will be deployed in?			
3	Did you put in place a series of steps to monitor, and document the AI system's accuracy?			
4	Did you consider whether the AI system's operation can invalidate the data or assumptions it was trained on, and how this might lead to adversarial effects?			
5	Did you put processes in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated?			
RELIABILITY, FALL-BACK PLANS AND REPRODUCIBILITY				
No.	**Question**	**YES**	**NO**	**Why?**
1	Could the AI system cause critical, adversarial, or damaging consequences (e.g. pertaining to human safety) in case of low reliability and/or reproducibility?			
1.1	Did you put in place a well-defined process to monitor if the AI system is meeting the intended goals?			
1.2	Did you test whether specific contexts or conditions need to be taken into account to ensure reproducibility?			
2	Did you put in place verification and validation methods and documentation (e.g. logging) to evaluate and ensure different aspects of the AI system's reliability and reproducibility?			
2.1	Did you clearly document and operationalise processes for the testing and verification of the reliability and reproducibility of the AI system?			

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

3	Did you define tested failsafe fallback plans to address AI system errors of whatever origin and put governance procedures in place to trigger them?			
4	Did you put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score?			
5	Is your AI system using (online) continual learning?			
5.1	Did you consider potential negative consequences from the AI system learning novel or unusual methods to score well on its objective function?			

ALTAI assessment questions considered for the Security Assessment

RESILIENCE TO ATTACK AND SECURITY				
No.	**Question**	**YES**	**NO**	**Why?**
1	Could the AI system have adversarial, critical or damaging effects (e.g. to human or societal safety) in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use?			
2	Is the AI system certified for cybersecurity (e.g. the certification scheme created by the Cybersecurity Act in Europe)19 or is it compliant with specific security standards?			
3.1	Did you assess potential forms of attacks to which the AI system could be vulnerable?			
3.2	Did you consider different types of vulnerabilities and potential entry points for attacks such as:			
3.2.1	Data poisoning (i.e. manipulation of training data);			

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

3.2.2	Model evasion (i.e. classifying the data according to the attacker's will);			
3.2.3	Model inversion (i.e. infer the model parameters)			
4	Did you put measures in place to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle?			
DATA GOVERNANCE				
No.	**Question**	**YES**	**NO**	**Why?**
1	Did you consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection?			
2	Depending on the UC, did you establish mechanisms that allow flagging issues related to privacy concerning the AI system?			
3	Did you consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data?			
4	Did you align the AI system with relevant standards (e.g. ISO, IEEE) or widely adopted protocols for (daily) data management and governance?			

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

SESAR HP arguments considered in the Human Factors Assessment

	Argument	Explanation
Arg. 1	The role of the human is consistent with human capabilities and limitations.	Roles are defined as the position(s) or purpose(s) that someone has in an organisation. Responsibility is defined as a duty or obligation to perform a set of tasks assigned to a specific role.
Arg. 1.2	Operating methods (procedures) are exhaustive and support human performance.	Operating methods (Procedures) are the accepted courses of actions to fulfil a certain responsibility. For an ANSP, they are normally laid down in an Operational Manual and need to be in line with ICAO provisions. For the flight deck, operating methods are described in Standard Operating Procedures which are part of Flight Manuals or other aircraft-specific documentation or airlines' operating manuals.
Arg. 1.3	Human actors can achieve their tasks (in normal & abnormal conditions of the operational environment and degraded modes of operation)	The proposed project changes must not negatively affect human performance and hence the ability of the human actor to perform & achieve their tasks in normal & abnormal operating conditions as well as degraded modes of operation. Several factors can have a significant impact on human performance, these include; subjective workload, error potential, situation awareness, trust, fatigue.
Arg. 2	Technical systems support the human actors in performing their tasks.	In order for the technical systems to support the human in carrying out their tasks, the usability of the technical system must be assured. Usability is the extent to which a system allows people to achieve goals (tasks) in an effective, efficient and satisfactory way (HF Case v2.0).
Arg. 2.2	The performance of the technical system	In order for the technical system to support the human actor(s) in their tasks all the information presented must be: relevant & necessary to the task(s) being

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

	supports the human in carrying out their task.	performed; accurate; and presented in a timely manner.
Arg. 2.3	The design of the human-machine interface supports the human in carrying out their tasks.	The human machine interface refers to the modes by which the human user and the machine communicate information and by which control is commanded (HF Case v2.0).
Arg. 3	Team structures and team communication support the human actors in performing their tasks.	Teams and communication relates to how people work together and communicate with each other on shared goals and tasks (The HF Case v2.0). Changes in team structure can include changes to the composition of a team in terms of roles, as well as, changes to the way in which tasks are allocated between the team members. Such changes may impact the communication flow within a team and way tasks are performed.
Arg. 3.2	The allocation of tasks between human actors supports human performance.	The allocation of tasks between human actors refers to the way in which tasks are distributed between the different team members.
Arg. 3.3	The communication between team members supports human performance	Communication can be said to support human performance if it enables the timely and accurate passing of all the necessary information between actors so that the communicated information is received and understood by the actor(s) that need it. Communication can be verbal and non-verbal (e.g., using gestures).
Arg. 4	Human Performance related transition factors are considered.	Although the transition to the proposed concept will only happen in V5, E-OCVM requires an assessment of the transition feasibility in V3. For this reason, HP related aspects of the operational concept that are crucial for the successful transition at a later stage should need to be identified.

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332

Arg. 4.2	Changes in competence requirements are analysed.	Competence refers to the skill, knowledge & experience required by the human actors to SESAR2020 HUMAN PERFORMANCE ASSESSMENT PROCESS V1 TO V3- INCLUDING VLDS Insert project logo here 83 ID Argument Explanation perform their tasks to the required standard/level of required performance.
Arg. 4.3	Changes in staffing requirements and staffing levels are identified.	Changes to the roles, tasks and responsibilities may impact and change the number of staff required, as well as the composition and organisation of staff within the organisation
Arg. 4.4	The impact on recruitment and selection processes has been considered.	Changes to the competencies (i.e. skills, knowledge & experience) required by the human actors to perform their work may have an impact on the strategies and criteria for recruitment and selection of staff.
Arg. 4.5	Training needs are identified for the affected human actors.	The training needs resulting from the proposed changes to the human actors' roles and tasks must be identified and defined for all affected operators

© Copyright 2022 HAIKU Project. All rights reserved

This project has received funding by the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332