

≡

Grant Agreement number: 101147454

Project acronym: DEMOQUAS

Project title: DEsigning, Manufacturing and Operating Quantification of Uncertainties to increase Aviation Safety

Type of action: HORIZON Research and Innovation Actions



WP6 “Uncertainty Quantification (UQ) in operational environment & safety risk assessment”

T6.3 “Airport and pilot operations under current and predictive scenarios”

Requirements definition to support aviation’s safety risk assessment [D6.1]

Delivery type:	Report
Lead beneficiary:	EGA
Lead author(s):	Dimitris Noufriadis, Konstantinos Papadopoulos, Elissavet Kapoutsi, Vasileios Gkoutzamanis (AUTH), Chrisalena Athanasiadou, George Triantafyllidis (EGA)
Contributions:	Francesco Orefice (TUD Reviewing)
Contractual delivery date:	31/08/2025
Delivery date:	10/09/2025
Dissemination level:	PU

Information Table

Project Title	DEsigning, Manufacturing and Operating Quantification of Uncertainties to increase Aviation Safety
Project Acronym	DEMOQUAS
GA n.	101147454
Project Coordinator	Aristotle University of Thessaloniki {AUTH}
Project Duration	36 months
Deliverable n.	D6.1
Deliverable title	Requirements definition to support aviation's safety risk assessment
Deliverable version v.	04
Deliverable description	Emphasis on airport and pilot occurrences, including risks of component technological malfunctions
Dissemination level	PUBLIC
Work Package	6
Task(s)	6.3
Lead Beneficiary	EGA
Contributing beneficiary/ies	AUTH, TUD
Due date of deliverable (month)	16
Submission date	10/09/2025

History of Changes

Version	Date	Author/Contributor	Changes
01	02/07/2025	Konstantinos Papadopoulos, Dimitris Noufriadis, Elissavet Kapoutsi Vasileios Gkoutzamanis	Table of Contents
02	29/08/2025	AUTH, EGA	First Draft (Review)
03	29/08/2025	AUTH, EGA, TUD	Revised Draft (Review)
04	10/09/2025	AUTH	Final version

Abbreviations and acronyms

Abbreviation	Definition
Dx.x	Deliverable number x.x
WP(s)	Work Package(s)
Tx.x	Task number x.x
Acronym	Definition
AUTH	Aristotle University of Thessaloniki
EGA	Egnatia Aviation
TUD	TU Delft

Disclaimer

The sole responsibility for the content of this publication lies with the authors. It does not necessarily reflect the opinion of the European Commission. The European Commission is not responsible for any use that may be made of the information contained therein.

Table of contents

- List of Figures 5
- Executive Summary..... 6
- Nomenclature 7
- 1. Introduction 9
- 2. EASA Safety Protocols..... 10
 - 2.1 Safety Reporting..... 10
 - 2.2 Data4Safety 21
- 3. Flight Operations – Pilot Evaluation 24
 - 3.1 Questionnaires 24
 - 3.1.1 Data Collection 24
 - 3.1.2 Data Handling and GDPR 24
 - 3.2 Flight Assessment 24
- 4. Machine Learning for Aviation Safety..... 31
 - 4.1 Literature Review (State-of-the-art)..... 31
 - 4.2 Pilot Operations 34
 - 4.2.1 Bayesian Network Introduction..... 34
 - 4.2.2 State-of-the-Art 36
 - 4.3 Airport Occurrences 39
 - 4.3.1 State-of-the-Art 40
 - 4.4 UQ with Machine Learning & Natural Language Processing 43
- 5. Conclusions..... 47

List of Figures

Figure 1: EASA Safety Reporting Data Flow.....	10
Figure 2: EASA Safety Data Management Process.....	12
Figure 3: The European SRM process.	13
Figure 4: Key risk areas by aggregated ERCS score and number of risk-scored occurrences involving commercial air transport airlines and air-taxi.....	14
Figure 5: Safety issues by aggregated ERCS score and numbers of accidents and serious incidents involving commercial air transport airline and air-taxi and non-commercial business operations.	15
Figure 6: Key risk areas by aggregated ERCS score and number of risk-scored occurrences involving aerodromes and ground handling.....	16
Figure 7: Safety issues by aggregated ERCS score and numbers of occurrences involving aerodromes and ground handling.	17
Figure 8: Key risk areas by aggregated ERCS score and number of risk-scored ATM/ANS occurrences.	18
Figure 9: Safety issues by aggregated ERCS score and numbers of accidents and serious incidents for ATM/ANS safety issues.	19
Figure 10: Reporting rate per 1000 aircraft movements ²	20
Figure 11: Number of occurrences per KRA and severity category in 2023.	20
Figure 12: Data4Safety development roadmap.....	21
Figure 13: Proof of concept D4S Data.....	22
Figure 14: Classification of factors influencing Pilot Performance.	25
Figure 15: Mid-air collision BBN (adapted from).....	37
Figure 16: Topology of a Dynamic Bayesian Network (adapted from)	38
Figure 17: Illustration of sources of uncertainty ⁵⁸	44

Executive Summary

This document presents the first report for WP6 “UQ in operational environment and safety risk assessment”. It describes work done on operational aspects of aviation, including safety reporting, incidents occurrence and pilot performance analysis, through the scope of machine learning applications and probability-driven modelling.

Firstly, the need is established, by reviewing the European Union Aviation Safety Agency’s (EASA) efforts and its corresponding safety reporting protocols that govern European air-travel and airspace. The context is specifically related to EASA’s Data4Safety, an under-development database domain for enhancing aviation safety through shared data and analytics. The corresponding review yields aspects of probability modeling and Natural Language Processing as key elements and are selected for development and use under Task 6.3 (T6.3).

A review around the state-of-the-art regarding Machine Learning (ML) methods for aviation-safety-related matters is conducted. It first seeks to establish a baseline in relation to the general application of ML for aviation safety, while analysis splits into two parts thereafter. The first part relates to the prediction of pilot performance using probability-driven methods like Bayesian Networks, establishing a more specialized threshold for the state-of-the-art and proposing a methodology to be applied within the context of T6.3. The second part relates to the use of BERT-like NLP methods to handle aviation safety reports. The state-of-the-art and proposed methodology are established for this part of the investigation as well.

As part of the development of machine learning methods for the prediction of pilot performance, the AUTH and EGA partners collaborate on the creation of a respective database through a questionnaire-type reporting system. The flight operations of EGA are described and the data collection process is defined. The data handling and flight assessments are also detailed, as required for the development of BN-based models.

The final segment of this document details the proposed application of UQ methods for the ML-based methodologies for pilot and airport operations. A brief state-of-the-art is established, evaluating other applications of UQ in operational aspects of aviation. The UQ methodology for T6.3 follows a similar approach as defined with D3.1 and WP3. Finally, the future steps for the work of T6.3 are also detailed within the context of this section.

Nomenclature

AI	Artificial Intelligence
ASIAS	Aviation Safety Information Analysis and Sharing
ASRS	Aviation Safety Reporting System
ATCo	Air Traffic Controller
ATM/ANS	Air Traffic Management / Air Navigation Services
BERT	Bidirectional Encoder Representations from Transformers
BIS	Best Intervention Strategy
BN	Bayesian Network
CPT	Conditional Probability Table
CSR	Confidential Safety Reporting
D4S	Data4Safety
DAG	Directed Acyclic Graph
DBN	Dynamic Bayesian Network
DPPO	Data Protection and Processing Organization
EASA	European Union Aviation Safety Agency
ECCAIRS	European Coordination Centre for Accident and Incident Reporting Systems
EPAS	European Plan for Aviation Safety
ERCS	European Risk Classification Scheme
FAA	Federal Aviation Authority
FDM	Flight Data Monitoring
FDX	Flight Data Exchange
FOSM	First-Order Second-Moment
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation
IATA	International Air Transport Association
IR	Information Retrieval
JSD	Jensen-Shannon Divergence
KLD	Kullback-Leiber Divergence
KRA	Key Risk Area
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MH	Metropolis-Hastings
ML	Machine Learning
MLE	Maximum Likelihood Estimation
MLM	Masked Language Modeling
NAS	National Airspace System
NASA	National Aeronautics and Space Administration
NER	Named-Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing

NLU	Natural Language Understanding
NTSB	National Transportation Safety Board
PCE	Polynomial Chaos Expansions
PoC	Proof of Concept
QA	Question Answering
RASP	Regional Aviation Safety Plan
RE	Relation Extrapolation
SB	Steering Board
SDM	Safety Data Management
SIA	Safety Issue Assessment
SPF	Safety Performance Framework
SPI	Safety Performance Indicator
TVD	Total Variation Distance
UQ	Uncertainty Quantification
UTFM	Uncertainty Transfer Function Model

1. Introduction

This report introduces the need, reviews the state-of-the-art and establishes the proposed methodology for applying Uncertainty Quantification (UQ) methods on operational aviation aspects, as part of Work Package 6 (WP6) and Task 6.3 (T6.3). The effort relates to two major elements of aviation operations, pilots and airports.

The need arises from the development of data-driven methods for aviation safety like European Union Aviation Safety Agency's (EASA) Data4Safety platform which include a variety of data types and quantities. As such, the purpose of T6.3 is to leverage data-driven Machine-Learning (ML) methods related to aviation safety, applying UQ methods on this kind of modelling tools. In this context, this report seeks to:

- Establish the need by data-driven domains and methods like EASA's Data4Safety to establish a benchmark for use of such processes.
- Describe on-site operational activities of EGA flight operations and review data-acquisition methods to be used as part of T6.3.
- Review the state-of-the-art surrounding ML methods around aviation safety and identify the specific methodologies to be employed within the context of T6.3.
- Define the aspects of pilot and airport operations to be investigated within T6.3 and describe the proposed methodologies.
- Review applications of UQ on operational aspects of aviation and its related safety matters and characterize the methods to be used within the context of T6.3.

In total, this report seeks to review concurrent efforts related to T6.3, define a set of requirements to be met regarding operational aviation safety and propose a methodology to be developed under its activities. It will serve as a reference point for future steps during the dynamic evolution of project efforts.

2. EASA Safety Protocols

EASA stands as the central regulatory authority regarding safety for air operations within the European Union and its directives constitute key targets for aviation safety research. The work performed by the agency is summarized with the following list of principal tasks:

1. Draft implementing rules pertinent to its mission.
2. Certify and approve products and organizations in fields related to its exclusive competence (i.e. Airworthiness).
3. Provide oversight and support to EU’s member states in fields related to its exclusive or shared competence (i.e. Air Operations, Air Traffic Management, etc)
4. Promote the use of European and International standards
5. Cooperate with international actors to achieve the highest level of safety for EU citizens.

This part of the investigation deals with describing the safety reporting pillars of EASA and reviewing the Data4Safety initiative. The latter is set to serve as a unified database for safety reporting and stands as the main motivation aspect for this work.

2.1 Safety Reporting

EASA emphasizes that accident prevention relies heavily on proactively gathering and analyzing safety-related data—incidents and deficiencies often precede accidents, signaling underlying hazards. To supplement reactive safety systems, EASA mandates that relevant civil aviation occurrences be reported, stored, shared, and scrutinized, with appropriate safety actions taken based on findings. This structured approach facilitates hazard detection, mitigation, and continuous safety improvement across the aviation sector. Among its pipelines, EASA allows for both organizations and individuals to contribute. A simplified workflow is presented with Figure 1.



Figure 1: EASA Safety Reporting Data Flow.

Aviation Safety Reporting for Organizations relates to entities whose Competent Authority falls under EASA. This includes holders of type-certificates, production or major repair approvals, maintenance, continuing airworthiness or pilot training organizations, and air navigation service providers. For entities not under directly EASA, the reporting responsibility shifts instead to the respective national authority.

The requirement for reporting pertains to safety-relevant "occurrences"; events that endanger, or could endanger, an aircraft, its occupants, or others, which must be systematically reported, documented, protected, and analyzed. These are vital to detecting safety hazards proactively, rather than reactively responding to accidents. In doing so, the pipeline supports broader system-level safety improvements and risk management across the EU.

Regulations complement this pipeline by defining mandatory and voluntary reporting structures, timelines, data quality protocols, and follow-up procedures. Organizations are required to have processes in place for collecting, evaluating, and storing occurrence data, and for identifying hazards. Notably, mandatory reports must reach the competent authority within 72 hours of awareness, whereas voluntary reports should be submitted in a timely manner. When hazards are identified, preliminary findings and mitigative actions should be communicated within 30 days, with final assessments delivered no later than three months after initial notification.

Importantly, the system stresses the value of a Just Culture encouraging reporting by protecting individuals from unwarranted penalties and ensuring confidentiality and trust in the process. This helps maximize meaningful reporting and organizational learning, ultimately feeding into EASA's European Plan for Aviation Safety and enhancing aviation safety governance at both national and EU levels.

The concept of a Just Culture extends to individuals, as EASA maintains anonymity protocols to encourage self-reporting. Specifically, it operates a Confidential Safety Reporting (CSR); an independent mechanism that supplements existing mandatory reporting. It allows individuals including pilots and air traffic controllers to maintenance and airport personnel to voluntarily report safety-related malpractices or irregularities without fear of retaliation. The system ensures strong confidentiality protections and encourages reporting of otherwise undisclosed safety hazards. The information collected is de-identified and used to enhance aviation safety through hazard detection and risk management improvements. A dedicated, specially trained CSR team handles these reports and protects the identity of the reporter.

To submit a CSR, reporters complete an online form and avoid sending the content elsewhere within EASA. Reports are handled securely, the reporter's identity is kept confidential under data protection regulations, namely Regulation (EU) 2018/1725 with only CSR team members having access to the information. Depending on the case, EASA may investigate directly or forward information to the competent national or third-country authority, always while preserving confidentiality. Reporters may be contacted for clarification, and the outcome of the investigation may be shared without jeopardizing anonymity.

Enabling this pipeline is the European Coordination Centre for Accident and Incident Reporting Systems (ECCAIRS). It is a web-based platform developed to support the standardized collection, sharing, and analysis of aviation safety occurrences across Europe. Its goal is to provide a seamless tool for reporting and exchanging data, while operating in a secure, centralized environment and maintaining a separation between national and central repositories. The original version has been around since the 1990s, while an updated version of ECCAIRS was launched in 2020 in a portal format. It uses open-source and cloud-based technology to deliver better usability, scalability, and interoperability, ensuring alignment with ICAO's global standards. It also strengthens data quality and integration through initiatives including Data4Safety, providing regulators and stakeholders with a robust evidence base for identifying trends, prioritizing risks, and defining preventive safety actions across the aviation system. A standardized file type is also used, mainly by organizations that can support the IT requirements. The E5X file format allows transfer of occurrence data from internal databases directly into the ECCAIRS environment.

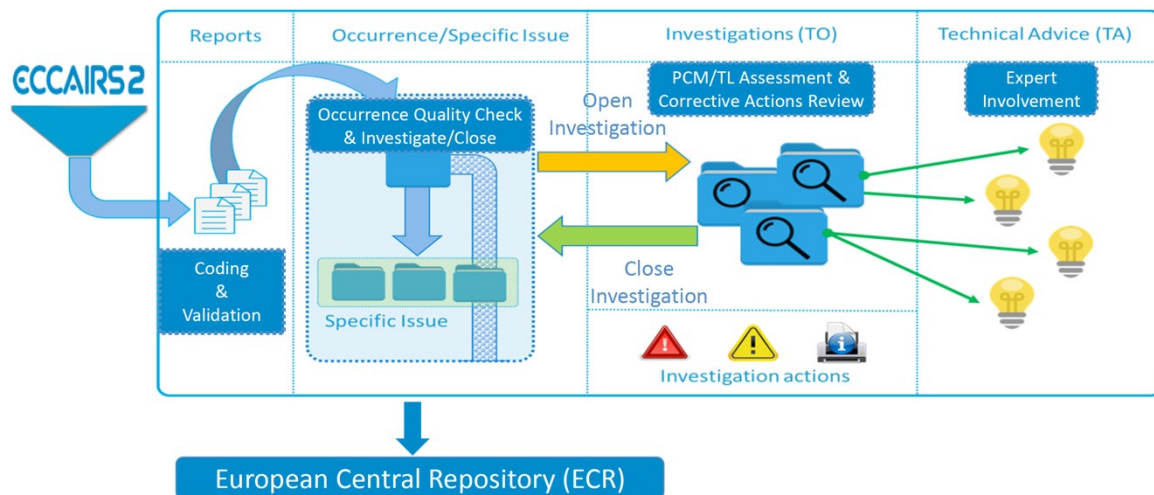


Figure 2: EASA Safety Data Management Process.

In the direction of data handling, a Safety Data Management (SDM) team operates within the Strategic & Safety Management Directorate, under the Safety Intelligence and Performance subdivision¹. The aim of SDM is to process all occurrences received by EASA in its role as a competent authority in a timely and proper manner. These data are then shared within EASA, while also providing monitoring, support and analysis. Occurrences are also grouped according to specific issues, enabling technical owners to monitor these issues over time and develop a sound understanding of the nature of the issue and the risks posed. Occurrences that are closed without review are systematically reviewed to identify any unaddressed safety issues. The complete process is depicted in Figure 2.

One of the goals of EASA's safety assessments is the reduction of accidents, serious incidents, fatalities, and injuries across all aviation domains like commercial and general aviation, aerodromes, and air traffic management - air navigation services (ATM/ANS). This is measured through Safety Performance Indicators (SPIs), including Tier 1 metrics like fatal accident rates and Tier 2 metrics focusing on Key Risk Areas (KRAs) such as runway excursions or airborne collisions. EASA deploys a suite of sophisticated methods to achieve this goal. Central is the European Risk Classification Scheme (ERCS), a standardized methodology that quantifies risk through a two-dimensional matrix assessing severity (potential consequences if an occurrence escalates) and probability (likelihood of escalation based on barrier failures). The ERCS enables consistent risk prioritization across domains, revealing hidden high-risk serious incidents that might be overlooked in traditional accident-focused analyses.

The European Plan for Aviation Safety (EPAS) serves as the Regional Aviation Safety Plan (RASP) for EASA Member States within the ICAO EUR region. Developed by EASA with input from national authorities and industry experts, EPAS outlines Europe's key safety priorities, identifies major risks, and sets targeted actions to address them. Using a systemic approach and the SRM process, the plan aims to continuously enhance aviation safety across the region. The key safety risks and related mitigation actions included in EPAS are identified through the five steps presented with Figure 3.

¹ "Occurrence Reporting for EASA Organizations" *Safety Data Management (SDM) Webinar Series*. November 6th, 2024.



Figure 3: The European SRM process².

- **Identification of Safety Issues:** Safety issues are identified through data analysis and undergo preliminary assessment by EASA, forming domain-specific risk portfolios with prioritised issues.
- **Assessment of Safety Issues:** High-risk issues are further evaluated via the Safety Issue Assessment (SIA) and Best Intervention Strategy (BIS), leading to recommended mitigation actions for EPAS.
- **Definition and Programming of Safety Actions:** Based on the SIA/BIS outcomes, proposed safety actions are reviewed, approved, and included in EPAS, with urgent actions fast-tracked if needed.
- **Implementation and Follow-up:** Approved EPAS actions are implemented and monitored, covering areas like rulemaking, research, national tasks, promotion, and evaluation.
- **Safety Performance Measurement:** Safety performance is monitored through the Safety Performance Framework (SPF) using key indicators, and findings feed back into the SRM cycle for continuous improvement.

In the following figures EASA categorization is represented based on frequency and ERCS Score for three main domains of aviation:

- Aeroplanes
- Aerodromes and ground handling
- Air Traffic Management / Air Navigation Services

Specifically, it demonstrates the safety issues by number of occurrences and aggregated ERCS Score and the Key Risk Areas by number of occurrences and aggregated ERCS Score.

A) Safety risks for large aeroplanes (CAT airlines, air taxi and NCC business)

The key risk areas for this domain are highlighted in Figure 4 and are defined by their potential accident outcome and by the immediate precursors of that accident outcome.

² European Union Aviation Safety Agency (EASA). (2022). Annual Safety Review 2022

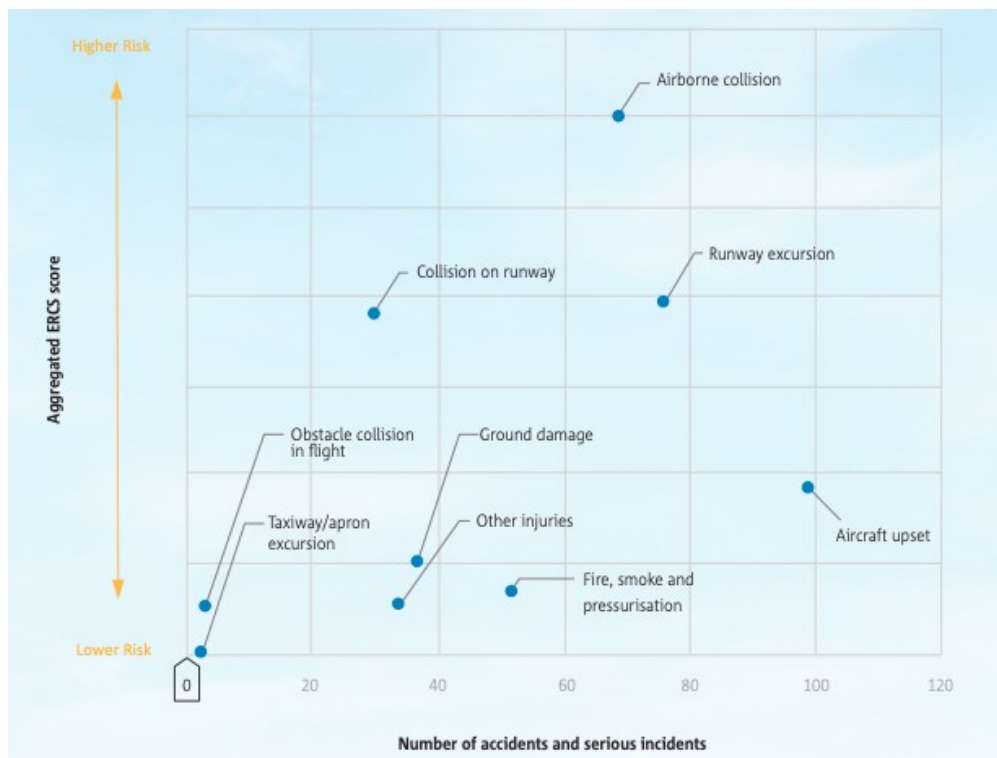


Figure 4: Key risk areas by aggregated ERCS score and number of risk-scored occurrences involving commercial air transport airlines and air-taxi

It can be concluded that the highest risk key risk areas are:

- **Airborne Collision:** Involves actual or potential mid-air collisions between aircraft or with other airborne objects (excluding wildlife); key 2021 risk factors included loss of separation with student pilots, TCAS resolution advisories, and near misses with drones. Managed mainly through ATM/ANS safety risk portfolio.
- **Runway Excursion:** Refers to incidents where an aircraft veers off or overruns the runway without becoming airborne; main 2021 contributors included incorrect take-off parameters, landing gear failures, tyre-bursts, and excursions during severe weather.
- **Collision on Runway:** Covers collisions or near-collisions on a runway involving aircraft, vehicles, or people (excluding wildlife); key 2021 risks involved runway incursions by aircraft and vehicles. Managed through ATM/ANS and Aerodrome/Ground Handling portfolios.

In Figure 5, the safety issues from the large aeroplanes data portfolio are listed along with their occurrence counts and risk scores.

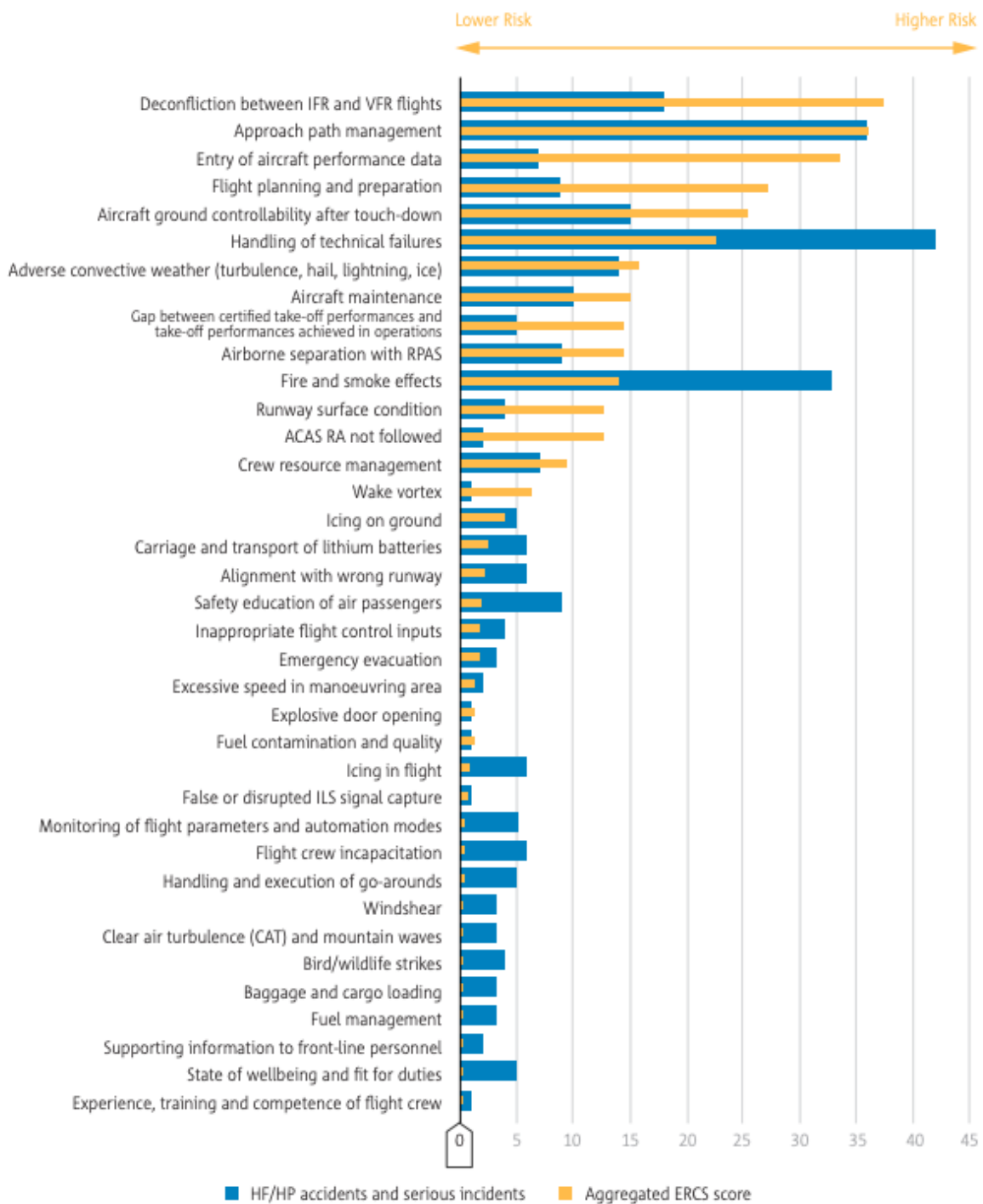


Figure 5: Safety issues by aggregated ERCS score and numbers of accidents and serious incidents involving commercial air transport airline and air-taxi and non-commercial business operations.

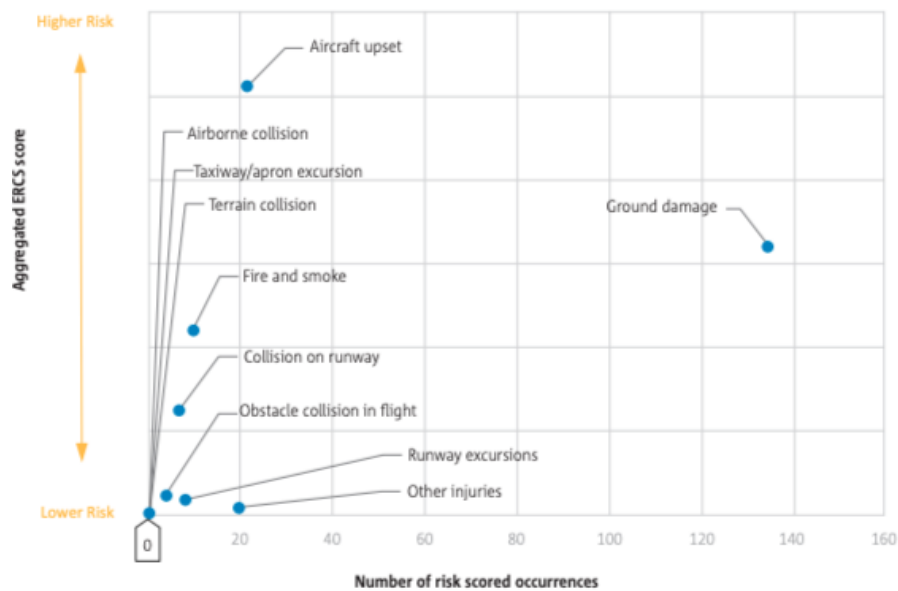


Figure 6: Key risk areas by aggregated ERCS score and number of risk-scored occurrences involving aerodromes and ground handling

B) Safety risks for aerodromes and ground handling

The safety risks related to aerodromes and ground handling are based on accident and serious incident data collected from the EASA Occurrence Repository and the European Central Repository, covering the period from 2017 to 2021 (a total of 185 occurrences). The main key risk areas for this domain are highlighted in **Figure 6**.

In Figure 7, a comparison is presented between the number of occurrences per safety issue and their corresponding aggregated ERCS (European Risk Classification Scheme) scores. A significantly longer yellow bar relative to the underlying blue bar indicates that a low number of occurrences contribute to a high overall risk level.

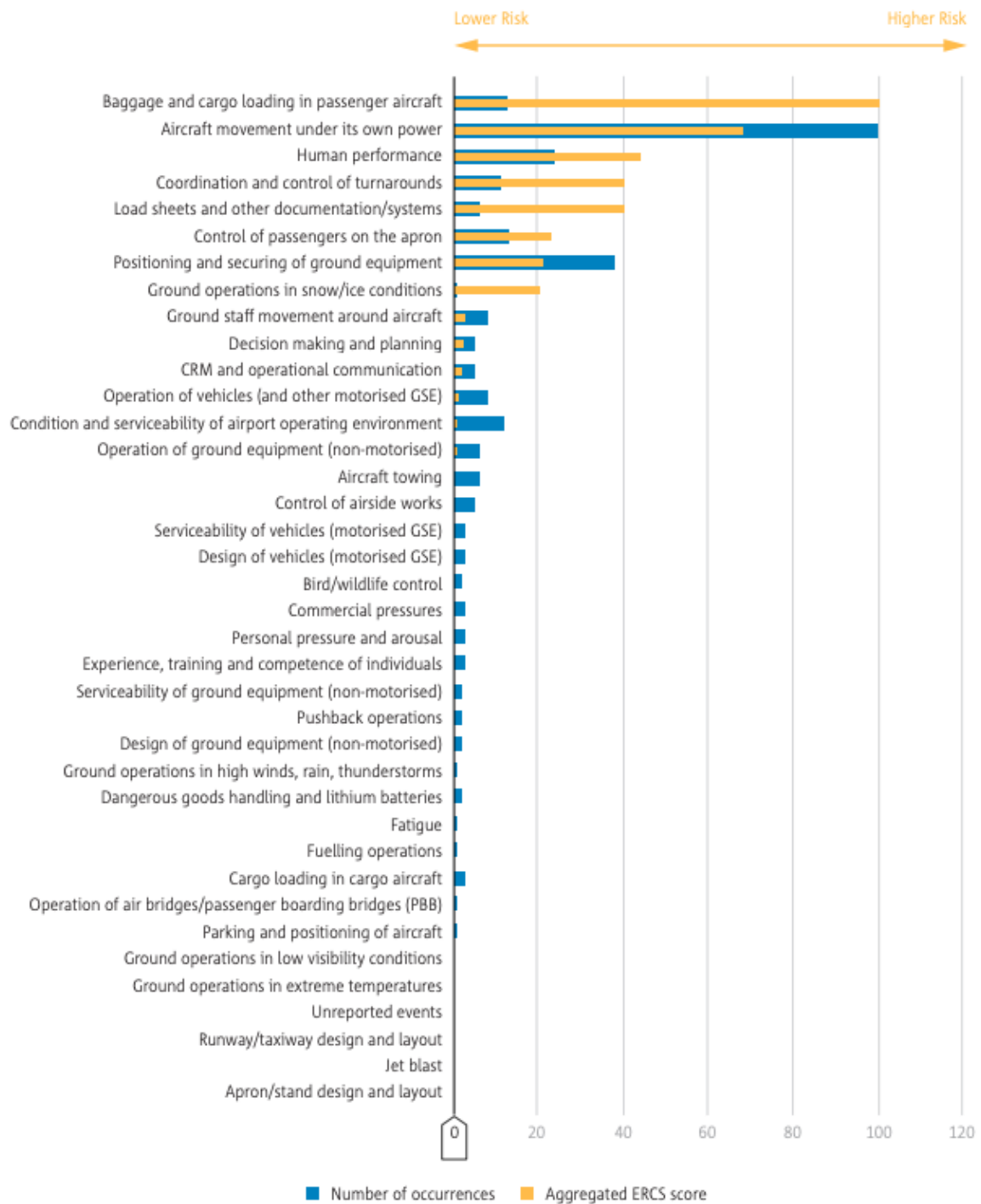


Figure 7: Safety issues by aggregated ERCS score and numbers of occurrences involving aerodromes and ground handling.

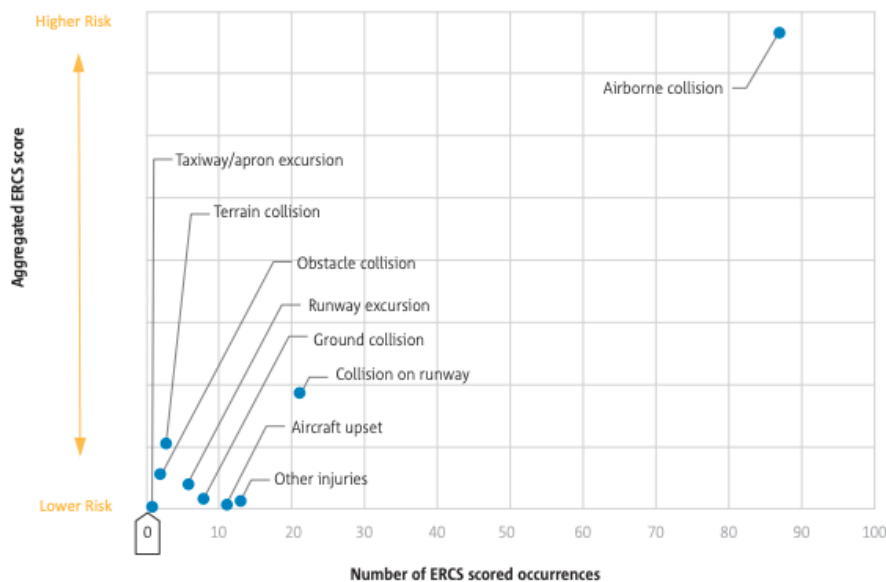


Figure 8: Key risk areas by aggregated ERCS score and number of risk-scored ATM/ANS occurrences.

C) Safety risks for Air Traffic Management / Air Navigation Services

Safety risks related to ATM/ANS are based on accident and serious incident data from the EASA Occurrence Repository and the European Central Repository, covering the five-year period from 2017 to 2021. Figure 8 highlights the main key risk areas in this domain, which are defined by their potential accident outcomes and the immediate precursors leading to those outcomes. Airborne collision and collision on runway are the top key risk areas in the ATM/ANS domain, highlighting the critical role of ATM/ANS in aircraft separation and navigation guidance.

In addition, key safety issues within the ATM/ANS domain were identified. Accidents and serious incidents were mapped to specific safety issues along with their associated ERCS scores to create the data portfolio. In the graph, a yellow bar significantly longer than the underlying blue bar indicates a small number of occurrences linked to a high risk. The results of this mapping are presented in Figure 9.

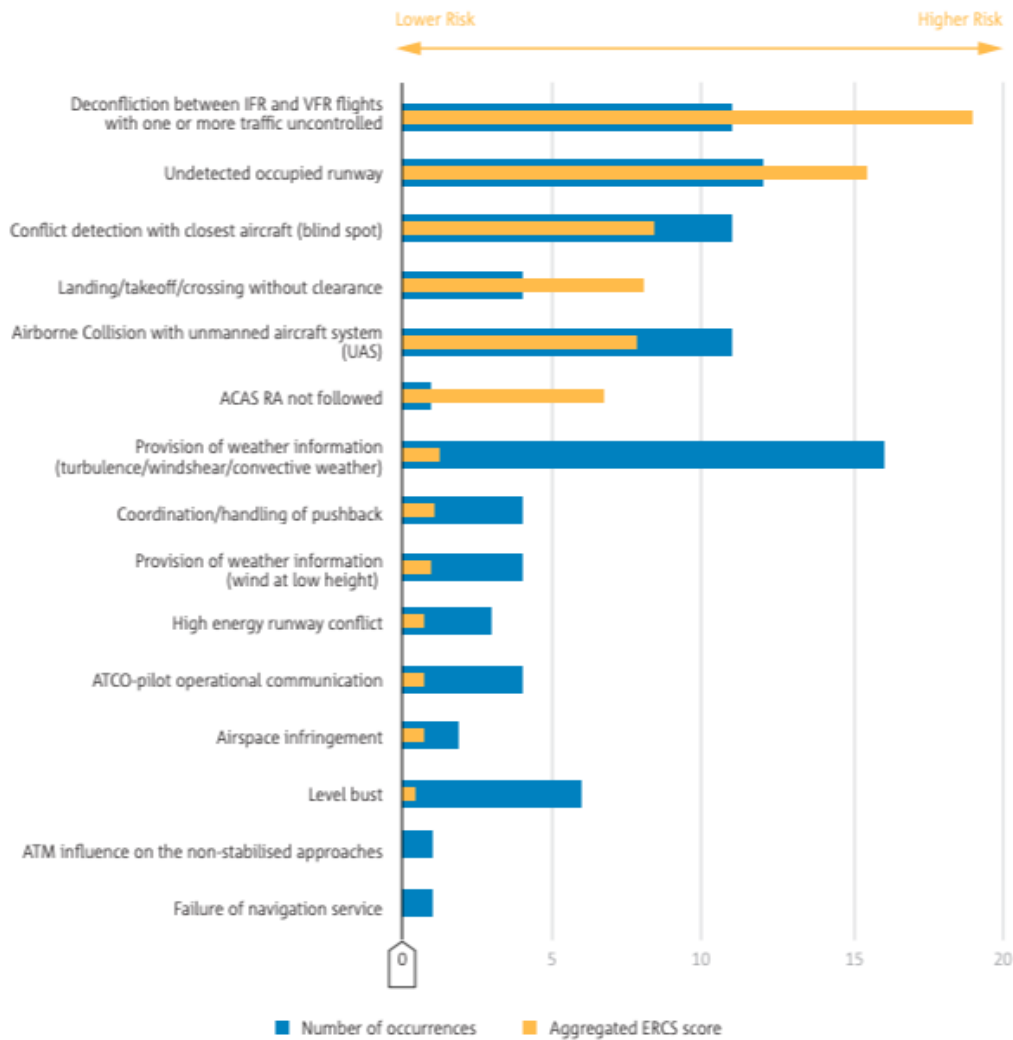


Figure 9: Safety issues by aggregated ERCS score and numbers of accidents and serious incidents for ATM/ANS safety issues.

Practical application of the methods and metrics is shared in EASA’s Annual Safety Review. According to the 2024³ report, the number of reports is steadily climbing, as is the rate of reports per 1000 movements, depicted with Figure 10. While the rate of reporting is increasing, this does not translate to increasing serious occurrences, as presented with Figure 11. It showcases that the predominant form of occurrence is an incident, followed by occurrences without safety effects. Serious incidents and accidents constitute very small percentages of the reported cases.

³ “Annual Safety Review 2024” EASA.

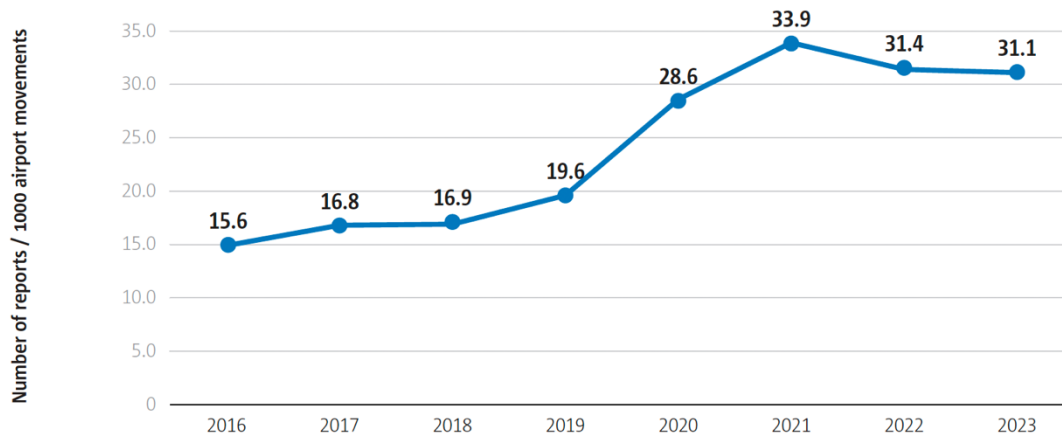


Figure 10: Reporting rate per 1000 aircraft movements².

With increasing data collection and availability, EASA recognizes the associated emergence of ML methods to complement the Big-Data analytics. The Artificial Intelligence (AI) Roadmap 2.0⁴, published in 2023, outlines the agency’s approach to certifying AI and ML systems in aviation. EASA also collaborates with the SESAR 3 Joint Undertaking to explore AI solutions for air traffic management, such as predictive maintenance tools and virtual assistant technologies.

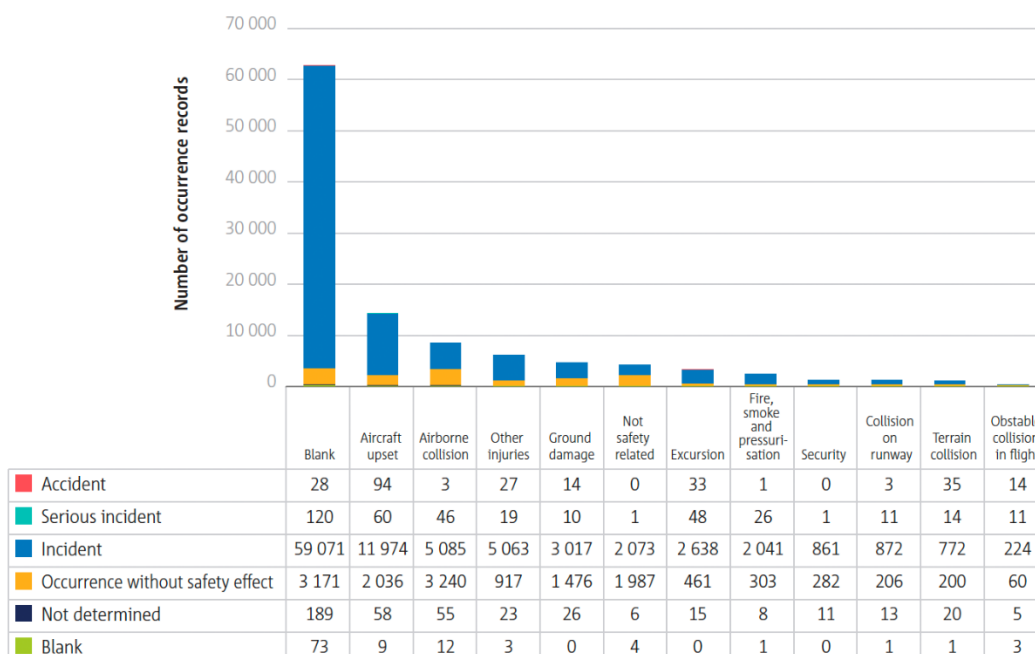


Figure 11: Number of occurrences per KRA and severity category in 2023.

⁴ “Artificial Intelligence Roadmap 2.0 – Human-centric approach to AI in aviation,” EASA, 2023.

2.2 Data4Safety

EASA’s Data4Safety (D4S) Programme represents a cornerstone of its research strategy, leveraging big data analytics to enhance predictive safety capabilities. It represents a voluntary, cooperative partnership within the European aviation community, designed to improve safety through the sharing and analysis of aviation data. By integrating diverse data sources including flight data, weather patterns, and occurrence reports, EASA aims to identify systemic risks before they escalate into accidents. A key milestone was the development of use cases aligned with the EPAS, focusing on high-priority risks such as runway excursions and air traffic control fatigue. The integration of D4S with ECCAIRS2, further enhances the Agency’s ability to analyze safety trends and implement proactive measures.

The development cycle for D4S is presented with Figure 12. EASA began with a feasibility study in 2015, which explored the potential of applying big data technologies to aviation safety. This study confirmed the technical viability of such a programme and laid the groundwork for the Proof of Concept (PoC) phase launched in 2016. During the PoC, EASA focused on validating the programme's governance model, data-sharing protocols, and analytical capabilities. Key outcomes included the development of standardized methodologies for data fusion and risk assessment, as well as the establishment of a dual governance structure involving industry and regulatory partners. D4S is currently at its development phase, with launch expected in 2026.

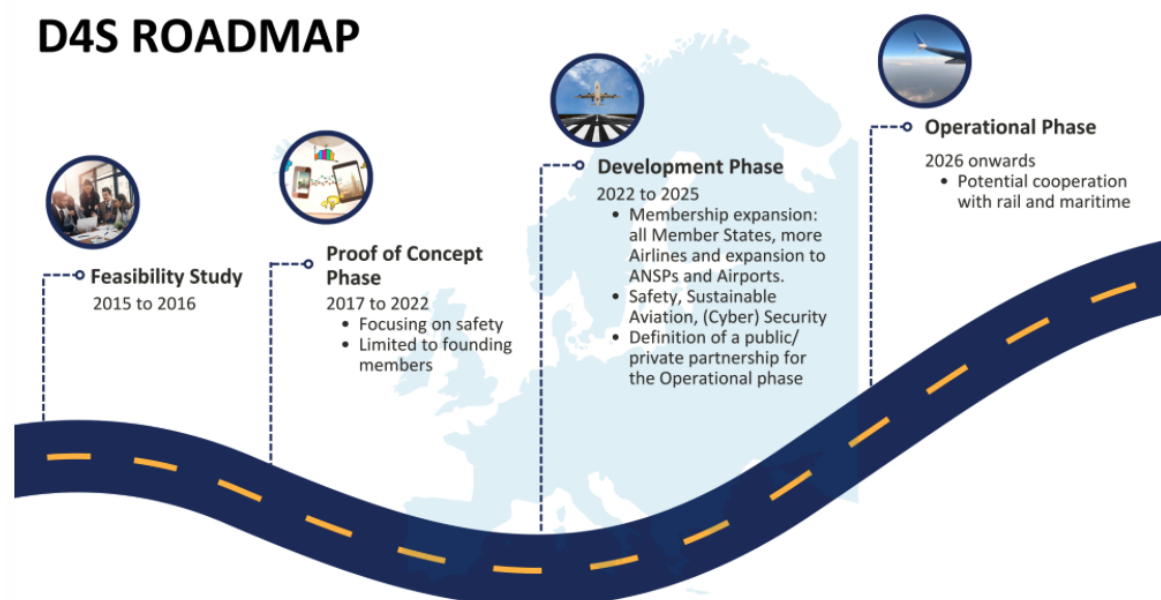


Figure 12: Data4Safety development roadmap.

The primary objective of D4S is to organize and analyze the vast and fragmented data stores scattered across European aviation organizations. The programme creates a critical mass of information enabling advanced data fusion techniques to evaluate safety risks comprehensively. Additionally, it focuses on predictive analytics using AI and machine learning to detect vulnerabilities and trends. D4S integrates machine learning models capable of identifying latent safety risks through pattern recognition in high-dimensional datasets. Applications include clustering of flight anomalies, predictive modelling of unstable approaches, and dynamic risk scoring based on weather and operational variables.

Key goals include structuring available data stores, facilitating data fusion, and providing a common platform for predictive risk assessment. The programme also supports members in improving their individual safety performance through metrics, blind benchmarking, and directed studies. D4S is strictly focused on safety enhancement and excludes use of targeted oversight or commercial gain, adhering to the principles of Just Culture as defined in EU regulations.

D4S operates under a dual governance model, reflecting its collaborative nature. The programme is overseen by a Steering Board (SB) of approximately 15 members, co-chaired by the EASA and another SB member. In 2023, the programme expanded its membership to include 41 organizations, fostering a collaborative approach to data sharing and analysis. Decisions are made consensually, ensuring alignment with the collective interests of the aviation community. Day-to-day operations are managed by a Technical Board (TB), which mirrors the SB in composition and co-chairmanship.

Data protection is a cornerstone of the D4S programme. The governance framework incorporates technical, procedural, and legal measures ensuring compliance with EU regulations and the General Data Protection Regulation (GDPR)⁵. Data is de-identified to safeguard privacy, though reversibility is maintained to allow for future data fusion. Members retain full control over their data and can request removal from the platform at any time.

Members of D4S can actively contribute to the diverse datasets. These include raw flight data from Flight Data Monitoring (FDM) systems, which provide detailed parameters such as speed, altitude, and engine performance, in addition to voluntary occurrence reports. Surveillance and weather data further enhance the contextual understanding of flight operations, enriching the platform’s analytical capabilities. The contents of Figure 13 describe the D4S platform at its proof-of-concept stage, as presented at EASA’s Aviation Safety Conference 2020⁶.

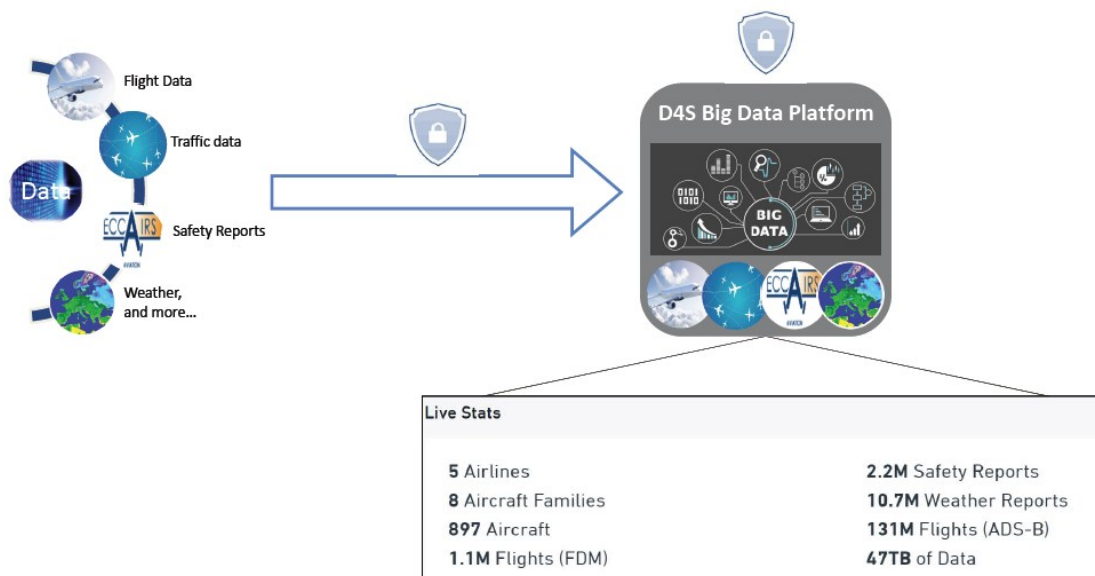


Figure 13: Proof of concept D4S Data.

⁵ REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL: “On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)”

⁶ “Data4Safety: The European Big Data programme for Aviation Safety” *Presentation at EASA Annual Safety Conference*. November 5th, 2020.

To ensure consistency and reliability, all contributors must comply with D4S-defined standards, including de-identification protocols and specific data retention periods. The Data Protection and Processing Organisation (DPPO) is responsible for all incoming data.

Although D4S is a European initiative, it aligns with international aviation safety programs such as the Federal Aviation Authority's (FAA) Aviation Safety Information Analysis and Sharing (ASIAS) and the International Air Transport Association (IATA) Flight Data Exchange (FDX). While the data itself is not exchanged across these platforms, D4S engages in continuous knowledge sharing to promote safety standards on a global scale.

One of EASA's significant research contributions is the development of detection algorithms for critical safety events, such as Unstable Approaches. Through collaborative efforts with airlines, manufacturers, and data experts, EASA defined a harmonized set of criteria and thresholds to identify these events across diverse fleets and operational environments. This work, documented in the Guidance for Identifying Unstable Approach with Flight Data (2022), involved analyzing over 1.4 million flights to refine parameters like airspeed deviations, vertical speed, and aircraft configuration. The resulting algorithm not only serves as a benchmark for operators but also demonstrates EASA's commitment to evidence-based safety solutions.

EASA has also spearheaded directed studies under the D4S framework to address emerging safety concerns. For example, research into Go-Arounds and their role as a mitigation barrier for Unstable Approaches provided actionable insights into operational practices and risk factors. These studies leverage the programme's unique ability to aggregate and anonymize data from multiple stakeholders, enabling a system-wide perspective that individual operators could not achieve independently. By publishing findings and best practices, EASA ensures that the broader aviation community benefits from these insights.

In addition to event-specific research, EASA is exploring the integration of advanced technologies into D4S as a big-data platform. Projects examining the use of AI / ML methods for predictive risk modeling highlight the agency's forward-looking approach. In this direction, the requirement for DemoQUAS and Task T6.3 is to develop AI/ML models leveraging data structures like those referenced for D4S and apply UQ methods on such tools. Focus is placed upon pilot and airport operations, as aspects with great variability in data.

3. Flight Operations – Pilot Evaluation

This part of the report describes the process conducted through collaboration of AUTH and EGA, using questionnaires, to evaluate pilot performance. This is done to collect data structured in accordance with EASA and international standardizations, to ultimately develop ML methods to proactively predict and evaluate the human factor in pilot operations. Student pilots constitute optimum test-subjects, given that they are still early in their path to pilot experts.

3.1 Questionnaires

3.1.1 Data Collection

EGA collects data from instructor pilots and trainee pilots through structured questionnaires and performance assessments. This data includes observations, self-assessments and feedback related to pilot training, competencies and operational factors.

Data is collected solely for research and training improvement purposes, in full compliance with ethical standards and GDPR. Data is also anonymized to safeguard individual identities.

All participants are informed by the Training Department about the purpose of the research project, how and when to complete the questionnaires and how to submit them. To support consistent engagement, the Training Department sends out a weekly email reminder with information about the ongoing project, encouraging voluntary participation.

The performance of the training system is assessed through a structured feedback process which serves to validate and improve the program, ensuring that it effectively builds pilot competencies and aligns with the organization's training objectives. The assessment materials are developed based on the ICAO competency model.

Participation in the project is completely voluntary, and individuals may choose to opt out at any time without any negative consequences. Access to the collected data is restricted to authorized research personnel only. All information is stored securely, handled transparently, and retained only for as long as necessary to meet the objectives of the project.

3.1.2 Data Handling and GDPR

All personal data collected is handled securely, used solely for research and training quality improvement purposes, and anonymized to protect individual identities. Access to data is strictly limited to authorized research personnel and is not shared with third parties outside the scope of the project.

Participants are informed of their rights under GDPR, including the right to withdraw consent at any time without penalty. Data is stored securely, retained only for as long as necessary, and disposed of responsibly in line with regulatory requirements.

3.2 Flight Assessment

The data structure for the questionnaires is developed in an approach for the eventual application of Bayesian Network modeling methods for the proactive prediction of pilot

performance given specific parameters to proactively predict potentially hazardous conditions. Toward this direction, an extensive literature review was conducted in close collaboration with 12 experienced pilot instructors from the pilot training academy, and 41 discrete factors were extracted, categorized into four classifications:

- Organizational Factors
- Human Factors
- Environmental Factors
- Pilot Competencies

Pilot Competencies encompass the core skills and capabilities that pilots are required to acquire and continually enhance through training, to perform flight operations safely and efficiently. These competencies are derived from the IATA White Paper on Competency-Based Training and Assessment⁷. In Figure 14 the categorization of the extracted factors is represented:

Organizational Factors	Environmental Factors	Human Factors	Pilot Competencies
Organizational Goals and Resource Allocation	Weather Conditions	Fatigue	Application of Knowledge
Safety Culture	Runway Conditions	Stress and Anxiety	Application of Procedures and Compliance with Regulations
Workload Management	Geomorphological Features	Attitude	Communication
Communication and Coordination within the Company	Bird Strikes and Wildlife	Personality Traits	Aeroplane Flight Path Management, Automation [FPA]
Organizational Climate	Air Traffic Density	Attitude towards Risk	Aeroplane Flight Path Management, Manual Control [FPM]
Procedures	Terrain and Obstacles	Physical Health / Body Size / Strength	Leadership and Teamwork
Supervision	Operational Materials	Mental Health	Problem Solving and Decision Making
Training Process	Facilities	Perception	Situation Awareness and Management of Information
Pressure	Communication with ATC	Complacency	Workload Management
Authority Gradient	Aircraft Design and Operational Suitability	Assertiveness	
	Aircraft Serviceability / Airworthiness	Personal Readiness	

Figure 14: Classification of factors influencing Pilot Performance.

This classification offers a solid basis for systematically analyzing how each factor influences pilot performance through the Bayesian Network. It enables a deeper understanding of training needs, helps uncover safety-related risks, and reveals patterns linked to performance variability. In addition, this categorization clearly defines each factor and ensures their mutual independence, which is essential for constructing an accurate and valid Bayesian Network.

To build the Bayesian network, it is essential to extract the relationships between the nodes and connect them with edges to represent probabilistic dependencies. In this direction, it was necessary to determine from real data both the way and the extent to which each factor affects pilot performance. Thus, four questionnaires were created, the results of which provided information not only on whether a factor from the categorization in Figure 14 affects pilot

⁷ IATA “Competency-Based Training and Assessment (CBTA) Expansion within the Aviation System” *White Paper*, March 2024.

performance but also included a rating that showed to what degree this factor influences pilot performance. The four questionnaires are:

A) Instructor Evaluation of Pilot Abilities - After Each Flight

The first questionnaire was completed by the pilot instructor after each dual flight. Specifically, the instructor evaluated the pilot's performance across several flight phases, including taxi-out, take-off, climb, cruise, descent, approach, landing, and taxi-in. For each phase, the student was rated on a 1–5 scale. A rating of 5 indicated that the student performed all exercises accurately following a single demonstration and minimal practice. A rating of 4 reflected that the student completed most exercises after the demonstration and several practice attempts, achieving a reasonable degree of accuracy within the allotted time. A score of 3 meant that the student required additional demonstration and/or practice to perform the exercise, with extra time being necessary. A rating of 2 indicated that the student had difficulty performing the exercise, needing frequent re-demonstration and additional practice, and that satisfactory completion within the allotted time was not possible, requiring further training. Finally, a rating of 1 reflected that the student experienced great difficulty in performing the exercise, with poor accuracy, inability to complete the exercise in the allotted time, and a clear need for additional training.

In the next section, the pilot instructor was requested to assess each human Factor related to the flight. These factors included attitude, personality traits, attitude towards risk, physical and mental health, perception, complacency, assertiveness, and personal readiness. Each factor was rated on a defined scale, reflecting the pilot's performance, behavior, and readiness during the flight.

For attitude, ratings ranged from very positive, indicating a highly constructive, cooperative, and safety-conscious approach, to very negative, reflecting dismissive, uncooperative, or potentially dangerous behavior. Personality traits were evaluated from very stable, representing extreme calmness, resilience, reliability, and openness to change, to very unstable, showing high impulsiveness, irritability, or inconsistency. Attitude towards risk spanned from very cautious, prioritizing safety, to very risk-prone, indicating adventurous or potentially unsafe risk-taking.

Physical health, body size, and strength were rated from excellent, denoting exceptional fitness and strength, to very poor, reflecting limitations that could significantly affect performance. Mental health was assessed from excellent, with no psychological issues, to very poor, where severe psychological challenges could reduce performance. Perception measured situational awareness, from highly accurate to highly distorted, indicating clear understanding or significant misinterpretation of the environment.

Complacency evaluated vigilance and engagement, ranging from highly proactive, showing active attention and error-checking, to overconfident, reflecting reliance on routine and potential disregard for safety protocols. Assertiveness covered communication and decision-making, from highly assertive, ensuring clear exchanges even under stress, to passive, indicating reluctance to contribute, which could impact teamwork and safety. Finally, Personal readiness, including rest, preparation, and fitness for duty, ranged from fully prepared, exemplifying optimal readiness, to unfit for duty, where lack of rest or preparation could compromise safety.

In the following sector, the pilot instructor is requested to evaluate each pilot ability for this specific flight. The assessment covers several key areas, including knowledge, procedural compliance, communication, flight path management, teamwork, decision making, situational awareness, and workload management. For each ability, the student was rated on a five-level scale. A rating of exceptional indicated that the student demonstrated thorough and detailed understanding and consistently applied it effectively in all situations with minimal assistance. A rating of proficient reflected that the student applied relevant knowledge and skills accurately and timely, showing good understanding and performance with little guidance. A rating of competent meant that the student generally applied knowledge and skills effectively but occasionally required guidance or reminders. A rating of basic indicated that the student struggled to apply knowledge and skills in complex situations, sometimes missing key information or making errors, requiring additional support. Finally, a rating of novice reflected that the student lacked sufficient knowledge or ability, frequently failing to apply it correctly, and needing significant improvement and supervision.

Finally, the instructor evaluated the overall pilot performance using the same grading scale applied to the flight phases. As highlighted in the questionnaires, the overall pilot performance is assessed relative to the pilot's experience and the objective of the lesson.

B) Pilot Assessment of Organizational Factors and Environmental Factors

In this questionnaire, trainee pilots were asked to evaluate management and organizational factors as well as environmental factors for a specific reporting period. The reporting period corresponds to the year in which the trainees completed the questionnaire.

In the first section, trainee pilots were asked to evaluate a series of management and organizational factors that influence pilot performance, based on their personal experience during a specific reporting period. For each factor, they were presented with several descriptive options and asked to select the one that best represented their perception of the organization.

They first assessed organizational goals and resource allocation, determining whether goals were aligned with safety and efficiency. A rating of well-aligned indicated that goals fully supported operational excellence and that resources were effectively allocated to safety. Mostly aligned reflected goals and resources that were generally supportive but with minor constraints. Partially aligned reflected inconsistency in prioritization, while conflicting indicated a tendency to prioritize profit over safety considerations, and incompatible described situations where goals and resources were entirely misaligned with operational standards.

The safety culture was rated according to the emphasis placed on safety and the effectiveness of the reporting system. A very high safety culture signified continuous improvement and full employee engagement, while high suggested strong participation for the most of employees. Moderate implied occasional lapses or delays in reporting, low reflected minimal emphasis on safety, and very low described an almost non-existent system.

Trainees also evaluated workload management. Highly effective indicated well-distributed workloads and strong measures against fatigue, whereas highly ineffective described poorly managed workloads that created severe stress. Ratings in between (effective, moderate, ineffective) reflected decreasing levels of balance and control over stress and fatigue. For communication and coordination within the company, the scale ranged from highly efficient,

with clear and consistent communication across departments, to highly inefficient, where miscommunication frequently led to significant errors. The organizational climate was described as very positive when pilots experienced a supportive and collaborative environment, and as very negative when the climate was toxic and marked by dissatisfaction. Intermediate ratings captured varying degrees of support and neutrality.

The procedures were assessed for clarity and adequacy: Comprehensive procedures were detailed and highly appropriate, while Absent meant procedures were almost non-existent or entirely ineffective, posing a high risk of error. Supervision was rated from Strong, reflecting supportive and proactive oversight, to Non-existent, indicating a lack of guidance and support. Similarly, the training process was rated from Thorough, ensuring excellent preparation, to Negligible, leaving pilots unprepared.

Finally, pilots assessed the authority gradient, which measured the balance between leadership and openness to input. Balanced indicated clear authority but open communication, while Rigid reflected a strict hierarchy where subordinates felt unable to voice concerns.

In the following section of the questionnaire, trainee pilots were asked to evaluate several aspects of the operational environment that influence performance. These included the technological environment, operational materials, facilities, and aircraft design and operational suitability. Each factor was assessed on a descriptive scale, ranging from highly supportive to severely inadequate, reflecting the support of safety and efficiency.

For the technological environment (equipment design, controls, interface characteristics, automation) ratings ranged from very user-friendly, where systems were intuitive, ergonomic, and automation enhanced decision-making with minimal effort, to highly incompatible, where systems were difficult to use, prone to error, and lacking operational effectiveness. Intermediate ratings such as generally suitable and adequate indicated mostly functional systems with minor challenges or occasional difficulties, while somewhat complicated reflected disruptive complexity requiring workarounds.

The operational materials category assessed manuals, checklists, documents, charts, and maps. Highly effective materials were complete, up-to-date, and readily available, while highly insufficient materials were outdated and regularly hindered performance. In between, effective and adequate indicated generally sufficient coverage with minor or occasional gaps, and insufficient reflected incomplete or outdated resources requiring significant workarounds.

Regarding facilities, responses ranged from excellent, where infrastructure fully supported operational needs, to severely inappropriate, where facilities created substantial challenges. Intermediate levels described conditions that were generally good but required minor improvements (good), merely functional but lacking important elements (acceptable), or frequently insufficient (inadequate).

Finally, aircraft design and operational suitability was rated from excellent, indicating designs perfectly aligned with training or operational requirements and ergonomically supportive, to unacceptable, describing aircraft entirely unsuitable for the mission due to severe ergonomic or usability issues. Good design supported most needs effectively, fair covered only the basics but introduced noticeable limitations, and poor indicated misalignment with training or operational goals.

C) Pilot Assessment - After Each Flight

In this questionnaire, student pilots were asked to assess factors influencing their performance during a specific flight. The evaluation covered management and organizational factors, environmental factors, and human factors.

In the first section, student pilots were requested to evaluate management and organizational factors for a specific flight. Only pressure was included for this questionnaire, and trainees assessed this factor using descriptive scales reflecting the degree to which the organizational environment supported safe and efficient operations. Thus, pressure was rated from balanced (well-managed, motivating without causing stress) to overwhelming (consistently unbearable, severely impacting performance). The other factors had similar descriptive scales appropriate to their nature, capturing the effectiveness, clarity, or alignment of organizational support and processes.

Student pilots were also asked to assess environmental factors for the flight, including communication with ATC, weather conditions, runway conditions, geomorphological features, bird strikes and wildlife, air traffic density, terrain and obstacles, and aircraft serviceability/airworthiness. Each factor was evaluated using scales specific to its characteristics, ranging from ideal or very favorable conditions to hazardous or severely inappropriate situations. For instance, weather conditions ranged from ideal (clear, calm, optimal visibility) to hazardous (very poor visibility, strong winds, precipitation, or freezing conditions posing danger). These assessments reflected how environmental conditions impacted safe and efficient flight operations.

Finally, trainees evaluated human factors affecting their performance, including fatigue and stress and anxiety. These factors were rated on descriptive scales indicating the degree to which pilots were rested, alert, or affected by stress. For example, fatigue ranged from well-rested (adequate rest, fully prepared and focused) to severely fatigued (minimal rest, late-night operation, significant safety risk). Stress and anxiety ranged from very low (minimal or no stress, optimal operation) to very high (extreme stress severely affecting performance). These ratings highlighted the influence of individual physiological and psychological states on flight performance.

D) Factor Interdependencies

In this questionnaire, instructors were asked to identify the interdependencies between various factors that include pilot performance, based on their experience. These factors were classified into four categories: management and organizational factors, environmental factors, human factors, and pilot abilities.

The organizational factors include goals and resource allocation, safety culture, workload management, communication and coordination, organizational climate, procedures, supervision, pressure, training process and authority gradient. Instructors indicated which human factors – such as fatigue, stress and anxiety, attitude, personality traits, attitude towards risk, physical and mental health, perception, complacency, assertiveness, and personal readiness – were affected by each organizational factor, highlighting the relationship between organizational environment and individual pilot performance.

Environmental factors such as communication with ATC, weather conditions, runway conditions, geomorphological features, bird strikes and wildlife, air traffic density, terrain and obstacles, technological environment, operational materials, facilities, aircraft airworthiness/serviceability, and aircraft design and operational suitability were evaluated to determine their effect on pilot abilities, including application of knowledge, application of procedures and compliance with regulations, communication, aeroplane flight path management (automation and control), leadership and teamwork, problem solving and decision making, situational awareness and management of information, and workload management.

Similarly, human factors such as fatigue, stress and anxiety, attitude, personality traits, attitude towards risk, physical and mental health, perception, complacency, assertiveness, and personal readiness were examined to understand their influence on the same pilot abilities, highlighting the combined impact of organizational, environmental, and individual human factors on pilot performance.

4. Machine Learning for Aviation Safety

As established, the aviation industry constitutes a vital part in modern transportation by facilitating both international trade and personal mobility. Each passing year, it maintains an increasing growth rate which threatens an already over-constrained system. The rise in air traffic highlights society's growing dependence on air transportation which, in turn, triggers increasing demand for commercial aircraft and related services. However, this increase in demand guarantees that the features required to maintain the sector's safety standards are challenged.

To ensure safety in air travel, existing safety protocols require the addition of radical technological advancements that help the industry keep up with modern challenges. Aviation-related accidents can be the result of adverse weather conditions, technical malfunctions or improper maintenance, among others. Despite the presence of numerous contributing factors, most aviation accidents are attributed to human error(s) (estimated about 70%^{8 9}).

Achieving optimal safety levels can be a complex task. Given the impact of human factors, efforts towards improving aviation safety are directed towards automation. This is driven by technological innovation and with a goal to enhance regulatory oversight. Such developments include advanced air traffic control systems and predictive maintenance algorithms¹⁰. With machine learning technologies in the forefront of technological affairs, their influence is starting to become evident in the aviation industry as well¹¹. This part of the current report delves into the specifics on how ML technologies are applied in support of aviation safety, along with the proposed implementations that are explored within the context of the DemoQUAS research project and Task 6.3.

4.1 Literature Review (State-of-the-art)

With the rise of ML technologies and the introduction of Artificial Intelligence (AI) to enhance many aspects of human affairs, it can be inferred that the concept has found its way into aviation-related matters. This is evidenced by the current state-of-the-art research regarding aviation, especially in the domain of pilot behavior and airport operations. By using ML techniques, research groups seek to harness high-dimensional datasets, which can often be beyond the scope of traditional statistical methods to interpret in a meaningful manner. Following are some categories that represent the pinnacle of ML application in aviation research:

- **Predictive Maintenance:**

This is the most effective use case of machine learning algorithms to prevent component failures in aircraft. By analyzing historical and real-time sensor data, machine learning can detect

⁸ European Union Aviation Safety Agency (EASA). (2022). Annual Safety Review 2022

⁹ Burns, K., and Bonaceto, C., "An empirically benchmarked human reliability analysis of general aviation," Reliability Engineering & System Safety, Vol. 194, 2020, 106227

¹⁰ Alreshidi, I., Moulitsas, I., and Jenkins, K., W., Advancing Aviation Safety Through Machine Learning and Psychophysiological Data: A Systematic Review, Journals & Magazines, IEEE Access, Vol. 12, 2024, pp. 5132 – 5150

¹¹ Demir G., Moslem, S., and Duleba S., Artificial Intelligence in Aviation Safety: Systematic Review and Biometric Analysis, International Journal of Computational Intelligence Systems, Vol. 17, No. 279, 2024

potential issues and enable maintenance before failures occur. In this context, Brown et al.¹² developed a project-based learning framework that integrated logistic regression methods and machine learning techniques to enhance engineering education while addressing aviation system vulnerabilities and cybersecurity threats. Their work demonstrates how real-world use cases, such as predictive maintenance, can be incorporated into interdisciplinary design activities to prepare engineering students for challenges in modern aviation systems.

- Flight operations optimization:

Machine learning improves flight operations by forecasting air traffic and adjusting flight schedules. It uses inputs like weather conditions, traffic density, and aircraft performance to recommend optimal flight paths and altitudes, enhancing fuel efficiency and reducing delays. In a related work¹³, a software environment was developed that integrated aircraft trajectory data with aeronautical information, enabling interoperability with machine learning tools for improved flight planning and airspace management.

- Anomaly Detection in Aviation Cybersecurity:

With aviation's growing reliance on digital systems, cybersecurity is critical. Machine learning is used to monitor avionics data channels in real time, detect anomalies, and respond to potential cyber threats. Garcia et al.¹⁴ proposed a roadmap for adapting AI and machine learning cybersecurity techniques to aviation security engineering and airworthiness, while addressing implementation challenges within the regulatory framework of the aviation industry. This also includes streamlining air-traffic control by introducing automation and enhancing the decision-support-tools for controllers.

- Human Factors Assessment:

Machine learning helps evaluate pilot performance, workload, and decision-making. By analyzing simulator and real-flight data, it supports optimized training programs and human reliability in operations. In a recent work, Flávio L. Lázaro et al.¹⁵ applied machine learning algorithms to aviation incident reports using NLP to identify patterns, including human factors, contributing to safety incidents. The results showed that machine learning models can accurately predict contributing factors and support the prevention of future incidents.

Integrating ML techniques into aviation engineering follows a prevailing methodology structure used for similar applications¹⁶. The key stages include:

¹² Brown, W. L., Dabipi, I., Sharma, D., Zhang, L., Zhu-Stone, W., Mei, L., Wiggins, U. T., Cornelius, T. T., Jones, S., Wescott, R., Sharp, J. A., and Glenn, F. "The Investigation of Logistic Regression Methods Applied to Engineering Education using Project Based Learning for Airport Systems Design," IEEE Frontiers in Education Conference (FIE), 2021

¹³ Morales, C., Sanz, J., & Moral, S., "Design of a software environment to support machine learning analysis of aircraft trajectories (EIWAC 2017). EIWAC 2017 5th ENRI International Workshop on ATM/CNS, 2017

¹⁴ Garcia, A. B., Babiceanu, R. F., and Seker, R. "Artificial Intelligence and Machine Learning Approaches For Aviation Cybersecurity: An Overview," Integrated Communications Navigation and Surveillance Conference (ICNS), 2021

¹⁵ Lázaro, F.L.; Nogueira, R.P.R.; Melicio, R.; Valério, D.; Santos, L.F.F.M. "Human Factors as Predictor of Fatalities in Aviation Accidents: A Neural Network Analysis". *Applied Science* 2024. <https://doi.org/10.3390/app14020640>

¹⁶ Koul, P., "A Review of Machine Learning Applications in Aviation Engineering," *Advances in Mechanical and Materials*, Vol. 42, No. 1, 2025

- **Data Collection and Preprocessing:**
Gathering of high-quality datasets. This is the first step for implementing ML in aviation engineering¹⁷. Depending on the application, they can come from aircraft sensors¹⁸, maintenance logs, or measurements of environmental data, among others. Preprocessing involves cleaning, normalization, transformation, outlier detection, and missing data imputation.
- **Model Selection:**
Choosing the right machine learning model depends on the application. Supervised learning (e.g., decision trees, neural networks) is used for applications like Predictive Maintenance¹⁹, while unsupervised learning (e.g., clustering) can help in anomaly detection²⁰. Reinforcement learning may be used for adaptive decision-making.
- **Training and Testing:**
Models are trained on existing data and tested on unseen data using performance metrics²¹ like accuracy, precision, recall, and F1-score. Iterative tuning through cross-validation and hyperparameter optimization is often needed.
- **System Integration:**
After the appropriate training process, ML models are integrated into existing aviation software systems used for flight operations, maintenance, and safety. This requires collaboration with engineers and compliance with regulations, along with ongoing model updates.
- **Human Integration and Model Trustworthiness:**
A critical aspect of ML in any application is the use of and impact on human users. Given the potential consequences of mishandling or misuse, it is paramount that all steps of ML development and employment are appropriately archived to ensure a trustworthy AI component that is interpretable and explainable²² through all its aspects. ML methods aim to enhance system performance and support human decision-making without posing critical threats to safety or efficiency. Clear interface design and understandable outputs are crucial for effective use by aviation professionals, along with training to foster system adoption.

¹⁷ Kabashkin, I., Misnevs, B., and Zervina, O. "Artificial Intelligence in Aviation: New Professionals for New Technologies," Applied Sciences, 13(21), 11660, 2023

¹⁸ Kerle, N., "Real-time data collection and information generation using airborne sensors," Geospatial information technology for emergency response, Book series 6, 2008, pp. 43-74

¹⁹ Hasan, G. M., Hasan, M., Liu, P., Rad, M., Bernier, E., and Hall, T. J., "Optical Wavelength Meter with Machine Learning Enhanced Precision," Centre for Research in Photonics, Cornell University, 2022

²⁰ Jacko, J. A., "Human-Computer Interaction. Interacting in Various Application Domains. 13th International Conference, HCI International San Diego, CA, USA, July 19-24, 2009, Proceedings, Part IV

²¹ Rahman, Md., A., Bhuiyan, T., M. Ameer Ali "Enhancing aviation safety: Machine learning for real-time ADS-B injection detection through advanced data analysis," Alexandria Engineering Journal, Vol. 126, 2025, pp. 262-276

²² Protopapadakis G., Apostolidis A., Kalfas A., 'Explainable and interpretable AI-assisted Remaining Useful Life Estimation for Aeroengines', [2022]

4.2 Pilot Operations

Understanding and improving pilot performance is essential to aviation safety, especially considering the wide range of factors that can influence human behavior during flight. Although numerous studies and books address this topic, much of the existing research tends to rely on interviews or cultural assessments, lacking a comprehensive, data-driven methodology capable of capturing the full scope of contributing variables. Within Work Package 6 and Task 6.3, the goal is to study pilot performance using Bayesian Network ML techniques based on data collected by EGA, as described in section 3.

4.2.1 Bayesian Network Introduction

A Bayesian Network (BN) is a probabilistic model that uses a directed acyclic graph (DAG) to illustrate a group of random variables and the conditional relationships between them. They are a subfield of ML that applies the principles of Bayesian inference to build predictive and decision-making models²³. The approach helps with modeling uncertainty and allows for updating prior beliefs using new data. It is particularly suitable to integrate factors from different domains or resources to examine a complicated issue. In BN each node represents a random variable, and the directed edges show how these variables are probabilistically dependent. The most important advantage that BNs provide is their contribution to uncertainty evaluation and reasoning, as well as their ability to revise and continuously update probability estimations as new evidence and information become available.

Bayesian networks are composed of nodes that are connected by directed edges in an acyclic structure. Each node corresponds to a random variable, which may be either discrete or continuous²⁴. The directed edges run from parent nodes to child nodes, representing probabilistic dependencies among the variables. Each node is associated with a conditional probability distribution, typically represented in the form of a Conditional Probability Table (CPT), which specifies the likelihood of its possible states given the states of its parent nodes. In a Bayesian network, the size of the CPT grows exponentially with the number of parent nodes, following the formula X^N , where X denotes the number of states per parent and N is the number of parent nodes.

Bayesian Networks utilize different forms of probabilities to represent and quantify the complex relationships and dependencies among random variables in a structured and compact way. By capturing both direct and indirect probabilistic influences through a directed acyclic graph, these networks allow for efficient reasoning under uncertainty, inference of hidden variables, and updating of beliefs considering new evidence. The main types of probabilities used in Bayesian Networks are represented below:

²³ Bharadiya, J., "A Review of Bayesian Machine Learning Principles, Methods, and Applications," *International Journal of Innovative Science and Research Technology*, Vol. 8, 2025, pp. 2033-2038

²⁴ Londner, E., H., and Moss, R., J., "Bayesian Network Model of Pilot Response to Collision Avoidance System Resolution Advisories," *JOURNAL OF AIR TRANSPORTATION*, Vol. 26, No. 4, 2028

- **Prior Probability – $P(A)$:**

This refers to the probability of a variable taking a specific value without considering any other variables. It applies to variables that have no parent nodes.

- **Conditional Probability – $P(A/B)$:**

The likelihood of a variable assuming a particular state, given that another variable (or a set of variables) is in a known state.

- **Posterior Probability – $P(A/Evidence)$:**

The probability of a variable after incorporating new evidence or observations. It is typically computed through **Bayes' Theorem**.

- **Joint Probability Distribution:**

This expresses the combined probability of a particular configuration of all variables in the network. It is calculated by multiplying the conditional probabilities of each variable based on its parent nodes. The joint probability distribution P is defined by:

$$P(X_1, X_2, \dots, X_n) = P(X_1 | X_2, \dots, X_n) P(X_2 | X_3, \dots, X_n) \dots P(X_{n-1} | X_n) P(X_n) \quad \text{Eq. 1}$$

The value P represents the joint probability distribution across all possible combinations of variable values in the network. In principle, once P is known, the outcome of any given scenario can be determined by summing out the irrelevant variables. In a Bayesian network consisting of n nodes, the full joint probability distribution could, in theory, involve at least 2^n possible outcomes. However, it's not necessary to compute every combination explicitly. This is because Bayesian Belief Networks inherently follow the local Markov property, which states that each node is conditionally independent of all other non-descendant nodes, provided its parent nodes are known. Thus, Eq. 1 takes the following form:

$$P(X_1, X_2, \dots, X_N) = \prod_{i=1}^n P(X_i | \text{Parents}(X_i)) \quad \text{Eq. 2}$$

Inference refers to the method of estimating the probability of a particular variable based on known or observed data. Within a Bayesian Network, two primary forms of inference can be applied: forward inference (from cause to effect) and backward inference (from effect to cause).

Forward inference, or predictive reasoning, uses prior knowledge embedded in the network's structure and CPTs to estimate the likelihood of a target variable—such as pilot performance—based on observed evidence like fatigue, weather, or air traffic. By propagating information from parent to child nodes and applying the law of total probability, the probabilities of the queried variable are updated, leading to effective predictions.

Backward inference, or diagnostic reasoning, estimates the probabilities of underlying variables—such as fatigue, weather conditions, or air traffic density—based on observed evidence like pilot performance. This method propagates information in the reverse direction of the network, moving from child nodes back to their parents. By applying Bayes' theorem, it recalculates the likelihood of potential causes conditioned on the observed effects, enabling the identification of factors that most likely contributed to the given outcome.

4.2.2 State-of-the-Art

The use of BNs in the field of aviation safety is well established. These models are particularly effective for reasoning under uncertainty, as they can capture non-linear dependencies in highly complex systems. In their work, Adedigba et al.²⁵ implement a BN within a non-sequential, barrier-based accident model, where each safety barrier is represented by a distinct BN, and their interactions are also governed by Bayesian principles. This modeling approach has been widely adopted in the literature for the identification of causal factors in aviation incidents, often through the analysis of safety reports derived from the Aviation Safety Reporting System (ASRS)²⁶ and the National Transportation Safety Board (NTSB) database²⁷. Greenberg et al.²⁸ utilize BNs to evaluate accident probabilities in large commercial aircraft, while other researchers have applied the method specifically to runway safety, identifying contributing factors in runway overruns²⁹ and excursions³⁰.

A more recent study³¹ develops a BN model to estimate the risk of mid-air collisions, focusing on the interplay between human, technical, and systemic factors. Their workflow is depicted with Figure 15.

²⁵ Adedigba, S. A., Khan, F. and Yang, M., "Process accident model considering dependency among contributory Factors," *Process Safety and Environmental Protection*, Vol. 102, 2016, pp. 633-647

²⁶ Zhou, Z., Yu, X., Zhu, Z., Zhou, D. and Qi, H., "Development and application of a Bayesian network-based model for systematically reducing safety risks in the commercial air transportation system," *Safety Science*, Vol. 157, 2023

²⁷ Zhang, X. and Mahadevan, S., "Bayesian network modelling of accident investigation reports for aviation safety Assessment," *Reliability Engineering and System Safety*, Vol. 209, 2021

²⁸ Greenberg, R., Cook, S. C. and Harris, D., "A civil aviation safety assessment model using a Bayesian belief network (BBN)," *The Aeronautical Journal*, Vol. 109, 2005

²⁹ Calle-Alonso, F., Perez, C. Z. and Ayra, E. S., "A Bayesian-network-based Approach to Risk Analysis in Runway Excursions. *Journal of Navigation*", Vol. 72, 2019, pp. 1121-1139

³⁰ Ayra, E. S., Insua, D. R. and Cano, J., "Bayesian Network for Managing Runway Overruns in Aviation Safety," *Journal of Aerospace Information Systems*, Vol. 16, 2019.

³¹ Bauranov, A., and Rakas, J., "Bayesian network model of aviation safety: Impact of new communication technologies on mid-air collisions," *Reliability Engineering and System Safety*, Vol. 243, 2024

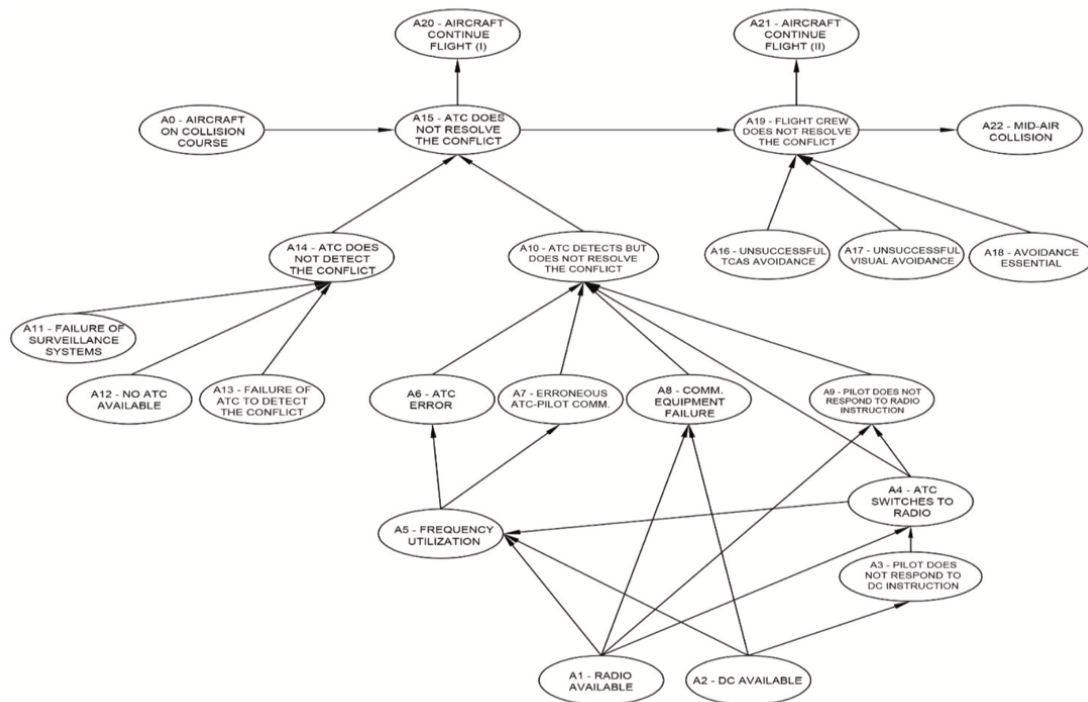


Figure 15: Mid-air collision BBN (adapted from³²)

Furthermore, work has been done³³ to explore the link between flight delays and aviation safety risks using a Bayesian framework. Bayesian Networks can also be applied to predict and assess factors influencing pilot and overall crew performance. In a related and recent study³⁴, proposed the implementation of Dynamic Bayesian Networks (DBNs) for real-time prediction of pilot fatigue during long-haul flights. This method presents an innovative approach, offering accurate, realistic, and actionable fatigue assessments, thereby contributing significantly to enhancing aviation safety, as presented with Figure 16.

³² Bauranov, A., and Rakas, J., “Bayesian network model of aviation safety: Impact of new communication technologies on mid-air collisions,” Reliability Engineering and System Safety, Vol. 243, 2024

³³ Wang, H. and Gao, J., “Bayesian Network Assessment Method for Civil Aviation Safety Based on Flight Delays,” Application of Discrete Mathematics in Urban Transportation System Analysis, 2013

³⁴ Zhou, Y., Chen, D., Xiao, J., Xiao, Y., Lu, Y., and Zhang, Y., “A Pilot Fatigue Prediction Method Based on Dynamic Bayesian Networks,” Human Factors and Ergonomics in Manufacturing & Service Industries, Vol. 35, 2025

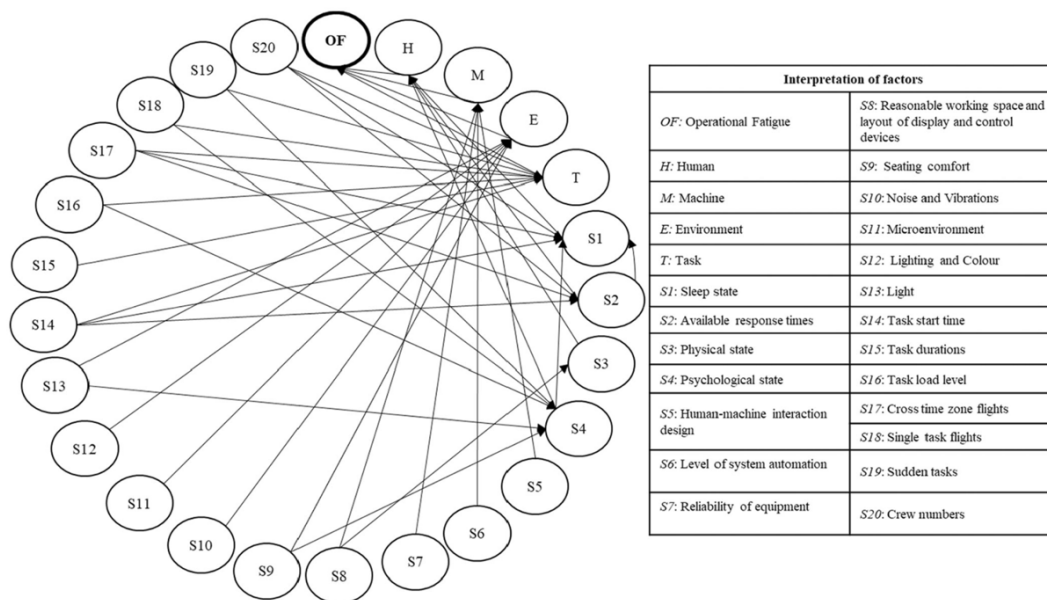


Figure 16: Topology of a Dynamic Bayesian Network (adapted from³⁵)

A different research team³⁶ introduced the use of Bayesian Networks to evaluate flight crew performance, enabling the integration of multidisciplinary sources of both objective and subjective data—even in cases where behavioral data may be limited. In their study, causal factors were identified based on the analysis of 484 aviation accidents attributed to human error. Lastly, Brooker³⁷ highlights the critical role of expert knowledge in the construction and validation of Bayesian Networks within safety-critical domains such as aviation.

Among the most recognized conceptual frameworks for analyzing human performance in aviation is the SHELL model³⁸, which conceptualizes pilot interaction with four components: Software (e.g., operating procedures, rules), Hardware (aircraft systems and tools), Environment (e.g., weather, airport conditions), and Liveware (interpersonal interactions, such as communication with crew members or air traffic controllers). This model emphasizes the importance of understanding human-system interaction.

In terms of performance categorization, Bandeira et al.³⁹ proposed four distinct categories that capture the complexity of influencing factors: Management and Organizational Factors, Human Factors, Environmental Factors, and Pilot Abilities. Their study also highlights the approach and landing stages as the most challenging flight phases, followed by take-off, due to their operational and cognitive demands.

³⁵ Zhou, Y., Chen, D., Xiao, J., Xiao, Y., Lu, Y., and Zhang, Y., "A Pilot Fatigue Prediction Method Based on Dynamic Bayesian Networks," *Human Factors and Ergonomics in Manufacturing & Service Industries*, Vol. 35, 2025

³⁶ Chen, W. and Huang, S., "Evaluating Flight Crew Performance by a Bayesian Network Model," *Entropy*, Vol. 20, 2018

³⁷ Brooker, P., "Experts, Bayesian Belief Networks, rare events and aviation risk estimates," *Safety Science*, 49, 2011, pp. 1142-1155

³⁸ U.S. Department of Defense, "Human Factors Analysis and Classification System (DoD HFACS)," Version 8.0., International Civil Aviation Organization (ICAO), *Safety Management Manual (SMM)*, 3rd Edition, 2012

³⁹ Bandeira, M., Correia, A. R. and Martins, M. R., "Method for measuring factors that affect the performance of pilots. *Transportes*," Vol. 25, No. 2, 2017, pp. 156-169.

Specific factor-level investigations have been conducted as well. Yazgan et al. (2017)⁴⁰ used multiple regression analysis to explore the causal links between human characteristics—including personality traits, psychomotor coordination, audio-visual memory, and numerical reasoning—and overall pilot performance. Similarly, Bai et al. (2017)⁴¹ gathered qualitative data through interviews with fighter and transport pilots, finding that physiological state, psychological condition, and environmental factors are among the most influential aspects affecting a pilot’s operational readiness.

Recent studies have also addressed the role of communication, a critical yet sometimes overlooked component. Bogdaneva et al. (2024)⁴² emphasize the importance of strengthening communication abilities—both internally within the cockpit and externally with Air Traffic Controllers (ATCo) - to identify subtle speech patterns related to situational awareness, decision-making, perception, stress, and fatigue.

Complementing this focus, the International Air Transport Association (IATA)⁴³ provides a Competency-Based Training framework that outlines nine key pilot competencies, such as Teamwork, Leadership, and Flight Path Management, encouraging a structured, performance-based approach to pilot development.

4.3 Airport Occurrences

Beyond the role of pilots during flight operations, a significant portion of risk lies within airport operations and services. The air transport system is inherently complex, consisting of key components—airlines, airports, and air traffic control services—that interact across various hierarchical levels⁴⁴. These interactions form an extensive, highly distributed network involving human operators, established procedures, and advanced technical systems. In such a system, accident risk and overall safety are heavily influenced by the way these elements work together. Ensuring an acceptable level of safety, therefore, goes beyond verifying the safe operation of each individual component; it requires managing the interdependencies among them. Given this complexity and the potentially severe consequences of accidents, risk and safety have consistently remained top priorities in the modern air transport sector.

As described in section 2.2, part of the D4S data includes safety reports. These are predominantly text-based documents, which require a Natural Language Processing (NLP) approach to handle and quantify. As such, this part considers the development of NLP-like models in the aviation and safety spaces, continuing with the proposed methodology for developing a model specifically for the respective application.

⁴⁰ Yazgan, E., Cilingir F. C., Erol, D. and Anagun, A. S., “An Analysis of the Factors Influencing Scoren Achieved during Pilot Training,” Transactions of the Japan Society for Aeronautical and Space Sciences, Vol. 60, 2017, pp. 202-211

⁴¹ Bai, S., Liu, J., Ye, J., Zhang, L., Du, J., Pan, W., Zhou, Y., Cheng, Q., Yang, L., Xiong, D., Du, P., Wang, R., Mu, H., Chen, X. and Ge, H., “Qualitative Analysis of the Factors Influencing Pilot Functional Status. Man-Machine-Environment System Engineering,” MMESE 2020, Lecture Notes in Electrical Engineering, Vol. 645, 2020, pp.119-126, Springer, Singapore.

⁴² Bogdanova, D., Giniiatullin, A. and Batth, J., S., “Existing Approaches & A Proposal to Content Analysis of Pilots’ Speech Activities,” International Russian Smart Industry Conference (SmartIndustryCon), 2024, pp. 703-708

⁴³ International Air Transport Association (IATA), “Competency Assessment and Evaluation for Pilots, Instructors and Evaluators – Guidance Material”. 2nd Edition, 2023

⁴⁴ Netjasov, F., and Janic, M., “A Review of the Research on Risk and Safety Modelling in Civil Aviation,” 3rd International Conference on Research in Air Transportation ICRAT, 2008

4.3.1 State-of-the-Art

Aviation regulations produce extensive documentation across various applications, including compliance records, safety reports, and system design specifications. While these documents contain valuable information, extracting it typically requires expert analysis due to the domain-specific language they use. The specialized terminology in aviation poses a challenge for general-purpose machine learning models, which often struggle to interpret context and meaning accurately. To address this, there is a growing need for language models tailored specifically to the aviation domain, capable of handling technical jargon and performing complex natural language processing tasks. Recent studies have shown that pre-training models on domain-specific corpora significantly improves their performance. These models have been successfully adapted for sectors like biomedicine, law, and finance, achieving superior results in a range of NLP tasks. Moreover, leveraging pre-trained models through transfer learning reduces the need for extensive computational resources and shortens training time for task-specific implementations.

Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language model based on the Transformer architecture that uses only encoder layers⁴⁵. It captures context from both directions (left and right of a word) in a sentence and is designed to be fine-tuned for a wide range of Natural Language Processing tasks, including text classification, question answering, and named entity recognition.

NLP is a field of machine learning focused on two core tasks: natural language understanding (NLU) and natural language generation (NLG)⁴⁶. In NLG, a common scenario involves a user interacting with an application that acts as the "speaker"—for example, when someone asks Siri a question and receives a natural language response. On the other hand, NLU covers a broader range of tasks, drawing from linguistic concepts such as phonology (sounds), morphology (word structure), syntax (sentence structure), semantics (literal meaning), and pragmatics (implied meaning). When processing large volumes of text—such as technical or incident reports—semantic analysis plays a central role in extracting relevant information. Regarding the aviation reports, NLP provides the following key techniques⁴⁷:

- Classification is one of the most used NLP tasks across various fields. In the aviation domain, it has frequently been applied to ASRS documents to automatically assign categories based on textual input.
- Named-Entity Recognition (NER) is a subtask of information extraction that identifies and classifies specific words or phrases into predefined entity types, such as locations ("California", "Los Angeles"), people, or organizations.

⁴⁵ Nakamura, Y., Hanaoka, S., Nomura, Y., Nakao, T., Miki, S., Watadani, T., Yoshikawa, T., Hayashi, N., and Abe, O., "Automatic detection of actionable radiology reports using bidirectional encoder representations from transformers," *BMC Medical Informatics and Decision Making*

⁴⁶ Khurana, D., Koli, A., Khatter, K., and Singh, S., "Natural Language Processing: State of The Art, Current Trends and Challenges," *Multimedia Tools and Applications*, Vol. 82, 2022, pp. 3713-3744

⁴⁷ Andrade, S., R., and Walsh, H., S., "SafeAeroBERT: Towards a Safety-Informed Aerospace-Specific Language Model AIAA," *Digital Modeling & Simulation with ML/AI and/or HPC*, 2023

- Relation Extraction (RE), another information extraction method, identifies relationships between entities or text segments. A relevant subtype is causality mining (CM), which detects causal links between entities
 - Information Retrieval (IR) helps users locate relevant documents based on keyword queries or similarity to other documents.
 - Question Answering (QA) differs from IR in that it returns a direct answer to a user query, often based on a single document rather than retrieving full documents. For instance, one might ask, “What was the aircraft type involved in the accident?” and receive a specific answer drawn from the report.
 - Summarization can be performed using two main methods: extractive and abstractive. Extractive summarization involves selecting key sentences directly from the original document, whereas abstractive summarization rewrites the content in a shorter form using rephrased language.

The success of BERT in natural language processing, combined with its open-source nature, has driven a rapid growth in models adapted to specific domains. These domain-specific BERT models are typically created by further training a general BERT model on large collections of text relevant to a particular field. Below, several examples of such models developed for different domains are summarized.

- **BioBERT⁴⁸**

Trained on biomedical texts (PubMed, PMC). Tailored for tasks like Named Entity Recognition (NER), Relation Extraction (RE), and Question Answering (QA) in the biomedical domain.

- **Clinical BERT⁴⁹**

Adapted for clinical notes and hospital records. Useful for disease prediction and summarization of complex medical records. Handles jargon, abbreviations, and grammar inconsistencies common in clinical data.

- **AraBERT⁵⁰**

Developed for Arabic language NLP. Addresses the complexity of Arabic morphology and lack of resources. Supports NER, QA, and Sentiment Analysis.

- **SciBERT⁵¹**

Trained on full-text scientific publications (Semantic Scholar). Optimized for scientific NLP tasks such as PICO extraction, NER, and Text Classification.

⁴⁸ Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C., H., and Kang, J., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, Vol. 36, 2019, pp.1234–1240

⁴⁹ Huang, K., AlTosaarand, J., and Ranganath, R., “Clinical BERT: Modeling Clinical Notes and Predicting Hospital Readmission,” arXiv:1904.05342v3 [cs.CL], 2020

⁵⁰ Antoun, W., Baly, F., and Hajj, H., “AraBERT: Transformer-based Model for Arabic Language Understanding,” arXiv:2003.00104v3 [cs.CL], 2020.

⁵¹ Beltagy, I., Lo, K., and Cohan, A., “SCIBERT: A Pretrained Language Model for Scientific Text,” arXiv:1903.10676v3 [cs.CL], 2019

- **RoBERTa**⁵²

An improved BERT variant with more data, longer sequences, and no next sentence prediction. Achieves better performance in many general NLP tasks.

- **AlphaBERT**⁵³

Uses characters (rather than words) as input units. Reduces model size while maintaining performance, particularly in hospital-based NLP systems.

- **DistilBERT**⁵⁴

A lightweight and faster version of BERT—40% smaller and 60% faster. Suitable for edge devices and mobile platforms.

- **ALBERT**⁵⁵

A "Lite" BERT using parameter reduction techniques (factorized embeddings, shared layers) to train faster and use less memory while maintaining performance.

- **BioALBERT**⁵⁶

Combines ALBERT's efficiency with biomedical domain knowledge. Optimized for biomedical NER and faster training on domain-specific data.

- **MobileBERT**⁵⁷

Designed for mobile devices with limited resources. Achieves BERT-like performance but is 5.5× faster and 4.3× smaller using a teacher-student training method.

⁵² Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692v1 [cs.CL], 2019

⁵³ Chen, Y.-P., Chen, Y.-Y., Lin, J.-J., Huang, C.-H., & Lai, F., "Modified Bidirectional Encoder Representations From Transformers Extractive Summarization Model for Hospital Information Systems Based on Character-Level Tokens (AlphaBERT): Development and Performance Evaluation," JMIR Med Inform 2020, Vol. 8, No. 4, 2020, p. 2

⁵⁴ Sanh, V., Debut, L., Chaumond, J., & Wolf, T., "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Hugging Face, EMC²: 5th Edition Co-located with NeurIPS'19, arXiv:1910.01108v4 [cs.CL], 2020.

⁵⁵ Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R., "albert: a lite bert for self-supervised learning of language representations," Computer Science, Computation and Language arXiv:1909.11942v6 [cs.CL], 2020.

⁵⁶ Nasee, U., Khushi, M., Reddy, V., Rajendran, S., Razzak, I., & Kim, J., "BioALBERT: A Simple and Effective Pre-trained Language Model for Biomedical Named Entity Recognition" Computer Science, Computation and Language, arXiv:2009.09223v1 [cs.CL], 2020

⁵⁷ Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D., "MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices," Computer Science, Computation and Language, arXiv:2004.02984v2 [cs.CL], 2020.

4.4 UQ with Machine Learning & Natural Language Processing

Machine Learning and Natural Language Processing have been incorporated into a wide range of applications across various domains, as highlighted in previous chapters. NLP heavily relies on neural networks, which are fundamentally based on probabilistic calculations and highly dependent on the quantity and quality of the data used during training. As a result, inaccurate predictions are inevitable. Therefore, it is of utmost importance to be able to evaluate the uncertainty associated with model outputs. Uncertainty quantification in a developed model assesses its reliability and its ability to make accurate predictions and provide trustworthy insights or decisions. While UQ methods to be used within DemoQUAS are thoroughly described in Deliverable D3.1⁵⁸, this part reviews basic principles close to the disciplines discussed.

Uncertainty in machine learning is typically categorized into two main types⁵⁹: aleatoric and epistemic uncertainty. It should be noted that the distinction between epistemic and aleatoric uncertainty is not always clear, as these two types of uncertainty can coexist or interact with each other. In general:

A. Aleatoric Uncertainty: Also referred to as data uncertainty, it arises from the inherent noise or randomness within the data. It is considered irreducible, meaning it cannot be eliminated through model improvements or parameter tuning. Such uncertainty may result from various factors, including noisy or imprecise observations, overlapping class distributions, labeling errors in the ground truth, or other unpredictable elements that make the data fundamentally uncertain.

B. Epistemic Uncertainty: Also called model uncertainty, occurs when there is insufficient knowledge about the model's behavior or internal design. It reflects gaps in understanding related to how the model is built or trained, including choices about architecture and parameter settings. This type of uncertainty is not fixed and can be reduced by improving the training process—mainly by exposing the model to more diverse and representative data. It often appears when the model encounters unfamiliar scenarios, such as out-of-distribution inputs, and may also result from training or design limitations.

⁵⁸ DemoQUAS Deliverable D3.1, Accessible at <https://www.demoquas.eu/news/>

⁵⁹ Hu, M., Zhang, Z., Zhao, S., Huang, M., and Wu, B., "Uncertainty in Natural Language Processing: Sources, Quantification, and Applications," arXiv:2306.04459, 2023

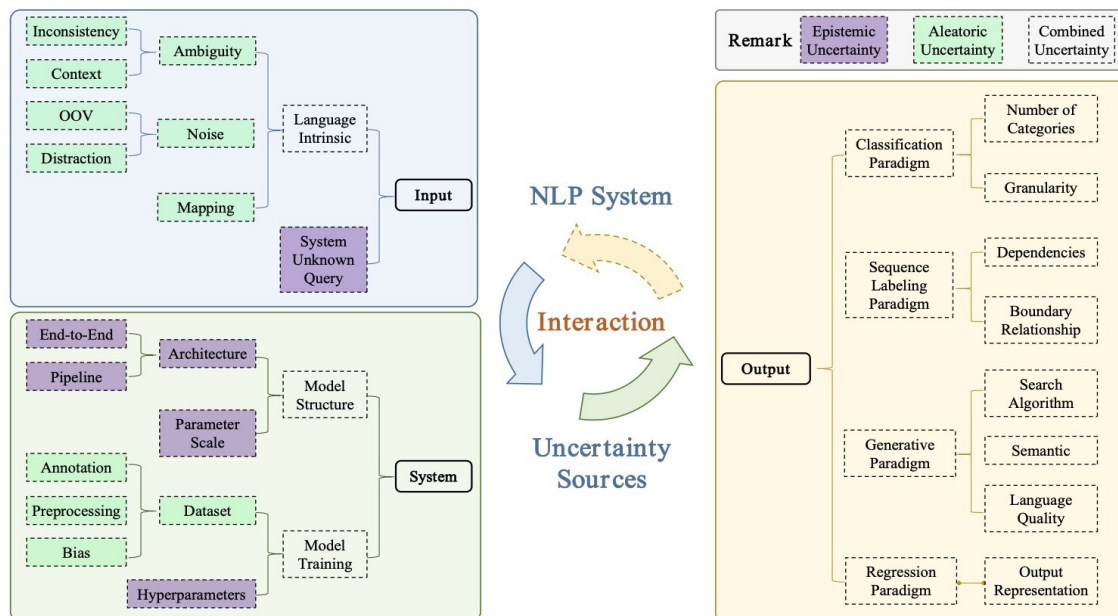


Figure 17: Illustration of sources of uncertainty⁵⁸.

Uncertainty estimation methods⁶⁰ are generally categorized into four main types: single deterministic approaches⁶¹, ensemble-based techniques⁶², Bayesian methods⁶³, and test-time augmentation strategies⁶⁴. Single deterministic methods estimate uncertainty using a single forward pass through a deterministic machine learning model. Ensemble methods, on the other hand, derive uncertainty by combining outputs from multiple different models. Bayesian methods focus on the model’s inherent randomness, such as the dropout layers in deep neural networks. Test-time augmentation techniques are model-independent and assess prediction uncertainty by applying various transformations to the input data during testing.

Uncertainty estimation in machine learning and natural language processing primarily focuses on model-driven and data-driven approaches adapted to prediction tasks. On the other hand, traditional uncertainty quantification methods, developed extensively in engineering and applied sciences, offer a broader mathematical framework for assessing uncertainty in complex systems. These classical methods complement and can enhance machine learning techniques by providing strict frameworks for analyzing uncertainty in diverse applications.

⁶⁰ Huang, Y., Song, J., Wang, Z., Zhao, S., Chen, H., Juefei-Xu, F., and Ma, L., “Look Before You Leap: An Exploratory Study of Uncertainty Analysis for Large Language Models,” arXiv:2307.10236, 2025

⁶¹ Oberdiek, P., Rottmann, M., and Gottschalk, H., “Classification uncertainty of deep neural networks based on gradient information,” in Artificial Neural Networks in Pattern Recognition: 8th IAPR TC3 Workshop, ANNPR 2018, Siena, Italy, September 19–21, 2018, Proceedings 8, Springer, 2018, pp. 113–125.

⁶² B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” Advances in neural information processing systems, vol. 30, 2017

⁶³ Barber, D., and Bishop, C. M., “Ensemble learning in bayesian neural networks,” Nato ASI Series F Computer and Systems Sciences, vol.168, pp. 215–238, 1998.

⁶⁴ A. Lyzhov, Y. Molchanova, A. Ashukha, D. Molchanov, and D. Vetrov, “Greedy policy search: A simple baseline for learnable testtime augmentation,” in Conference on Uncertainty in Artificial Intelligence. PMLR, 2020, pp. 1308–1317.

Among the techniques employed for UQ are Monte Carlo simulation⁶⁵ (MC), perturbation approaches⁶⁶, Neumann expansions⁶⁷, weighted integral methods^{68,69}, first-order second-moment methods (FOSM), first and second-order reliability methods¹¹, and spectral stochastic finite element methods⁷⁰. Among these, spectral expansion techniques such as Polynomial Chaos expansions (PCE) have become increasingly popular in recent decades due to their mathematical elegance, mean square convergence, and ability to handle random inputs with strong variation. The numerous benefits of PCE over other methods for UQ have led to applications in many disciplines, such as structural analysis, fluid dynamics, composite structures, and stability and control.

Specifically for the methods described within this report, Bayesian models themselves can be applied not only for predictions but also for uncertainty quantification in aviation applications. Due to their probabilistic nature, they have the capability to make predictions with a confidence interval, thus providing humans with the ability to accept or reject a prediction based on the estimated probability of its occurrence. For this reason, these models have been applied in areas where human factors play a crucial and decisive role in the safe performance of flights.

In a study by Weiss et al.⁷¹, the goal was to improve the reliability and speed of aircraft engine fault detection by using full-flight data and uncertainty quantification techniques. Towards this direction, the researchers applied artificial neural networks with aleatoric uncertainty estimation and extended them to include epistemic uncertainty via Out-of-Distribution Detection and finally conducted statistical evaluation and parameter optimization using grid search. They managed to achieve 2.8 times higher true positive detection rates and approximately 6.9 times faster response times compared to methods accounting only for aleatoric uncertainty. For another application, Ogunsina et al.⁷² proposed an uncertainty transfer function model (UTFM) framework for managing airline schedule disruptions proactively and reactively. The UTFM used hidden Markov models, a type of Bayesian probabilistic graphical model, to characterize and transfer uncertainty across different phases of disruption management. By leveraging historical data from a major U.S. airline, the framework enabled intelligent, quantitative decision-making to handle complex scheduling disruptions more robustly.

In a more general analysis, (Han et al.)⁷³ analyzed failure causal factors and modes of UAVs during urban delivery operations using data from over 20,000 flight hours. They developed a risk

⁶⁵ Hammersley, J. M., and Handscomb, D. C., *Monte Carlo Methods*, Methuen's Monograph on Applied Probability and Statistics, 1964

⁶⁶ Liu, W. K., Belytschko, T., and Mani, A., "Probabilistic Finite Elements for Non linear Structural Dynamics," *Computer Methods in Applied Mechanics and Engineering*, Vol. 56, No. 1, 1986, pp. 61–81

⁶⁷ Matthies, H. G., Brenner, C. E., Bucher, C. G., and Soares, C. G., "Uncertainties in Probabilistic Numerical Analysis of Structures and Solids-Stochastic Finite Elements," *Structural Safety*, Vol. 19, No. 3, 1998, pp. 283–336.

⁶⁸ Deodatis, G., and Shinozuka, M., "The Weighted Integral Method, II: Response Variability and Reliability," *Journal of Engineering Mechanics*, Vol. 117, No. 8, 1991, pp. 1865–1877.

⁶⁹ Ditlevsen, O., and Madsen, H. O., *Structural Reliability Methods*, Chichester: Wiley, 1996.

⁷⁰ Ghanem, R., and Spanos, P., *Stochastic Finite Elements: A Spectral Approach*, Springer Verlag, 1991.

⁷¹ Weiss, M., Staudacher, S., Mathes, J., Becchio, D., and Keller, C., "Uncertainty Quantification for Full-Flight Data Based Engine Fault Detection with Neural Networks," MDPI, *Machines*, 2022

⁷² Ogunsina, K., Papamichalis, M., and DeLaurentis, D., "Relational Dynamic Bayesian Network Modeling for Uncertainty Quantification and Propagation in Airline Disruption Management" arXiv:2102.05147, 2021

⁷³ Han, P., Yang, X., Zhao, Y., Guan, X., and Wang, S., "Quantitative Ground Risk Assessment for Urban Logistical Unmanned Aerial Vehicle (UAV) Based on Bayesian Network," MDPI, *Sustainability*, 2022

assessment model based on Bayesian networks to calculate the probabilities of ground-impact accidents and intermediate events under various conditions. By performing posterior probability analysis, they identified the main causes of accidents and proposed mitigation measures aimed at achieving safety levels comparable to manned aviation.

Regarding the flight trajectory, in this paper⁷⁴ a multi-fidelity deep learning approach for en-route flight trajectory prediction that explicitly quantifies prediction uncertainty using a Bayesian framework, was developed. By processing large-scale flight data, the authors trained both feedforward neural networks for short-term, accurate predictions and LSTM networks for longer-term forecasts. They then blended these models, quantified the discrepancy between their outputs at each time step to capture uncertainty, and corrected long-term predictions accordingly. Computational results demonstrated the effectiveness of this Bayesian uncertainty-aware multi-model approach in improving both prediction accuracy and safety assessment.

Bayesian methods have been applied for exploring safety incident data to efficiently detect anomalies and assess risk levels. Specifically, (Valdés et al.)⁷⁵ developed and analyzed five Bayesian statistical models, two basic and three hierarchical, to improve aviation safety analytics. The models allowed comparison of different risk sources and, importantly, provided a quantification of the uncertainty associated with those risks.

Furthermore, Bayesian models can be applied to evaluate uncertainties in external factors that affect the airport performance. A respective example is represented in this work⁷⁶, where the researchers applied Bayesian network analysis by combining airport-level traffic data from OAG with city- and country-level economic indicators to model how economic factors affect passenger and cargo volumes. Results indicate that Gross Domestic Product (GDP) and inflation impact both passenger and cargo volumes, fuel prices affect only cargo, and the Bayesian network provides probabilistic outputs for uncertainty quantification and informed aviation planning.

⁷⁴ Zhang, X., and Mahadevan, S., "Uncertainty quantification in Neural Networks by Approximate Bayesian Computation: Application to fatigue in composite materials," *Decision Support Systems*, ELSEVIER, Vol. 131, 2020

⁷⁵ Arnaldo Valdés, R. M., Gómez Comendador, V. F., Perez Sanz, L., and Rodríguez Sanz, A., "Prediction of aircraft safety incidents using Bayesian inference and hierarchical structures," *Safety Science*, ELSEVIER, Vol. 104, 2018

⁷⁶ Wang, Y., Wong, C. W. H., Cheung, T. K.-Y., and Wu, E. Y., "How influential factors affect aviation networks: A Bayesian network analysis," *Journal of Air Transport Management*, ELSEVIER, 2021

5. Conclusions

In conclusion, this report detailed a requirements definition process that included i) defining the need, ii) describing a data collection method, iii) reviewing machine learning methods to be used and iv) discussing the application of UQ methods. Specifically:

- 1) Section 2 described the safety reporting process of EASA. It detailed that due to increasing volume and types of data collected and specifically through the development of the Data4Safety (D4S) initiative, machine learning methods are prominent as big-data handling tools. As a result, this created the following requirements:
 - a) Review and recreate D4S data types to apply UQ methods.
 - b) Pin-point the machine learning methods tailored to handle each evaluated data type.
 - c) Create a list of UQ-based safety propositions and processes to be applied with the D4S framework for the handling of data and evaluation of ML methods.
- 2) Section 3 detailed the flight operations and data collection process as part of EGA operations. The data was collected using questionnaires in post-flight assessment and impact factor evaluation. The requirements for the follow-up steps include:
 - a) Re-evaluate the data structure as collected using the questionnaires.
 - b) Expand and collect additional data to fit the requirements of the developed ML models.
 - c) Coordinate and apply mitigation actions in a case study to evaluate potential improvements.
- 3) Section 4 discussed the proposed machine learning methods to be utilized within T6.3. It included reviewing applications of Bayesian Network models for proactively predicting pilot performance and proposing mitigation actions. The chapter also investigated instances of natural language processing models, stemming from the need to rapidly evaluate vast amounts of text in the form of aviation safety reports. Requirements include:
 - a) Developing a methodology for Bayesian Network framework to predict pilot performance based on the data collected by EGA, described in section 3.
 - b) Proposing safety requirements and mitigation actions based on outcomes from the pilot performance studies.
 - c) Evaluating the impact of pilot experience on the quality of data and respective results, given data reflects the performance of student pilots.
 - d) Developing data collection and methodology of a BERT-based NLP framework to process text-based report data.

According to the presented requirements and proposed methods, T6.3 will continue towards its multi-faceted goals in enabling UQ methods towards enhancing aviation safety through the respective milestones and deliverables set post-M16.