

EXPLAINABILITY AT ALL LEVELS OF SYSTEM DEVELOPMENT: INSIGHTS FROM THE DESIGN OF INTELLIGENT SYSTEMS IN AVIATION

Jakob Würfel[1,2], Boris Djartov[1], Max Friedrich[1], Sabrina Schog[1] & Anne Papenfuss[1]

[1]German Aerospace Center (DLR) - Institute of Flight Guidance, Lilienthalplatz 7, 38108 Braunschweig, Germany
[2]RWTH Aachen University

Abstract

In anticipation of the future integration of AI into aviation systems, European Union Aviation Safety Agency (EASA) has published the AI Roadmap 2.0 which, among other objectives, calls for the explainability of AI in operational contexts. The goal is to enable human end users, such as pilots and air traffic controllers, to understand the cause behind the results of AI-based systems and apply them in operational contexts. This paper presents a multi-level approach to achieving explainability for the Intelligent Pilot Advisory System (IPAS), an AI-based cockpit assistant developed at the German Aerospace Center (DLR). Explainability is addressed both at the AI-core level, where an interpretable multi-criteria decision-making model extracts feature-attribution information, and at the human-AI interaction level, where a human-centered approach derives interpretability features to support cockpit interface design. A Wizard-of-Oz study with twelve airline pilots evaluated four display variants of the human-AI interaction system, each implementing different combinations of interpretability features. The results show that while more interpretability features increase perceived explainability, compact displays with fewer features are rated more useful for decision support. The paper concludes that operational explainability must be considered at all levels of system design, from AI model selection to interface development, and that parallel development of both the AI system and the human-AI interaction, along with early and recurrent involvement of end users, is essential to address explainability and task-specific needs. Additionally, the transferability of these findings to a second use case, the military Manned-Unmanned Teaming (MUM-T) mission manager is discussed.

Keywords: Operational Explainability, AI-based Decision Support, Human–AI Interaction, Multi-Criteria Decision Support

1. Introduction

Integrating AI into aviation systems, particularly for safety-critical applications such as cockpit decision support, requires operational explainability to ensure calibrated trust, transparency, and effective human-AI collaboration. This need is explicitly addressed in the AI Roadmap 2.0 published by the European Union Aviation Safety Agency (EASA), which defines operational explainability as the provision of understandable, reliable, and context-relevant explanations of AI outputs to end users during real-time operations [1, 2]. While existing eXplainable AI (XAI) methods provide technical insights into model behavior, they often fail to meet the practical needs of end users, such as pilots who require easily comprehensible, task-specific, and operationally meaningful explanations under time constraints and uncertainty [3, 4, 5, 6].

There are many potential applications for AI-based systems in aviation. This paper presents the Intelligent Pilot Advisory System (IPAS) as an use case for implementing AI in the airline cockpit environment. Research was conducted on operational explainability for the IPAS, which is described and discussed in detail in this paper. In conclusions, the transferability of the IPAS findings to a

second use case, the Manned-Unmanned Teaming (MUM-T) mission manager, a display for future air combat operations, is discussed.
The IPAS consists out of two subsystems: the AI Core Module (AICOM), which serves as the AI core implementing the multi-criteria decision logic, and the AI-Crew Interaction System (AICIS), which acts as the interface between pilots and the AI. The AICIS presents AI outputs in an explainable and operationally useful manner, with the aim of ensuring that operators can understand the AI's reasoning, validate its recommendations, maintain situation awareness, and ultimately support human-AI performance [7, 6]. Nevertheless, concrete guidelines and processes for integrating explainability into operational systems for these domains are still lacking.
This paper contributes to closing this gap by:

1. Presentation of the fundamentals of operational explainability of AI as well as interpretable and explainable AI methods.
2. Presentation of two aviation use cases for human-AI interaction: MUM-T and IPAS.
3. Presenting a multi-level approach to achieve operational explainability for the IPAS, combining the AI perspective (AICOM) and the human-centered interface development perspective (AICIS);
4. Evaluating four explainable display variants of the AICIS in a Wizard-of-Oz study with twelve airline pilots, focusing on explainability, operational suitability and usefulness;
5. Deriving design implications for balancing explainability with operational usability in Artificial Intelligence (AI)-based aviation systems; and
6. Discussing lessons learned from IPAS development regarding operational explainability and their potential transferability to the MUM-T use case.

The objective of the Wizard-of-Oz study presented in this paper (Section 5), is to explore the implementation of operational explainability for the AICIS. To this end, the study addresses the following three research questions:

1. How do different Explainability-Information and their visualisations influence pilots' perceived explainability of AI results?
2. How do different Explainability-Information and visualisations affect perceived usefulness and overall operational suitability for decision-support?
3. Which design implications can be derived for AI-based cockpit interfaces to provide operational explainability, decision-making usefulness, and operational suitability?

2. Theoretical Background

2.1 AI Operational Explainability

The evolution of cockpit automation, from early airline cockpits with dozens of analog gauges and manual steering to modern glass cockpits with complex cockpit systems, has shifted the human role from an active, manual operator to that of a system manager [8]. While automation enhances efficiency and safety, it also introduces human-factor risks such as loss of situation awareness (out-of-the-loop effects), skill degradation, and automation bias [9, 10, 11, 12]. AI-based automation intensifies these challenges because many AI models operate as *black boxes*, because their inner working is opaque to human operators [13]. To reveal the black box, the field of explainable AI (XAI) research has developed.
The EU Commission specifies that an AI system is explainable when its functionality and operations can be explained in non-technical terms to a person not skilled in the art [14]. A clear distinction must be made regarding the target audience for AI explainability. While the research on XAI focuses on developing mathematical models to open the black box, it often prioritizes technical aspects over user-friendliness and practical applicability for end users [3, 15]. End users are primarily interested in the "Why?", the reasoning behind the AI's decisions, rather than the "How?", the technical details [16, 5]. However, most XAI models are designed for technical users, such as AI researchers, data scientists, and developers, who use them for debugging and validation. End users, by contrast, are largely ignored in XAI research [4, 17].

Consequently, research in this area is needed, as explainability is a key aspect of Human-AI teaming and human-centered AI, enabling effective, efficient, and safe collaboration between humans and AI [7, 13]. For instance, explainability supports situation awareness [7], calibrated trust (preventing both under- and over-trust in AI systems) [18, 19, 20, 21], enhances user satisfaction [13], and ultimately improves Human-AI-Team performance [7, 22]. Providing too much explanatory detail can overwhelm users, negatively impacting performance and usability [18, 19]. Striking the right balance between explainability and cognitive load is crucial to ensure that AI systems remain both explainable and operationally applicable for users.
The EASA AI Roadmap 2.0 introduced the term *Operational AI Explainability*, emphasizing it as a fundamental requirement for all AI automation levels, from assistant systems (Level 1A) to shared decision-making authority (Level 2B) and future advanced automated systems (Level 3B) [1, 2]. This concept specifically refers to the provision of understandable, reliable, timely and context-relevant explanations about an AI system's results to end users during operation [1].

2.2 Interpretable and explainable AI Methods

The field of explainable and interpretable AI comprises a wide variety of methods, each differing in scope, underlying assumptions, and how they relate to the predictive model in question. A fundamental distinction exists between explanation techniques applied post hoc to opaque models and models that are inherently interpretable by design. While complex architectures such as neural networks and ensemble methods generally require additional explanation techniques to make their behaviour accessible, simpler model classes like linear models or shallow decision trees allow their decision logic to be inspected directly, affording a degree of intrinsic transparency. Several dimensions are commonly used to categorise this diverse methodological landscape. One such dimension concerns the target of explanation: global approaches seek to characterise the overall behaviour of a model, while local approaches focus on explaining individual predictions or specific decision instances. Another key distinction lies between model-agnostic and model-specific methods. The former treat the model as a black box and derive explanations purely from input-output observations, offering broad applicability across different model types. The latter leverage the internal structure of specific models to produce more targeted, though less generalisable, explanations. Together, these dimensions provide a useful framework for navigating the growing range of XAI techniques available.
Among the most widely adopted explanation techniques in practice are Permutation Feature Importance (PFI) [23, 24], Integrated Gradients [25], and Shapley-value-based methods as implemented in SHAP (SHapley Additive exPlanations) [26].

2.2.1 Permutation Feature Importance (PFI)

PFI is a model-agnostic, global method that estimates the degree to which a model relies on a given input feature by measuring the degradation in predictive performance when that feature's values are randomly permuted, thereby breaking its statistical association with the target [23, 24]. The importance of feature j is computed as either the difference or ratio between the permuted error $e_{\text{perm},j}$ and the baseline error e_{orig}:

$$FI_j^{(\Delta)} = e_{\text{perm},j} - e_{\text{orig}} \qquad \text{or} \qquad FI_j^{(\times)} = \frac{e_{\text{perm},j}}{e_{\text{orig}}} \tag{1}$$

While simple and broadly applicable, PFI does not provide information about the direction of a feature's effect on predictions, and can produce misleading results in the presence of correlated features, as permuting a feature breaks the correlation structure and can generate unrealistic input combinations [27].

2.2.2 Integrated Gradients

Integrated Gradients is a gradient-based attribution method applicable to differentiable models such as neural networks. It assigns a contribution to each input feature by aggregating gradient information along a straight-line path from a reference baseline $\mathbf{x}'$ to the input $\mathbf{x}$, formally defined as:

$$\mathrm{IG}_i(\mathbf{x}) := (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial f(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\partial x_i} d\alpha \tag{2}$$

The method satisfies desirable axioms including Sensitivity and Implementation Invariance, and guarantees that attributions sum exactly to the difference between the prediction and the baseline output [25]. However, the choice of baseline can substantially affect the resulting attributions, and the assumption of a straight-line path through the input space can generate implausible intermediate points, particularly for tabular data with complex feature relationships [27].

2.2.3 Shapley Values and SHAP

SHAP [26] provides additive feature attributions grounded in cooperative game theory, and is available in both model-agnostic and model-specific variants. Each input feature is treated as a player in a cooperative game, where the payout is the model's prediction. The Shapley value ϕ_i for feature i is defined as its weighted average marginal contribution across all possible feature subsets:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \tag{3}$$

Intuitively, this asks: on average, how much does adding feature i to any subset of features change the model output? A positive ϕ_i indicates the feature pushed the prediction above the baseline; a negative value indicates it pulled the prediction below it.

2.2.4 Challenges and summary of Interpretable and Explainable Machine Learning

The open challenges surrounding Interpretable and Explainable Machine Learning have been the subject of considerable research attention, with calls for a more rigorous scientific foundation for interpretability [28] and detailed accounts of the practical and technical difficulties that arise in applied settings [29]. Central among these are questions concerning the reliability and faithfulness of explanations, the degree to which they accurately reflect underlying model behaviour, and their practical utility for end users. The most prominent and recurring challenges can be summarised as follows:

1. Statistical uncertainty: The uncertainty of explanations is rarely quantified by XAI methods, which can lead practitioners to draw overconfident conclusions, particularly when working with noisy or sparse data.
2. Causal interpretation: As most XAI techniques capture correlational rather than causal relationships, their use in contexts requiring actionable or causal insight remains limited.
3. Feature dependence: Methods that assume feature independence, such as permutation-based approaches, can produce misleading attributions when input features are correlated.
4. Lack of definition: The absence of a precise and widely accepted definition of interpretability makes systematic evaluation and meaningful comparison across methods difficult.
5. Evaluation metrics: No agreed-upon ground truth exists for assessing explanation quality, leaving evaluation largely dependent on qualitative or subjective human judgement.
6. Local vs. global explanations: Reconciling instance-level local explanations with coherent global accounts of model behaviour remains an open methodological challenge.
7. Scalability and complexity: The computational demands of many XAI methods limit their applicability in high-dimensional settings or scenarios requiring real-time inference.
8. Human factors: What constitutes an understandable explanation is inherently user-dependent, varying considerably with domain expertise, background knowledge, and cognitive capacity.

Notwithstanding these challenges, XAI plays an important role in the responsible deployment of AI systems. Table 1 provides an overview about the discussed explanation techniques.

Table 1 – Comparison of XAI methods

Aspect	SHAP	PFI	IG
Explanation Type	Additive feature attributions based on Shapley values	Performance change after permuting a feature's values	Gradient-based attribution from a baseline to the input
Output Format	Feature contributions summing to the prediction	Importance score before vs. after permutation	Attribution vector per feature
Advantages	Theoretically grounded; fair; local and global insight	Simple, fast; no retraining; clear global rankings	Computationally efficient; faithful for deep models
Limitations	Computationally intensive	Sensitive to feature correlation; no directionality	Requires differentiable models; sensitive to baseline
Use Cases	Feature attribution; auditing decisions	Global importance ranking; model monitoring	Interpreting deep neural networks

3. Use Cases for AI Applications in Aviation

This section introduces the two use cases, the MUM-T and the IPAS. Sections 4 and 5 present the multi-level research activities regarding the implementation of operational AI explainability for the IPAS, while the transferability of the findings to the MUM-T is discussed at the end of the paper.

3.1 AI-Supported MUM-T Mission Management in Future Air Combat

MUM-T will be a central component in future air combat. In MUM-T, manned assets such as fighter aircraft operate together with unmanned aircraft systems (UAS) in coordinated missions [30, 31, 32]. Even as UAS become increasingly autonomous, the human pilot will remain an important element, especially during unexpected events that fall outside pre-planned use cases ([32].

Building on current technological trends, next-generation MUM-T systems will not only connect autonomous UAS with human pilots but they will embed AI-enabled planning and decision-support tools that work side-by-side with the pilot [30, 33, 34]. These intelligent aids are designed to support the entire mission lifecycle from pre-mission planning through real-time monitoring to real-time mission replanning when the circumstances demand [35, 36].

However, integrating UAS with autonomous capabilities in air combat missions raises assurance but also trust concerns. The human pilot needs to trust that the autonomy will respect doctrinal aspects, rules of engagement (ROE), ethical principles, and legal constraints, especially when human oversight cannot be continuously guaranteed. From an ethical-legal perspective, autonomous weapon systems can only comply with ROE and ethical norms if their autonomy is bounded and their decision-making processes remain interpretable and abortable by the human. However, the complexity of a system comprising multiple autonomous UAS operating in a contested environment will likely make efficient human oversight difficult, because the human pilot can no longer process the flood of information in time to make time and safety critical decisions.

In some missions the human may even have to "send" a swarm of UAS into a contested area where communications are jammed or spoofed, eliminating any real-time oversight. The swarm would then operate completely autonomously and may identify and engage targets and perform battle-damage assessment (BDA) on its own. Only after the communication link is restored would the system transmit a post-mission report. Since the human only sees the mission results after it has already ended, the system should provide a short, clear summary of the swarm's actions covering the most important observations, which targets were chosen, how they were engaged, and if any ROE were breached.

Methods of explainable and interpretable AI paired with a user-centered human-machine interface (HMI) can make the reasoning behind autonomous decisions and actions transparent, supporting human situation awareness and meaningful human control (MHC). Furthermore, [37] emphasize the

importance of adopting a comprehensive Human-Systems-Integration approach in the design of AI-supported mission management systems for future MUM-T missions, as successfully integrating AI into existing mission management processes and systems requires careful consideration of human operators' needs in order to fully realize the benefits of MUM-T [38].
An example of an HMI concept that provides the human pilot with different plan alternatives that were created by an intelligent mission management system incl. an icon and color coding based visualization of key mission parameters of each plan is presented in [39]. Figure 1 depicts the HMI concept (for further details, please refer to [37, 39, 40]). The HMI allows human pilots to compare the proposed plan alternatives and choose the plan that matches best the current mission objectives. The human would thus remain strategically involved in the mission by planning and launching the mission and analyzing the collected data after the mission, while the autonomous systems handle the tactical execution. Consequently, explainability serves as a bridge to enable human oversight even for complex autonomous systems operated in contested environments.
While the HMI offers a simplified, high-level overview of the mission parameters associated with each proposed plan, it does not yet present the underlying reasoning that led the intelligent mission management system to assign those values. Consequently, human pilots cannot assess why a particular mission plan received the respective parameter values. To close this gap, XAI-methods, like those presented in Table 1 could be implemented and visualized on a separate, on-demand screen. When a pilot selects a plan and requests further insight, the system would generate a concise, visual explanation of the factors and internal decision processes that contributed to each plan proposal and mission parameter. This approach would preserve the simplicity of the primary, high level view of the HMI while enabling access to more detailed explanations about the proposed mission plans.



Figure 1 – Illustration if the HMI concept for icon-based visualization of key parameters of proposed mission plans from [37, 39].

3.2 The Intelligent Pilot Advisory System

The IPAS is an AI-based cockpit assistance system developed to serve as a research platform to investigate and demonstrate the integration and operational use of AI in modern flight decks, with a primary objective of supporting flight crews in time-critical decision-making, particularly during abnormal and emergency situations such as unplanned diversions. According to EUROCONTROL, in 2024 approximately 30000 flights (0.32% of all European flights in 2024) landed at a different airport than originally planned [41]. Such diversions can be caused by a variety of factors, including adverse weather conditions, technical issues, medical emergencies, air traffic constraints, or limitations related to the destination airport [42]. In such situations, flight crews must make time-critical decisions by rapidly gathering and evaluating a wide range of information. They also carefully weigh technical constraints, operational limitations, and safety considerations while managing the aircraft. To ensure structured and effective decision-making, crews often rely on established models such as FORDEC

(Facts-Options-Risks-Decision-Execution-Controll) [43].
To assist crews in these demanding situations, the IPAS includes a function that supports the flight crew in finding an suitable alternate airport. This function generates, evaluates, and presents potential diversion options by gathering and evaluating relevant decision factors, considering operational constraints, the current state of the aircraft and environmental conditions.
Figure 2 illustrates a simplified swimlane diagram of the IPAS, showing the functional allocation between the subsystems AICOM (AI core) and the AICIS (crew interface). This paper focuses on the area highlighted by the dashed red frame in the diagram, the Option Assessment Support functionality.

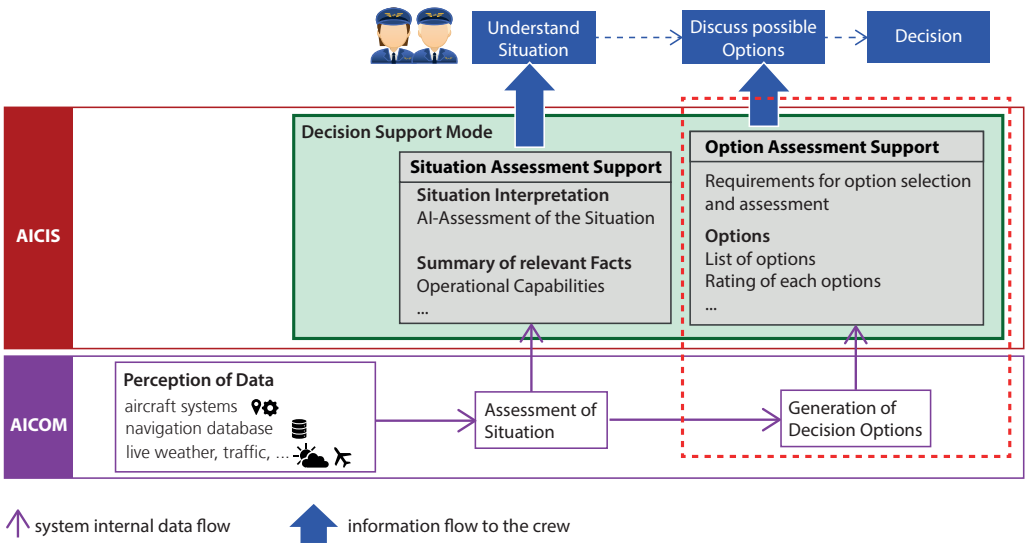


Figure 2 – Simplified swimlane diagram of Decision Support Mode. Adapted from [44]

The Decision Support Mode operates in two steps:

1. **Data Collection & Situation Assessment Support**
   The AICOM continuously collects data from aircraft systems, navigation databases, real-time weather updates, and air traffic information. This data is used to perform a comprehensive situation assessment, which is then presented to the pilots via the AICIS.

2. **Option Assessment Support**
   The AICOM generates alternate airport options based on the collected data. The AICIS presents these options, along with supporting information, to the flight crew. The pilots then discuss the options and make a final decision.

Previous research on the IPAS has established the foundations for its conceptual design [45, 46], technical architecture [47, 48], and operational explainability [5].

4. Explainability Considerations in the Development of the IPAS

A central research goal of the IPAS development is to investigate how operational AI explainability can be achieved for AI-based cockpit assistance systems. To explore this, two complementary approaches were followed in parallel. A top-down approach was followed to derive operational explainability requirements tailored to the needs, expertise and task demands of pilots. These high-level requirements guided the design of the interface of the AICIS, so that the AI results are presented explainable to the pilots. A bottom-up approach focused on the AICOM, exploring interpretable methods that fulfill the AI's functional requirements while making the system as explainable to humans. A key challenge is to expose the relevant Explainability-Information (e.g., feature importance) so that it can be fed into the user interface. The next step is to combine the pilot-centred explainability requirements with the constraints and possibilities of the underlying AI model. In this way it is possible

to demonstrate how operational explainability can be achieved from the AI-model level to the user interface. In the following both approaches are described in more detail.

4.1 Bottom-up - AI Level perspective: Explainability in the Design of AICOM

The AICOM constitutes the model-facing core of the IPAS architecture. It houses the trained AI models that produce rankings of candidate alternate airports. Given the safety-critical nature of the system and the need to pass meaningful explanation outputs to the AICIS for presentation to the flight crew, interpretability and explainability were given high priority throughout the development of the AICOM, spanning the choice of training methodology and model classes through to the selection of explanation techniques. These priorities are reflected in the design of the underlying methodology itself.

The AICOM is built around the ExACT-MCDM methodology [49], which formulates the alternate airport selection problem as a supervised ranking task. This is more suitable than a classification approach, as flight crews typically reason about multiple alternatives in parallel rather than selecting a single option. Each option is described by a set of operational features, including runway length, crosswind and tailwind conditions, fuel required, runway surface status, and distances, alongside context features characterising the current aircraft state, such as airline type, altitude, and initial fuel.

A further key aspect of the methodology is its cyclical nature and use of synthetic data generation [49]. Rather than relying on scarce real-world diversion data, training data is generated through simulations informed by domain expertise. This addresses the data scarcity problem common in safety-critical aviation applications, while also contributing to transparency: since the data is generated by the developers themselves, there is a deeper understanding of its structure and the assumptions embedded in it, making the overall development process more traceable and auditable.

4.1.1 Hierarchical Constraint-Aware Evaluation Function

The ranking targets are derived from a hierarchical, constraint-aware evaluation function that encodes domain expert knowledge through structured scoring and penalty terms [49]. This function is a central explainability mechanism: because it is explicit and traceable, stakeholders can compare model-generated rankings against the scoring logic and identify precisely where the learned model deviates from the intended decision process.

Feature grouping and transformation

Let $A = \{a_1, a_2, \ldots, a_n\}$ be the set of alternatives, and let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ denote the feature vector of option a_i. The features are partitioned into two groups:

- C: *constraint-critical* features, encoding safety and feasibility requirements (e.g. visibility, runway margin, fuel reserve),
- P: *preference-based* features, encoding operational efficiency criteria (e.g. distance to destination, distance to airline hub).

For each constraint-critical feature $j \in C$, a transformation function ϕ_j quantifies the degree of violation. For a lower-bound constraint with soft threshold s_j and hard threshold h_j, this takes the form:

$$\phi_j(x_{ij}) = \begin{cases} 0, & \text{if } x_{ij} \geq s_j \quad \text{(no violation)}, \\ s_j - x_{ij}, & \text{if } h_j \leq x_{ij} < s_j \quad \text{(soft violation)}, \\ s_j - x_{ij}, & \text{if } x_{ij} < h_j \quad \text{(hard violation, } \lambda \leftarrow \lambda + 1). \end{cases} \tag{4}$$

Upper-bound constraints (e.g. crosswind) are handled symmetrically, with the violation measured as $x_{ij} - s_j$ once the soft threshold is exceeded. Each hard-constraint breach increments a counter $\lambda_i \geq 0$ for option a_i, acting as a large, discrete penalty that forces infeasible options to the bottom of the ranking regardless of their preference profile. For preference-based features $j \in P$, an analogous transformation $\psi_j(x_{ij})$ may be defined, typically encoding a monotone preference direction (e.g. minimising distance).

Aggregation and scoring.

After applying the transformation functions, the resulting values are normalised feature-wise across all options using min–max normalisation. The aggregated group scores are:

$$S_i^{\text{constr}} = \sum_{j \in C} \phi_j(x_{ij}), \qquad S_i^{\text{pref}} = \sum_{j \in P} \psi_j(x_{ij}). \tag{5}$$

A second normalisation is then applied across options at the group level:

$$\text{Norm}(z_i) = \frac{z_i - \min_k z_k}{\max_k z_k - \min_k z_k}. \tag{6}$$

The final score for option a_i, which serves as the ranking target passed to the AI model, combines the normalised constraint and preference components with the hard-violation counter:

$$\text{Score}_i = \underbrace{\text{Norm}\left(S_i^{\text{constr}}\right)}_{\text{constraint group (penalty-dominant)}} + \lambda_i + \underbrace{\text{Norm}\left(S_i^{\text{pref}}\right)}_{\text{preference group}}. \tag{7}$$

Lower scores are better. The structure guarantees that a single hard-constraint violation (adding at least 1 to λ_i) outweighs any improvement in the normalised preference component, which is bounded in $[0, 1]$. Feasible alternatives are therefore always ranked above infeasible ones, while the preference terms differentiate among the safe options.

Explainability, robustness, and uncertainty

The hierarchical scoring function underlying the ranking models has direct implications for the explainability of AICOM outputs. Because the models are trained on targets derived from this explicit function, a well-trained model should assign feature importances that broadly mirror its structure, with constraint-critical features dominating and preference features playing a secondary role. This provides a built-in sanity check: if an attribution method assigns negligible importance to a feature that carries hard-constraint penalties in the scoring function, this is a signal that the explanation should be interrogated before being surfaced to operational users. The function's transparency also supports stakeholder communication, as developers and domain experts can inspectand modify the transformation functions without requiring access to internal model weights, separating what the model learned from what it was trained to do.

The trained models themselves act as fast surrogates that approximate the scoring logic in real time [49]. Robustness and uncertainty are incorporated directly into the training data through scenario simulation: since factors such as wind, weather, and fuel estimates cannot be known exactly at decision time, controlled perturbations are introduced into each scenario and the evaluation function is applied across the resulting variants. The models are therefore not trained on a single deterministic label per scenario, but on a robustness-aware estimate that reflects realistic input variability. By handling this computationally expensive process offline during training, the surrogate implicitly encodes both the domain expertise of the scoring function and the uncertainty of the simulation, and can deliver a reliable ranking almost instantaneously when it is needed most.

4.1.2 Shapley Values and Model-Internal Explainability

Multiple model classes were trained and evaluated within the ExACT-MCDM framework, spanning the interpretability–performance spectrum [49]. A weighted-sum model was included as a fully transparent baseline, chosen for its direct alignment with classical MCDM approaches and its fully inspectable decision logic. Learning classifier system (LCS) models based on the XCSF framework were selected for their rule-based, interpretable representations. Two variants were considered: a simple constrained variant with a small classifier population that remained largely inspectable, and a larger performance-oriented variant with a significantly bigger population in which interpretability was traded for greater representational capacity. Similarly, XGBoost was considered in two variants: a simple configuration with a limited number of shallow trees that retained partial interpretability, and a full configuration with many deeper trees in which predictive capacity was prioritised at the expense of human comprehensibility. Finally, a shallow feedforward neural network was included as

the least interpretable of the considered models, for which post-hoc explanation techniques were required entirely. This range of model classes and variants allowed the explainability analysis to span from fully transparent models to those requiring external tools, providing insight into what explanation information could be reliably extracted from each and passed to the AICIS.
To extract feature-level insights from the trained models and assess what explanation information can realistically be passed to the AICIS, three complementary explanation techniques were applied, all of which are described in detail in Section 2.: PFI, SHAP, and Integrated Gradients [49]. The results of this analysis are summarised in Figure 3.

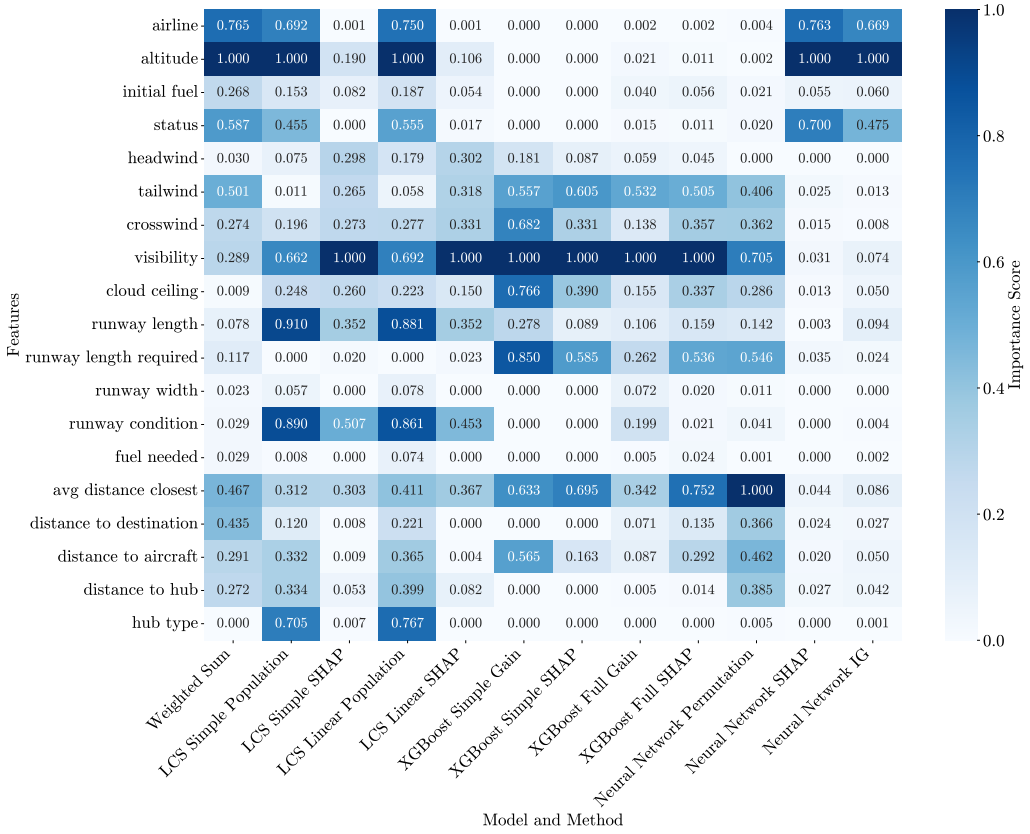


Figure 3 – Min-max normalised feature importance across models and explanation methods, taken from [49].

Across the interpretable model classes, the methods produced broadly consistent results. The globally dominant decision factors, namely average distance to the closest airports, visibility, crosswind, and tailwind, appeared reliably across methods and aligned with the safety-critical structure of the evaluation function. For the weighted-sum model and the simple LCS variant, SHAP-based feature contributions closely matched the explicit weight vectors, providing a natural sanity check and confirming that the explanation outputs would be meaningful if passed to the AICIS.
The picture changed markedly for the neural network. In the ranking formulation, context features are repeated before each airport's feature block in the input vector. SHAP and Integrated Gradients both assigned overwhelming importance to these context features and near-zero attribution to the airport-specific decision factors, a pattern that directly conflicts with the model's strong ranking performance and with the structure of the evaluation function [49]. This is consistent with documented failure modes of gradient-based attribution methods when the input structure introduces systematic repetitions [50]. PFI avoided this bias and produced a more balanced and operationally plausible importance profile, suggesting the problem lay with the explanation method's interaction with the input structure rather than with the model itself.
This finding has direct implications for what can realistically be extracted from the AICOM and passed to the AICIS. The factor ratings and feature contributions that the AICIS display variants visualise must be sourced from explanation methods that produce reliable outputs for the specific model and input

structure in use. For the neural network, SHAP and Integrated Gradients cannot provide this reliably in the current formulation, whereas a permutation-based approach can. The choice of explanation method therefore cannot be decoupled from the model architecture: selecting the highest-performing model does not automatically make it the best source of explanation information for the crew interface, and this must be validated before any outputs are surfaced to pilots.

4.2 Top-down - User-Interface Level perspective: Operational AI Explainability in the Design of the AICIS

As Degen [51] argues, explainability cannot be implemented without a understanding of the needs, goals, and tasks of the end users. Explanations must be user-specific, tailored to the characteristics, knowledge, and operational context of the target user group [3]. Furthermore, the required type and granularity of explanations depend on the users' specific tasks and prior knowledge [4]. To address this challenge, a user-centered and iterative Human Systems Exploration (HSE) development approach [52] is adopted. This methodology enables the identification of pilots' needs, goals, and task-specific requirements for operational explainability while supporting the parallel development of the AICIS. By integrating pilots early and frequently into the development process through interviews, user tests, and small-scale studies, a concept for operational explainability is explored, defined and getting more detailed step by step. At the same time the AICIS is continuously adapted to implement and demonstrate operationally suitable explainability based on the defined concept [53].
To find a starting point, the concept of so-called *interpretability features* was developed. The idea behind the concept of operational AI explainability via interpretability features and what this concept is based on is described in detail in Würfel et al. [5].

4.2.1 Interpretability Features to support Operational AI Explainability

First of all, the term *interpretability* is used in the literature with a broad spectrum of meanings. To identify an approach that ensures operational explainability, the definition proposed by Stevens and De Smedt [54] was adopted:

> "Interpretability is the ability to provide an explanation that consists solely of single chunks of information, preferably in a human-understandable fashion."

Lipton [55] refers to this as *post-hoc interpretability*: the extraction of individual pieces of data from a trained model so that its inner workings become transparent without exposing the full black-box. Hoffman et al. [56] argue that explanatory systems are most effective when they empower users to self-explain rather than presenting a whole statistical representation of the algorithm's output. These ideas form the basis of the *interpretability cues* introduced for the AICIS.
The approach aims to bridge the gap between the algorithmic and operational levels, making the reasons behind AI's results explainable to pilots (Why a certain result is shown) while remaining usable in decision situations with high operational workloads.
To prevent pilots from being overloaded in workload-intense cockpit situations, the approach avoids presenting the full set of model parameters, statistical data, or a detailed decision tree. Instead, the explanation is narrowed to a small number of relevant information chunks that explain the reason behind an AI-based decision, keeping the AI-based System operationally feasible while still providing the explanability necessary for the pilot to understand the AI recommendation. This trade-off between explanatory detail and operational usefulness is the one highlighted by Stevens and De Smedt [54].
The term *Interpretability Cues* was later renamed *Interpretability Features* to reflect that they describe the structure, content, and presentation as system features [53]. The interpretability features, identified through an exploratory, user-centred approach in Würfel et al. [5], are:

1. The system should display the decision factors typically used by the flight crew to make decisions, adapted to the current information needs of the decision process phase.
2. The system should provide information about the assessment that the flight crew can compare with their own perception, experience, and assessment.

3. The system should show whether displayed information is the result of an AI process and whether displayed information (AI assessment or AI-generated information) is based on real-time data or on an AI forecast of the situation in the future.
4. The system must indicate which factors and factor combinations are included in the assessment of the AI and how they affect a system output.

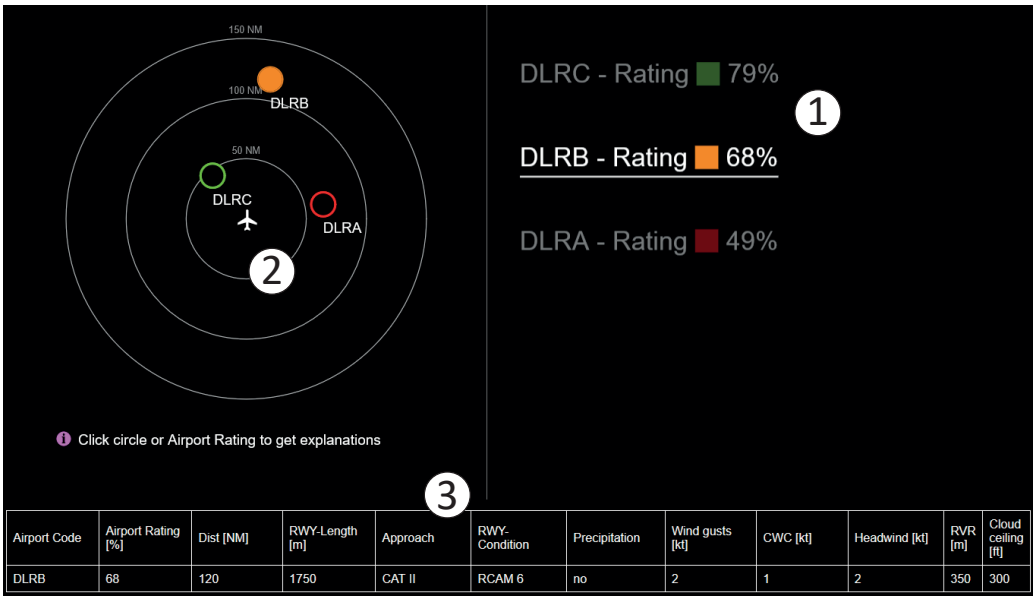
Depending on the use case and context, the development of human-AI systems in the aircraft cockpit can be guided by these high-level requirements to support operational explainability.

4.2.2 Mockups: Explainable Display Variants

To evaluate the proposed concept for operational explainability, the identified Interpretability Features were used to design four display variants as mockups. The displays fulfill the previously presented functionality of the Option Assessment Support by presenting potential alternate airports along with their AI-based assessment. Based on the Interpretability Features, explanatory information and visual elements were implemented to explain the reasoning behind the AI's recommendations and assessments to the user in an operational context:

- Decision-factor set: the list of decision factors that influence the airport rating (refers to Interpretability Feature 1).
- Factor rating indication: a color-coded (green/amber/red) indication of whether each factor contributes positively or negatively to the overall rating (Features 2 & 4).
- Feature importance indication: a quantitative magnitude that shows how strongly a factor pushes the rating up or down (Features 2 & 4).
- Threshold visualisation: the distance of a factor's value to its rating thresholds (Feature 2).

Interpretability Feature 3 was not implemented because the study focuses on the explainability of the airport rating itself rather than on information about the underlying data and data sources.
All four variants consist of the same radar diagram ①, which visualises the distance and heading of each airport relative to the aircraft, as well as a hierarchical list ② that orders the airports by their rating. The way in which the ratings are explained to pilots differs between the four display variants due to the use of different explanatory information and visual elements.



Airport Code	Airport Rating [%]	Dist [NM]	RWY-Length [m]	Approach	RWY-Condition	Precipitation	Wind gusts [kt]	CWC [kt]	Headwind [kt]	RVR [m]	Cloud ceiling [ft]
DLRB	68	120	1750	CAT II	RCAM 6	no	2	1	2	350	300

Figure 4 – Explanation Visualisation 1: rtb - Baseline

The baseline variant rtb shows only the raw decision-factor values in a plain table, with no additional explanatory information or visual element ③. As shown in Figure 5, the color-enhanced baseline variant rtbc adds the factor-rating information by color-coding the factors in the table according to their rating.

Airport Code	Airport Rating [%]	Dist [NM]	RWY-Length [m]	Approach	RWY-Condition	Precipitation	Wind gusts [kt]	CWC [kt]	Headwind [kt]	RVR [m]	Cloud ceiling [ft]
DLRB	68	120	1750	CAT II	RCAM 6	no	2	1	2	350	300

Figure 5 – Explanation Visualisation 2: rtbc - Table with colors

The two variants fp and fi further expose feature importance, each using a different visual style. Variant fp, shown in Figure 6, is based on a SHAP-inspired force plot: each decision factor is represented by an arrow ④ whose length indicates the magnitude of the factor's contribution, while the arrow's color and direction indicates whether the contribution is positive or negative.

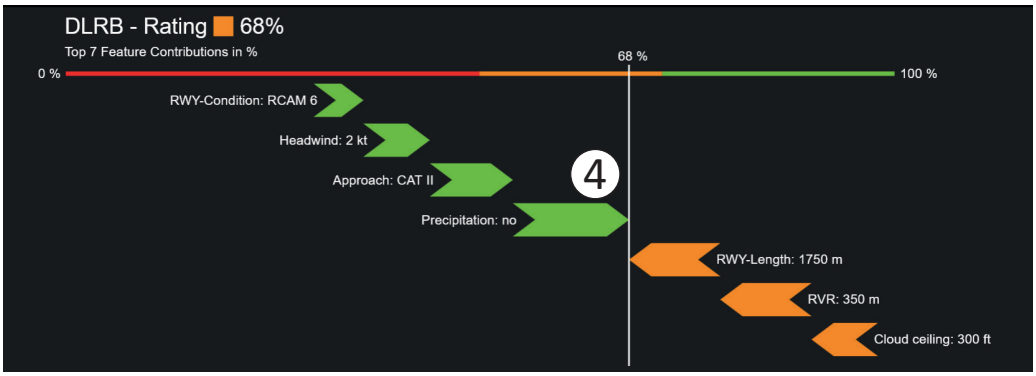


Figure 6 – Explanation Visualisation 3: - fp - Shapley inspired visualisation

Variant fi, shown in Figure 7, presents feature contributions as a bar chart ⑤ combined with a color scale ⑥ that indicates how close each factor is to its critical threshold.

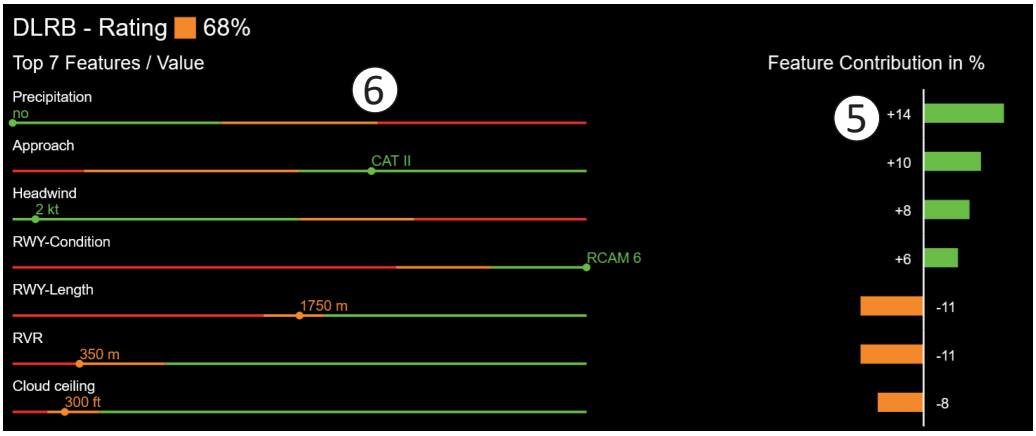


Figure 7 – Explanation Visualisation 4: fi - Feature contribution and rating areas

In both variants fp and fi, only the seven most important factors are displayed, as the significance of the unimportant factors is very low and only those that have an actual influence are shown.
Here is an overview of all the display variants and their differences:

- rtb (baseline table): shows only the raw decision-factor values; no color coding or contribution information is provided (Fig. 4).
- rtbc (color-coded table): adds the color-coded factor rating to the baseline table, enabling pilots to see at a glance which factors are favourable or unfavourable (Fig. 5).
- fp (force-plot): builds on the color-coded table and adds a SHAP-inspired force-plot that visualises feature importance as directional beams, thereby communicating both sign and magnitude of each factor's effect (Fig. 6).
- fi (feature-contribution bar chart): combines the color-coded factor rating with a bar-chart that explicitly displays the quantitative contribution of each factor and overlays the threshold visualisation, indicating how close a factor is to a critical limit (Fig. 7).

At the time of the study the actual AI model was still under development and did not provide the necessary data. Therefore the airport-rating results shown on the displays-mockups were produced by hand, according to the Wizard-of-Oz approach [57]. The displays therefore present AI-like ratings that enable the evaluation of explainability without a fully operational AI model.

5. Evaluation of Explainable User Interfaces for the AICIS

The study in this paper builds on the work of Schog [58]. In the study, the four explainable display variants described in Section 4.2.2 were evaluated by airline pilots with respect to perceived explainability, perceived usefulness for decision-support, and operational suitability.
Operational suitability refers to a display's suitability for use for decision-making in an operational cockpit environment. Usefulness for decision support means that the displayed information and its visualization support the pilots in decision-making in a useful way.

5.1 Method

5.1.1 Participants

Twelve airline pilots participated in the study. All were male, native-German speakers and held a commercial pilot licence with an Airbus type-rating. Four pilots were younger than 30 years, six pilots aged 30-39 years and two pilots were older than 39 years. Four pilots had logged between 1501 and 3000 flight hours and eight pilots had logged 3001 and more flight hours. Eight pilots served as First Officer, two as Senior First Officer and two as Captain.

5.1.2 Study Design

Experimental design The study was design as within-subjects design: every pilot used and evaluated all four display variants. All participants started with the baseline variant (rtb). To avoid learning or order effects the order of the three enhanced displays (rtbc, fi, fp) was rotated according to a Latin-square design. In addition, the underlying decision scenario was varied for each trial. Each scenario contained a different configuration of airport attributes (e.g., runway length, weather, visibility, cross-wind) together with the corresponding weightings and ratings.

Wizard-of-Oz implementation of AI-results At the time of the study the AI core (AICOM) and the cockpit interface (AICIS) were still under development and had not yet been implemented as a fully connected system. In line with the Wizard-of-Oz paradigm, the AI outputs displayed to the pilots were produced manually by domain experts who applied the multi-criteria decision-making logic that drives the AI core. The resulting airport-rating values were then injected into the prototype displays. No machine-learning algorithm were used. This set-up allowed the evaluation of the explainable display variants without requiring a fully operational AI backend.

Test environment The study was conducted on a desktop workstation; no cockpit simulator or additional hardware was used. No secondary tasks (e.g., navigation, communication, or other operational duties) had to be fulfilled, so participants could focus exclusively on the information presented on the screen.

Procedure The study procedure comprised the following steps: 1. Briefing: participants received an overview of the study purpose, the four display concepts and the task they would perform. 2. Experimental blocks: for each of the four display variants the pilot performed the decision task and completed questionnaires. 3. Ranking: after the four blocks participants ranked the displays with respect to perceived explainability and operational suitability.

Decision task During each block, the pilot was presented with a fictitious situation in which the originally planned destination became unavailable, the aircraft remained airworthy, and all passengers were safe. The system offered three alternate airport options, along with an AI-generated rating. Depending on the display variant, it also offered associated explainability information, such as color

coding, feature importance, and threshold visualization. Pilots were instructed to first try to find explanations to understand how the AI arrived at each rating (operational explainability), and second, to select the most suitable alternate airport for the simulated emergency.

5.1.3 Metrics

The following metrics were used to capture the pilots' evaluations of each display variant.

- **Explanation Satisfaction Scale (ESS).** The ESS assesses perceived explainability of the interface. Details of the scale are provided by Hoffmann et al. [56].
- **Tailor-made usefulness questionnaire.** This short questionnaire examines the perceived usefulness and information provision of the displays. The three items and their wording are listed in the Results section (see Table 3).
- **Post-experiment ranking.** After completing all trials, participants ranked each display on two dimensions (1 = best, 4 = worst):
  1. *Explainability*: How well does the display explain the rationale behind the AI's recommendation?
  2. *Operational decision-support suitability*: How suitable is the display for making the final airport-selection decision in an operational context?

5.2 Results

5.2.1 Explanation Satisfaction Scale

The satisfaction of the explainability of the four display variants was evaluated by the participants using the Explanation Satisfaction Scale (ESS). Figure 8 shows the boxplot of the rating results of the ESS.

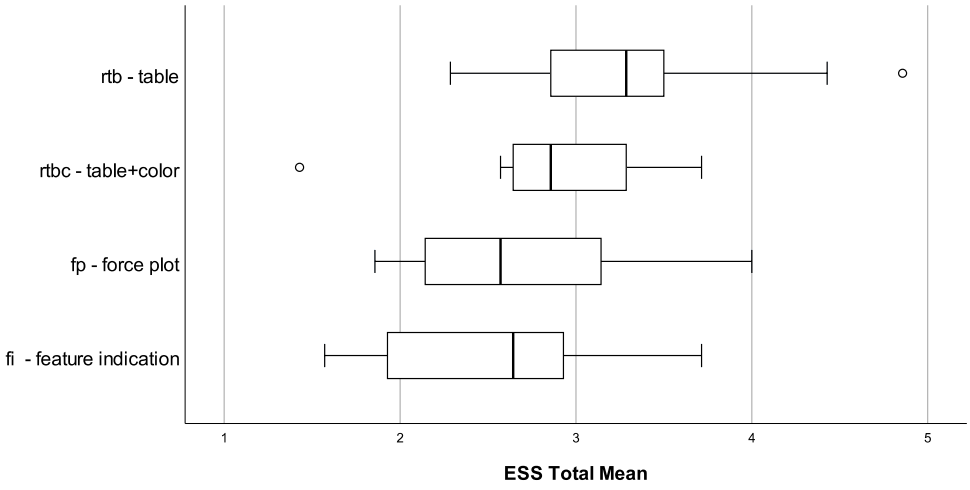


Figure 8 – Explanation Satisfaction Scale (ESS) - Boxplot

Descriptive statistics are shown in Table 2. Participants rated each variant on a 5-point Likert scale (1 = strongly agree, 5 = strongly disagree). Lower mean scores indicate higher perceived explainability.

Table 2 – Explanation Satisfaction Scale (ESS) scores for the four interface variants.

Display Variant	Mean	SD	Levene p-Value	Shapiro-Wilk	N
rtb – table	3.32	0.73	0.927	0.34	12
rtbc – table+color	2.89	0.60		0.17	12
fp - force plot	2.67	0.67		0.33	12
fi – feature indication	2.51	0.67		0.72	12

The feature-indication variant (fi) obtained the lowest mean score (M = 2.51), indicating the highest level of satisfaction with the explanation. This was followed by the force-plot (fp) and the color-enhanced table (rtbc), both of which also yielded mean values below the neutral midpoint (3.0). The raw-data table (rtb) received the highest mean (M = 3.32), reflecting the lowest perceived explainability among the four designs.
Levene's test shows homogeneity of variances ($p = .927 > .05$). The Shapiro–Wilk test is non-significant for every condition ($p > .05$), confirming normality. Mauchly's test of sphericity is also non-significant ($W(5) = .621,\ p = .465$), so the sphericity assumption holds. A repeated-measures ANOVA therefore reveals a significant difference between the display variants: $F(3,33) = 4.672,\ p = .008,\ \eta_p^2 = .298$. According to Cohen [59], $\eta_p^2 = 0.298$ represents a large effect size.
Bonferroni-adjusted post-hoc t-tests identify a single significant difference between the fi and the rtb, t(11) = 2.96, p = .029. fi receives significantly higher satisfaction scores than rtb. All other pairwise comparisons are non-significant.
In summary, participants rated the feature-indication display variant as significantly more satisfying than the raw-data table variant. The remaining display variants did not differ statistically from one another.

5.2.2 Tailor-made questionnaire on perceived operational usefulness

The participants were asked to rate the four display variants with respect to their usefulness for operational decision-making. All items were answered on a 5-point Likert scale (1 = strongly agree, 5 = strongly disagree). Table 3 presents the descriptive statistics for the three questionnaire items.

Table 3 – Perceived operational usefulness of the four interface variants.

Item	rtb		rtbc		fp		fi	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
The display provided me with useful support in selecting a suitable alternative airport.	2.83	0.99	2.50	0.65	3.00	1.00	2.75	0.83
The display provided me with too much information.	1.50	0.50	1.50	0.50	2.08	0.95	2.00	0.71
Important information were missing, to make an informed decision.	3.92	0.86	3.42	1.04	3.33	1.03	3.08	1.19

All four interface variants were rated between 2.50 and 3.00 with respect to their perceived usefulness for supporting the selection of an alternate airport. The color-enhanced table variant was rated most useful (rtbc = 2.50 ± 0.65), followed by the feature-indication variant (fi = 2.75 ± 0.83), the raw-data table (rtb = 2.83 ± 0.99), and finally the force-plot variant (fp = 3.00 ± 1.00). Regarding the "too much information" item, the baseline variant (rtb) and the color-enhanced variant (rtbc) obtained identical means of 1.50 (rtb = 1.50 ± 0.50; rtbc = 1.50 ± 0.50). The force-plot (fp) and feature-indication (fi) variants yielded slightly higher means of 2.08 (± 0.95) and 2.00 (± 0.71), respectively. Because the Likert scale is anchored at 1 = strongly agree and 5 = strongly disagree, the shown lower scores represent stronger agreement with the statement that the display provides too much information. The participants stated that the variants rtb and rtbc provided too much information, whereas the variants fp and fi were perceived as slightly less overwhelming despite still being information-dense.
For the item addressing missing important information, the mean ratings ranged from 3.08 to 3.92. The raw-data table variant showed the highest mean (rtb = 3.92 ± 0.86; least perceived missing information), whereas the feature-indication variant showed the lowest mean (fi = 3.08 ± 1.19; most perceived missing information).

5.2.3 Rankings of display variants

The twelve participants ranked each display from 1 (best) to 4 (worst) for explainability and for decision-support suitability. The individual ranks were added together, so that a lower sum repre-

sents a better rank.

Table 4 – Summed ranks of the four display variants in terms of explainability and operational decision-support suitability.

Display	Explainability	Decision-support suitability
rtb	42	41
rtbc	31	25
fp	20	30
fi	19	24

For explainability the feature-indication variant (fi) achieved the best total rank, closely followed by the force-plot (fp), while the raw-data table (rtb) received the worst rank. For operational decision-support suitability, fi again ranked best, with the color-enhanced table (rtbc) only slightly behind, whereas fp and especially rtb were rated less favourably.

5.3 Limitations

The limitations relate to two key aspects. First, the displays were tested as click dummies on a computer, meaning that there was no operational context in a cockpit environment. The participants were therefore not engaged in any other operational tasks, as would be the case in a real cockpit in decision-making situations. This means that no conclusions can be drawn about the operational usefulness and suitability of the display variants in a real cockpit environment. In future, it would also be advisable to evaluate such displays in an operational cockpit environment in order to assess their operational usefulness, suitability and explainability in this context, where factors such as parallel tasks, information from cockpit systems and other operational factors may influence explainability and decision support.
Secondly, the study did not use data generated by an AI model, but rather data simulated by humans (Wizard of Oz paradigm), so that the explainability and usefulness of AI-generated data was not captured, but rather the explainability and usefulness of the way AI-based results are presented. However, once the basic appropriateness of the display variants for the purposes of explainability and usefulness on decision making has been demonstrated, the display variants should be tested again with real data generated by AI as soon as technically possible in order to confirm their suitability.

6. Discussion

The development of the AICOM and the findings from the AICIS mockups study offer relevant insights into introducing explainability in operational AI systems, bridging technical feasibility with user-centered design. The following discussion synthesizes the key lessons, their implications for cockpit interface design, and their potential transfer to future applications such as AI-supported MUM-T mission management.

6.1 Lessons from the development of the AICOM

6.1.1 Explainability considerations should be embedded from the outset, not added after the fact.

A central lesson from the AICOM development is that explainability cannot be treated as an add-on. Decisions made early in the design process, including the choice of feature representation, the structure of the evaluation function, and the formulation of the ranking task, all had direct consequences for what could later be explained and how reliably. In particular, the deliberate avoidance of dimensionality reduction techniques such as PCA was an important design decision. Although such techniques can improve model performance, the resulting latent features are not directly interpretable by domain experts or flight crews, which would undermine the transparency of the system and make it harder to validate explanation outputs against operational expectations [27].

6.1.2 The choice of explanation method cannot be decoupled from the model architecture and input structure.

The feature importance analysis revealed that SHAP and Integrated Gradients produced misleading results for the neural network due to the repeated context features in the ranking input structure, while PFI provided stable and operationally plausible outputs. This was not a failure of the model but a failure of the explanation method to handle the specific input representation. The practical implication is that explanation pipelines must be validated not only technically but also against domain expectations before their outputs are surfaced to end users. No single method can be assumed to work reliably across all model classes and input structures.

6.1.3 The performance–interpretability trade-off extends to explanation reliability.

The neural network achieved the strongest ranking performance but proved the most difficult to explain reliably. The weighted-sum model and simple LCS variant, while lower in raw performance, produced stable and consistent explanation outputs that could be directly verified against the evaluation function. This suggests that in operational contexts where the quality of explanation information passed to the crew interface is critical, a more interpretable model may in some cases be preferable to the highest-performing one, depending on the deployment priorities.

6.1.4 Multiple explanation methods should be applied in parallel and cross-validated.

No single explanation technique provided a complete and reliable picture across all model classes. The combination of model-specific methods, such as gain-based importance for XGBoost and population-based importance for LCS, with model-agnostic methods such as PFI and SHAP provided a richer and more trustworthy overall assessment. Discrepancies between methods should not be dismissed as noise but treated as signals that warrant further investigation before explanation outputs are used operationally.

6.2 Lessons from the AICIS Mockup Study

The study investigated four display variants, explaining AI-generated alternate-airport ratings. Professional pilots evaluated each variant addressing explanation satisfaction and perceived usefulness.

6.2.1 Color-coding indicating individual decision-factor ratings improves explainability.

The table that presents the decision-factors with color-coded ratings (rtbc) and the two SHAP-inspired variants (fi-feature-contribution bar chart and fp-force plot), which also color-code the factors, achieved higher ESS scores than the raw-data baseline (rtb), which lists all decision factors without any indication of their individual rating. This visual cue, which distinguishes between positive and negative factor contributions to the overall airport-rating, notably increases pilots' satisfaction with the explanation.

6.2.2 Adding information about decision-factor contribution leads to better perceived explainability.

The two SHAP-inspired variants (fp and fi) outperformed both table-based variants in terms of explanation satisfaction. The feature-contribution bar chart (fi) variant achieved the best ESS score and was the only design that differed statistically significantly from the raw-data baseline (rtb). This finding supports the suggestion that pilots are more satisfied with explanations that provide more information about how each decision factor influences the overall rating, including the extent of each contribution, than with those that provide only a simple positive/negative indication. When participants ranked the displays from most to least suitable for explanation, the SHAP-inspired variants (fp and fi) were clearly preferred, whereas the raw-data baseline (rtb) variant was ranked last. Thus, the ranking results are fully consistent with the ESS results.

6.2.3 The color-enhanced table (rtbc) is judged most useful and suitable for decision support.

In the tailor-made questionnaire, the color-enhanced table (rtbc) variant received the best score in terms of usefulness and suitability for selecting an alternate airport. Additionally, the decision-support

suitability ranking results show that the feature-contribution bar chart (fi) and color-enhanced table (rtbc) variants are the most suitable for decision support, while the force plot (fp) and baseline (rtb) variants are less suitable. This indicates that a table showing all the decision factors and their positive or negative contribution to the overall airport rating is sufficient for making a decision, even if the explanation is not the most satisfactory. By contrast, the SHAP-inspired force plot (fp) variant was rated as the least useful, but still achieved good ESS-results.

6.2.4 SHAP-based visualisations lessen pilots' perceived information overload even though they provide richer explanatory detail.

The two table-based variants (rtb and rtbc) were considered to provide too much information, whereas the SHAP-based variants (fp and fi) were rated more favourably. Nevertheless, even the SHAP-based displays were considered to contain excessive information. Although the SHAP-inspired variants (fp and fi) provide more explainability-related information about the AI results, such as feature contributions and rating-threshold visualisations, their presentation format, which uses colored bars, directional beams and scaled arrows, appears to reduce the perceived information load compared to tables containing numeric data. Additionally, the SHAP-inspired variants (fp and fi) variants display only the seven most significant factors, rather than the entire set of twelve decision factors, as is the case with table-based variants (rtb and rtbc). This reduces the amount of information that pilots must process.

6.2.5 Feature indication and color coded tables striking a good balance between operational explainability and operational usefulness.

The results show that the most explainable design (fi-feature-contribution bar chart) is not automatically the most useful for decision-support: color-enhanced table (rtbc) obtained the highest usefulness scores, whereas feature-contribution bar chart (fi) achieved the best explainability rating. The feature-contribution bar chart (fi) variant provides greater operational explainability, while the color-enhanced table (rtbc) variant is more useful for decision support. Consequently, a single visualisation cannot optimise both criteria simultaneously. Designers must either prioritise one of the criteria or combine the two approaches. Looking at the ranking, both feature-contribution bar chart (fi) and color-enhanced table (rtbc) were ranked as the most operationally suitable for decision support. This suggests that both displays are suitable for operational use, striking a good balance between operational explainability and usefulness.

6.2.6 Derived Design Implications

Integrating the findings on explainability, the usefulness of decision support, perceived information overload, completeness, and the balance between explainability and operational usefulness suggests a possible design direction: a cockpit display combining two visualisations that allows pilots to adjust the level of explanation according to their current needs:
Primary quick-look view that...

- shows the overall airport rating, together with color-coded decision factors indicating the rating for each factor.
- provides a fast, operationally useful overview without overloading the pilot's cognition.

Secondary "explain-on-demand" view that ...

- reveals the extend of the contribution of each factor and the proximity of factor values to their rating-thresholds.
- displays only the most influential factors, thereby keeping the amount of information manageable while delivering deeper insight when required.

This dual-layer approach respects the explainability-performance trade-off identified in the study: a quick-look table (rtbc) satisfies the need for rapid, operationally useful information, with superficial explainability, whereas the detailed feature-indication view (fi) supplies the depth of explanation required while containing operational suitability and preventing the display from becoming cluttered. Consequently, the cockpit interface remains useful, explainable, and information-reduced, ensuring real-world flight-operation suitability.

6.2.7 Summary of the Explainability-information derived from the Interpretability Features.

Table 5 summarizes the results of the user study on how the implemented Explainability-Information and its visual presentation contribute to operational explainability. These pieces of information were originally derived from, and map directly back to, the high-level Interpretability-Features introduced in Section 4.2. They can therefore be formalised as UI-level requirements for future cockpit-based AI-system interfaces.

Table 5 – Impact of the three explainability information derived from the Interpretability Features on explainability

Explainability information	**What it tells the user**	**Observed impact on ratings**
Raw input data (presentation of raw data)	Raw decision-factor values as they are received by the AI-core	Removing the individual raw values caused pilots to report missing information, demonstrating that displaying the raw data is especially important for operational suitability. Moreover, every subsequent Explainability-Information (color-coding, feature-importance indicator, rating threshold visualisation) are based on these raw inputs, making them an integral part of the overall Interpretability Features concept.
Indication of decision-factor rating (color-coded: green = positive, amber/red = negative)	A brief visual cue indicating whether each factor contributes positively or negatively to the overall airport rating.	Produced the largest single improvement over the baseline (rtb $\rightarrow$ rtbc). Color-coding alone raised the ESS scores notably.
Feature-importance indicator (fp: arrow-based plot, or fi: bar-chart)	Shows how strongly a factor pushes the overall rating up or down.	Adding this information further increased perceived explainability. Between the two visualisations, the fi-display, which shows feature-importance on a separate bar chart, was preferred to the force-plot (fp), which uses arrows of varying length and direction.
Threshold visualisation (display of rating boundaries on a scale)	Makes the distance of a factor's value to its rating thresholds visual. Shows when a factor would receive a different rating.	The visualisation of thresholds in the fi variant helped it achieve the highest overall explainability score, with a significant difference to the baseline (rtb).

6.2.8 Discussion on Technical Feasibility from the AI Perspective

A recurring theme in the AICOM development is the question of what explanation information can realistically be extracted from a trained model and passed to the AICIS in a form that is both technically reliable and operationally meaningful. The answer depends strongly on the model class in use, the intended purpose of the explanation, and the computational constraints of the deployment context.
An important distinction must be made between the two main purposes for which explanation information is needed. During system development and validation, global explanations are required: developers and domain experts need to understand overall model behaviour, verify that the learned logic aligns with the evaluation function, and compare models across the interpretability-performance spectrum. During operational use, local explanations are needed: for each individual diversion decision, the AICIS must be able to present the flight crew with a meaningful breakdown of why a specific airport was rated as it was, based on the current situation.

SHAP is primarily a local explanation method. It provides per-prediction feature attributions that show how much each decision factor contributed to a specific airport rating relative to the expected baseline output, making it well suited in principle for the operational use case. However, as demonstrated in the AICOM experiments, SHAP did not produce reliable results for the neural network in the ranking formulation. The repeated context features in the input vector caused SHAP to assign overwhelming importance to context factors and near-zero attribution to the airport-specific decision factors, a result that contradicts both the model's strong performance and the structure of the evaluation function. This means that for the neural network, SHAP cannot currently be recommended as a source of local explanation information for the AICIS, regardless of its theoretical properties.
For tree-based models the situation is more favourable. TreeSHAP computes exact local attributions efficiently by exploiting the tree structure, and the results were consistent and operationally plausible in the AICOM experiments. From a computational standpoint, TreeSHAP is feasible for real-time use, making it a viable option for passing local feature contributions to the AICIS for XGBoost-based deployments [27]. For neural networks, KernelSHAP would require hundreds of forward passes per prediction, introducing latency that would be problematic in time-critical operational settings, and as noted above, the reliability of the results is also in question for the current input structure [60].
For global explanation needs during development, permutation-based importance proved to be the most reliable and consistent method across all model classes in the AICOM experiments. It carries no dependency on model architecture or input structure, and for the neural network it was the only method that produced stable and domain-consistent importance signals. It is therefore the recommended method for validating explanation pipelines and assessing overall model behaviour during development, even if it does not provide the per-prediction granularity needed for operational explanations.
In summary, the current feasibility picture is as follows. For interpretable models such as the weighted-sum and simple LCS variants, both local and global explanation information can be extracted reliably and at low computational cost. For XGBoost, TreeSHAP provides feasible and reliable local attributions suitable for operational use. For the neural network, local explanation via SHAP is currently not reliable in the ranking formulation, and permutation-based global importance is the recommended fallback. Addressing this gap, either through input restructuring, alternative local explanation methods, or a shift to a more interpretable model, is an important direction for future work before the full explanation pipeline from AICOM to AICIS can be considered operationally ready.

6.3 Transfer to AI-Supported MUM-T Mission Management in Future Air Combat

The results of the current study, namely that the feature-indication visualisation achieved the highest explanation-satisfaction score while the color-coded table was rated most useful for rapid decision support, can be directly integrated onto the HMI concept shown in Figure 1. That HMI-concept already employs icons and empirically validated color coding to convey key mission parameters and their states. The feature-indication visualisation approach can be integrated into the HMI-concept by adding another on-demand layer showing the quantitative contribution of each parameter (e.g., the TreeSHAP-derived impact of a pop-up surface-to-air missile on the risk parameters) alongside the existing icons, thereby providing pilots with deeper, model-based explanations when needed. This design would leverage the proven color-coding for quick situation awareness and enrich it with the feature-indication approach for improved interpretability.

7. Conclusion & Outlook

This paper introduces two use cases for AI-based assistant systems in aviation, developed at the DLR: a military Manned-Unmanned Teaming (MUM-T) mission management system and the Intelligent Pilot Advisory System (IPAS). The growing use of AI for safety-critical systems in aviation demands operational AI explainability for end-users that supports calibrated trust, situation awareness, and effective human-AI collaboration. To address this need, this paper presented a multi-level approach for the IPAS that couples explainability-oriented design on the AI-side (AICOM) with a user-centred development process for the cockpit interface (AICIS). Four prototype displays, a raw data table, a color-coded table, a SHAP-inspired force-plot, and a feature-contribution bar chart, were

evaluated in a Wizard-of-Oz study with twelve professional airline pilots. Participants rated the displays on perceived explainability, perceived usefulness for decision support, and overall operational suitability. The results demonstrate a clear trade-off between explanatory depth and operational usability. The study indicates that explainability satisfaction benefits the most from concise visualisations of factor importance and threshold proximity. The color-coded table display received the highest usefulness ratings, suggesting that a quick overview of factor ratings is most suitable for rapid decision-making under workload.By contrast, the baseline, which was a table of raw data, performed worst across all metrics. This confirms that raw data without any explanation does not satisfy the explainability needs of pilots. A hybrid interface that offers a quick-look view for immediate operational support together with an on-demand, detailed explainability view appears most promising for future AI-enabled cockpit systems. The findings of the study, combining quick color coded overviews with detailed feature-contribution visualizations, can be directly applied to MUM-T HMI to provide pilots with adaptive explainability tailored to mission demands. Integrating SHAP- or TreeSHAP-based parameter impacts into existing icon- and color-coded interfaces could enhance trust without compromising operational speed in time-critical scenarios.
However, the integration of SHAP-based methods requires confirming that the method provides accurate information from the model and that Shapley values can be computed sufficiently fast for operational use. This depends on the type of model involved: for XGBoost, a custom TreeSHAP variant exists that should prove sufficiently fast, but for models without a dedicated SHAP implementation, a model-agnostic approach would be required, which can be considerably slower. This may be partially offset by the fact that only a single local explanation needs to be produced per inference, a task for which SHAP is well suited. Nonetheless, the feasibility of real-world deployment should be validated and tested before operational integration.

7.1 Multi-Level Implementation of Explainability

The development of the IPAS demonstrates that addressing operational AI explainability at all levels of system design, from AI model selection to human-machine interface development, can be beneficial. A hand-in-hand top-down and bottom-up approach was applied: In the top-down approach, human-centered explainability requirements were derived from pilot needs, task analyses, and operational constraints. From the bottom-up side, interpretable AI methods and explanation extraction techniques were explored to ensure technical feasibility. This parallel development enabled continuous alignment between the technical capabilities of the AI itself (AICOM) and the operational requirements of the crew interface (AICIS). At the AI-core level, this meant carefully selecting models and input representations that enabled reliable extraction of explanation information, accepting that the most performant model might not always be the most explainable one. The development process itself benefited from early and recurrent end-user involvement through interviews, user tests, and iterative prototyping. This was essential for identifying the operational context and explainability needs of the end users. It enabled the AICIS to explain AI results in the 'language of airline pilots'.
While there is no one-size-fits-all solution for operational explainability, aspects from all levels of the IPAS explainability research can be transferred and adapted to new use cases and AI-based systems. High-level requirements, such as the interpretability features identified can be adapted to a specific use case by developing and implementing design features derived from them. Similarly, concrete design decisions and implemented features, which have been developed and detailed within the AICIS, only need minor adaptations to fit the context of other use cases. This enables each development team to adopt different components as required to meet the specific needs of their system end users, taking into account their individual system development status, operational context and explanation requirements.
This approach appears to work, as elements of the IPAS explainability effort can be integrated into the MUM-T HMI. There, existing icon- and color-based visualizations can be enhanced with AICIS findings, such as model-derived parameter impacts, to provide operators with adaptive explainability tailored to mission demands.

7.2 Future Work

The next step in developing the IPAS will be to bridge the gap between algorithmic and operational explainability by integrating the AI core with the crew interface. A middleware component will be developed to extract artefacts (e.g. SHAP and TreeSHAP exports, factor ratings) and translate them into the visual elements and interpretability features validated in this study. This approach must be tested for feasibility to ultimately merge the bottom-up and top-down approaches into a single demonstrator of operational AI explainability.

To ensure broader applicability, the findings and lessons learned from the DLR work on operational AI explainability will be transferred to other aviation use cases and safety-critical domains, such as future air combat scenarios like MUM-T. To enable this transfer, guidelines that support the development of AI-based, operationally explainable assistant systems will be defined and tested. The guidelines may consist of a set of development-process tasks, processes and methods, a catalogue of design considerations to take in the development process, a set of required interface features (e.g. interpretability features and implementation examples), or any combination of these elements.

Beyond usability, explainability is increasingly a regulatory prerequisite for certifying AI-based systems in aviation. EASA's roadmap and concept paper explicitly list AI explainability as a core requirement alongside learning assurance and safety risk assessment [1, 2]. This positions methods such as SHAP not only as tools for human-centred interface design, but as potential contributors to the broader certification evidence required for AI deployment in safety-critical aviation contexts.

8. Contact Author Email Address ans phone number

Jakob Würfel jakob.wuerfel@dlr.de, +49 531 2951 091
Boris Djartov boris.djartov@dlr.de
Max Friedrich max.friedric@dlr.de
Sabrina Schog sabrina.schog@dlr.de
Anne Papenfuss anne.papenfuss@dlr.de

9. Copyright Statement

The authors confirm that they, and/or their company or organization, hold copyright on all of the original material included in this paper. The authors also confirm that they have obtained permission, from the copyright holder of any third party material included in this paper, to publish it as part of their paper. The authors confirm that they give permission, or have obtained permission from the copyright holder of this paper, for the publication and distribution of this paper as part of the ICAS proceedings or as individual off-prints from the proceedings.

References

[1] European Union Aviation Safety Agency. Easa-ai-roadmap 2.0: A human-centric approach to ai in aviation, 2023.

[2] European Union Aviation Safety Agency. Easa concept paper: Guidance for level1 &2 machine-learning applications, 2024.

[3] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G. Michael Youngblood. Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8, 2018.

[4] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. Euca: the end-user-centered explainable ai framework, 2022.

[5] Jakob Würfel, Anne Papenfuß, and Matthias Wies. Operationalizing ai explainability using interpretability cues in the cockpit: Insights from user-centered development of the intelligent pilot advisory system (ipas). In Helmut Degen and Stavroula Ntoa, editors, *Artificial Intelligence in HCI*, pages 297–315, Cham, 2024. Springer Nature Switzerland.

[6] Gabriella Gigante, Arthur Clavière, Dmitrii Kirov, ELISA SPILLER, Ana Ferreira, Anne Papenfuss, Jakob Würfel, Elena Rankova, Antoine Alaverdov, and Rolf Zon. Let ai fly. the jarvis project: Executive summary of ai challenges whitepapers, May 2026.

[7] Mica R. Endsley. Supporting human-ai teams:transparency, explainability, and situation awareness. *Computers in Human Behavior*, 140:107574, 2023.

[8] Charles E. Billings. Human-centered aviation automation: Principles and guidelines. Technical Memorandum NASA-TM-110381, NASA Ames Research Center, Moffett Field, CA, USA, 1996.

[9] Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, 1983.

[10] Eearl L. Wiener and Renwick E. Curry. Flight-deck automation: promises and problems. *Ergonomics*, 23(10):995–1011, 1980.

[11] Mica R. Endsley. Automation and situation awareness. In *Proceedings of the 1996 International Conference on Human-Computer Interaction*, 1996.

[12] European Union Aviation Safety Agency. Crm training implementation, 2017.

[13] David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. Darpa's explainable ai (xai) program: A retrospective. *Applied AI Letters*, 2(4), 2021.

[14] European Commission, Content Directorate-General for Communications Networks, and Technology. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Publications Office, 2020.

[15] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[16] Laurent Dupont, Olivier Fliche, and Su Yang. Governance of artificial intelligence in finance, 2020. Aufgerufen am 07.10.2025.

[17] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences, 2017.

[18] Neta Ezer, Sylvain Bruni, Yang Cai, Sam J. Hepenstal, Christopher A. Miller, and Dylan D. Schmorrow. Trust engineering for human-ai teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1):322–326, 2019.

[19] Chloë Droogenbroeck, Elena Rankova, Anne Papenfuss, Tanja Bos, and Rolf Zon. Human-ai teaming – challenges from a practitioner's perspective. In *Engineering Psychology and Cognitive Ergonomics*, Springer Nature Switzerland, pages 337–354, 2025.

[20] Raja Parasuraman and Dietrich Manzey. Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52:381–410, 06 2010.

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[22] Robert R. Hoffman, Todd Miller, Gary Klein, et al. Increasing the value of xai for users: A psychological perspective. *Künstliche Intelligenz*, 37:237–247, 2023.

[23] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[24] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[25] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017.

[26] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774. Curran Associates, Inc., 2017.

[27] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2 edition, 2022.

[28] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[29] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. Pitfalls to avoid when interpreting machine learning models, 2020.

[30] Jimmy Hammarbäck, Jens Alfredson, Björn JE Johansson, and Jonas Lundberg. My synthetic wingman must understand me: modelling intent for future manned–unmanned teaming. *Cognition, Technology & Work*, 26(1):107–126, 2024.

[31] T Mehling, F Viertler, and T Paul. Potential of the unmanned wingman in manned-unmanned teaming tested in flight. In *Deutscher Luft-und Raumfahrtkongress*, 2019.

[32] Tobias Paul and Emanuel Brämer. Operational considerations for teaming manned and unmanned helicopter. *Journal of Intelligent & Robotic Systems*, 69(1):33–40, 2013.

[33] Nancy J Cooke, Mustafa Demir, and Nathan McNeese. Synthetic teammates as team players: Coordina-

tion of human and synthetic teammates. Technical report, 2016.

[34] Mica R Endsley. Supporting human-ai teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior*, 140:107574, 2023.

[35] Batuhan Özcan, Max Friedrich, Jan-Paul Huttner, and Kevin Dwinger. Enhancing trust in human-ai interaction through explainable decision support systems for mission planning of uas swarms. In *AHFE 2025 International Conference*, 2025.

[36] Qiang Peng, Husheng Wu, and Ruisong Xue. Review of dynamic task allocation methods for uav swarms oriented to ground targets. *Complex System Modeling and Simulation*, 1(3):163–175, 2021.

[37] Birte Thomas-Friedrich, Johannes Maria Ernst, and Max Friedrich. Human system integration in future ai-supported mum-t mission management. *Intelligent Human Systems Integration (IHSI 2026): Disruptive and Innovative Technologies*, 200:307–317, 2026.

[38] T. Niebuhr. Enhancing NATO's strategic edge: A human-centric approach to multi-domain operations. *The Journal of the JAPCC*, 38:28–35, 2024.

[39] Johannes Maria Ernst, Birte Thomas-Friedrich, and Max Friedrich. Mission management in human autonomy teams–an hmi design-concept for managing multiple uncrewed aerial systems from a fighter cockpit. *Intelligent Human Systems Integration: Disruptive and Innovative Technologies*, page 227, 2026.

[40] Max Friedrich and Mark Vollrath. Human–machine interface design for monitoring safety risks associated with operating small unmanned aircraft systems in urban areas. *Aerospace*, 8(3):71, 2021.

[41] EUROCONTROL. Cost of diversion, 2025.

[42] SKYbrary. Diversion, 2025.

[43] H. J. Hörmann. For-dec - a prescriptive model for aeronautical decision making. In *21st WEAAP Conference, Dublin*, 1994.

[44] Boris Djartov and Jakob Würfel. *Navigating Decisions in the Cockpit: The Intelligent Pilot Advisory System*, pages 305–325. Springer Nature Switzerland, Cham, 2025.

[45] Jakob Würfel, Boris Djartov, Anne Papenfuß, and Matthias Wies. Intelligent pilot advisory system: The journey from ideation to an early system design of an ai-based decision support system for airline flight decks. In *AHFE 2023*, 2023.

[46] Sarah Ternus, Jakob Würfel, Anne Papenfuß, Matthias Wies, and Martin Rumpler. Exploring functionalities fornbsp;annbsp;intelligent pilot advisory system innbsp;normal operation. In *Engineering Psychology and Cognitive Ergonomics: 21st International Conference, EPCE 2024 (HCII 2024), Washington, DC, USA*, page 235–247, Berlin, Heidelberg, 2024. Springer-Verlag.

[47] Boris Djartov, Sanaz Mostaghim, Anne Papenfuß, and Matthias Wies. Description and first evaluation of an approach for a pilot decision support system based on multi-attribute decision making. In *IEEE Symposium Series on Computational Intelligence*, pages 141–147, 2022.

[48] Boris Djartov, Sanaz Mostaghim, Anne Papenfuß, and Matthias Wies. A learning classifier system approach to time-critical decision-making in dynamic alternate airport selection. In *2024 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, 2024.

[49] Boris Djartov. *Interpretable AI-Enhanced Multi-Criteria Decision-Making for Time-Critical Emergency Applications, with a Practical Application to Dynamic Alternate Airport Selection in Aviation*. PhD thesis, Otto von Guericke University Magdeburg, Faculty of Computer Science, Chair of Computational Intelligence, 2026. Submitted December 2025, defended April 2026.

[50] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9505–9515, 2018.

[51] Helmut Degen, Christof Budnik, Gregory Conte, Andrew Lintereur, and Seth Weber. How to explain it to energy engineers? In Jessie Y. C. Chen, Gino Fragomeni, Helmut Degen, and Stavroula Ntoa, editors, *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, pages 262–284, Cham, 2022. Springer Nature Switzerland.

[52] Frank O. Flemisch, Michael Preutenborbeck, Marcel Baltzer, Joscha Wasser, Christoph Kehl, Reinhard Grünwald, Hans-Martin Pastuszka, and Anja Dahlmann. *Human Systems Exploration for Ideation and Innovation in Potentially Disruptive Defense and Security Systems*, pages 79–117. Springer International Publishing, Cham, 2022.

[53] Jakob Würfel and Frank Ole Flemisch. Human system exploration for the ai-based flight deck decision support system ipas. In *In D. de Waard, D. Manzey, K. Brookhuis, F. Siebert, K. Karrer-Gauß, S. Winkler, F. Di Nocera, A. Toffetti, and T. Franke (2024). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2024 Annual Conference. ISSN 2333-4959 (online). Avail*, 2024.

[54] Alexander Stevens and Johannes de Smedt. Explainability in process outcome prediction: Guidelines to

obtain interpretable and faithful models. *European Journal of Operational Research*, 2023.

[55] Zachary C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, 2018.

[56] Robert R. Hoffmann, Shane T. Mueller, Gary Klein, and Jordan Litman. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5, 2023.

[57] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. Wizard of oz studies - why and how. *Knowledge-Based Systems*, 6(4):258–266, 1993.

[58] Sabrina Schog. Erklärbare KI im Flight Deck: Entwicklung und Evaluierung verschiedener Darstellungsformen für nutzerzentrierte Erklärungen in KI-gestützten Entscheidungsunterstützungssystemen. Bachelor's thesis, FernUniversität Hagen, 2024. `https://elib.dlr.de/210088/`.

[59] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

[60] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4765–4774, 2017.