Human-AI Teaming – Challenges from a Practitioner's Perspective

Chloë Van Droogenbroeck[1], Elena Rankova[2], Anne Papenfuss[2], Tanja Bos[1] and Rolf Zon[1]

[1] Royal Netherlands Aerospace Centre, Anthony Fokkerweg 2, 1059 CM Amsterdam, The Netherlands
[2] German Aerospace Center, Institute of Flight Guidance, Lilienthalplatz 7, 38108 Braunschweig, Germany

Abstract. The integration of Artificial Intelligence (AI) in aviation is gaining momentum, with potential benefits for pilots, air traffic controllers, and airport operations. However, the adoption of AI in safety-critical tasks poses dilemmas for developers and researchers, who must balance the need for proper processes and standards with the complexities of human-AI teaming. Recent research has highlighted the importance of transparency, explainability, and trust in AI systems. When in the future, AI collaborates with the human operator to work towards a shared goal, both the AI and the human need to communicate their motivations and consideration, which poses a different challenge from other applications with higher levels of automation. The EU-funded SESAR project JARVIS has explored these challenges through a workshop with digital assistant designers, identifying key questions such as the need for human-AI teaming versus using AI as a mere tool, how to instil trust in the system and how to facilitate smooth interaction between human and AI. This paper provides an overview of the topics and challenges faced by designers working on higher levels of automation in aviation, with a focus on comparing the identified research gaps and the challenges faced by the practitioners in order to bridge the gap between theory and practice. A comparison with National Academies of Sciences, Engineering, and Medicine's guidance on human-AI teaming reveals areas of agreement and highlights key research directions, while also indicating that current levels of automation are insufficient for effective teaming between human and AI, underscoring the need for further collaboration to address remaining challenges.

Keywords: Human-AI teaming, human factors, digital assistant design, JARVIS.

1 Introduction

Slowly but surely Artificial Intelligence (AI) is making its way into our daily lives. In the field of aviation, digital assistants promise to bring significant benefits by supporting pilots, air traffic controllers and airport operations with advanced automated functions using technology such as machine learning. Even though AI and machine-learning algorithms are still in the early stages of adoption in aviation, automation has been in

use in this domain for decades, especially in avionic systems, providing technology such as the autopilot. Although the use of automation in aviation has been recognised for its positive impact on safety, it has brought certain drawbacks and risks, such as mode confusion [1] and overreliance on automation [2]. Taking into account the possible risks that automation has brought, it is crucial to consider potential negative implications from AI solutions early on.

Although a universally accepted definition of AI does not exist, this technology is frequently related to systems that perform cognitive functions associated with humans [3]. It is thus commonly ascribed a certain level of intelligence. These intelligent qualities of AI solutions introduce new possibilities for interactions, role-division, responsibilities, and ultimately new organisational structures between humans and machines [4], meaning that AI solutions could introduce challenges clearly different from other applications with higher levels of automation

Although not a real human, the growing abilities of AI solutions might eventually enable them to interdependently collaborate with a human operator. This type of dynamic and interdependent shared effort towards a common goal is frequently associated with teaming constructs, and while a universally accepted definition of human-AI teaming does not yet exist, these three aspects are considered a necessary basis for the emergence of human-AI teams [5, 4].

In order to achieve such level of joint effort, a true understanding of each other's intentions and abilities is needed. This requires both the AI and the human to communicate their motivations and considerations to each other, posing novel challenges, implications and topics, such as transparency and explainability that must be accounted for.

In an attempt to provide guidance for safely including AI into systems that incorporate human-AI interaction, human factors scientists are actively developing research agendas and guidance materials addressing the design, development, and validation of AI solutions. However, the question arises whether the existing guidelines and general direction of current research manage to target the actual questions, dilemmas and challenges of developers and practitioners. With the aim of addressing this question, this study set out to gather and analyse current challenges faced by developers working on AI solutions for the aviation domain.

As part of the EU funded SESAR project JARVIS (Just A Rather Very Intelligent System) multiple digital assistants are being developed to support operators in cockpit, ATC and airport domains. The technologies developed in this project, will be referred to as AI *solutions* in the following text. Within this project, a workshop was held, during which team leaders of digital assistant designers (referred to as "practitioners") were asked to share current challenges and main concerns from their projects. This study reports on findings from this workshop and compares the outcomes to relevant research work.

The main objective of the study is to provide insights into similarities and differences between the challenges that practitioners are facing, and current topics addressed in state-of-the-art research.

2 Challenges in Human-AI Teaming: A Literature Review

While used interchangeably in literature, automation and AI are quite different. Automation typically follows predefined rules or scripts to carry out repetitive tasks. It thus does not "learn" or adapt but operates based on set instructions provided by humans. On the other hand, AI systems can learn from data, adapt to new situations and improve their performance over time. AI encompasses a wide range of techniques, including machine-learning, deep-learning, and natural language processing, enabling systems to perform tasks that require cognitive functions, such as reasoning, problem-solving, and decision-making. Issues such as overreliance and confusion regarding the automation's decision making have yet to be solved for the collaboration between humans and automation but will also arise in human-AI teaming and perhaps become even harder to solve due to the more unpredictable nature of AI. This section highlights some of the most researched topics when it comes to human-AI teaming.

2.1 Trust and Overreliance

In general, the definition of trust describes a relationship between two parties – a trustor and a trustee – where the trustor anticipates that the trustee will act in the trustor's best interest, while also accepting the potential vulnerabilities of the trustor's actions [6]. In the context of human-AI teaming, this definition has been viewed from the perspective of contractual trust, meaning that trust between a human and an AI is based on a human believing that an AI system, even in uncertain conditions, will comply to a contract associated with its usage [6]. Here, a contract is understood as any functionality of the AI system and the commitment of ensuring its proper operation. Additionally, contractual trust could also include compliance with general standards relevant to the specific type of AI application. A foundational effort towards developing standardised trustworthiness expectations can be found in the Ethics Guidelines for Trustworthy AI by the European Commission's High-Level Expert Group on Artificial Intelligence [7].

Trust in automation as a human factor topic has been described and researched for several years and is not unique to the introduction of AI into aviation [8]. While trust in automation is a valuable characteristic supporting the acceptance of AI technology, it has also been observed to potentially lead to unforeseen negative effects on the operation of a system. Overreliance, complacency and automation bias are commonly recognised threats to safety, which are highly related to disproportional trust in automation [9, 7]. In relation to AI, complacency is interpreted as suboptimal or lacking monitoring behaviour, where a human is overseeing an automated system with an insufficient frequency, resulting in a directly observable negative effect on performance, such as a failure to identify or a belated identification of a malfunction, anomalous condition or system failure [10]. Automation bias is a closely related phenomenon, emerging from the interaction between a human and a decision-support system, in which the human favours the system's suggestions and uses them as a replacement for information-seeking and thorough analysis of the available information [10].

2.2 Situation Awareness

Situation awareness (SA) is a term originally used in the aviation community and refers to the ability to perceive and understand the environment and situation around oneself, including the location, activities, and status of relevant entities, such as people, objects, and events. It involves being aware of the context and dynamics of a particular situation and using this awareness to make informed decisions or take appropriate actions. In human-AI teaming it is of importance to look at both individual and team SA.

Various definitions of individual SA exist, however, the most commonly used is the Three Level Model from Endsley [9] with:

- Level 1 "Perception": Perceiving the environment and situation;
- Level 2 "Comprehension": Processing and understanding the information; and
- Level 3 "Projection": Projecting future status and outcomes.

A central issue associated with automated systems is the "out-of-the-loop" performance problem. When monitoring an automated system, people are frequently slow in detecting that an intervention is needed. Once they do detect it, they need additional time to determine the state of the system and sufficiently understand it to effectively deal with the issue. This slight delay in performance can lead to catastrophic failures with major consequences.

For team SA, two main concepts exist: shared SA [11] and distributed SA [12]. Shared SA refers to a common understanding or mental model of the situation among a group of individuals. It implies that all members of the group or team are on the same page and have a mutual understanding of the current situation, its implications, and the actions required to respond to it. Distributed SA, on the other hand, refers to a system where SA is spread across multiple individuals or systems, often working independently but coordinating their actions. Instead of having a common mental model, distributed SA involves different participants having parts of the awareness and contributing pieces of information to form a collective understanding.

To optimise human-AI collaboration, fostering team SA is essential, where both humans and AI maintain compatible, complementary perspectives of the situation. Encouraging proactive communication, enhancing transparency, and addressing the tensions between AI and human agency are vital for effective interaction. Furthermore, strategies that support the management of system uncertainty and reduce complexity are key to improving users' understanding and confidence in AI systems [13].

2.3 Transparency and Explainability

Transparency and explainability are often used interchangeably but have a very different meaning. Transparency provides users insight into the system, including the inputs, algorithms and decision-making process whereas explainability ensures users can understand the reasoning behind actions or decisions made by the system. For both transparency and explainability, it is important not to provide the system's user with too much information and/or irrelevant information as this can overwhelm the user and create uncertainty. Additionally, increased transparency combined with decreased AI

reliability can lead to a decrease in trust. Although one could argue that a decrease in trust is justified if the AI is less reliable [14].

Chen, et al. [15] discuss three transparency requirements that highlight the importance of communicating information about:

1. The agent's current state and goals, intentions, and proposed actions;
2. The agent's reasoning process behind those actions and the constraints/affordances that the agent considers when planning those actions; and
3. Information regarding the agent's projection of the likelihood of success/failure and any uncertainty associated with the aforementioned projections.

The uncertainty referred to in point three can be caused by a lack of information and can therefore by reduced by gathering more data or further refining models. This could allow the human operator to compensate where the AI falls short. The uncertainty could however also be caused by the intrinsic randomness of a phenomenon. These levels of transparency were tested in two experiments [16, 17], comparing Level 1, Level 1+2 and Level 1+2+3. The experiments showed that providing more transparency does not necessarily lead to better outcomes.

Explainable AI was developed by researchers to understand the reasoning behind the AI model, much like Level 2 transparency. The explanations do not necessarily reflect the ground truth decision-making of the system, it may simply offer insight into how the decision was reached. According to Rosenfeld and Richardson [18] ,when designing an AI model, the designer must ask themselves the following questions: Does the system need to be explainable, who is going to use my system, when will the explanation(s) need to be presented, and what information should be given in the explanation? As mentioned before, more information does not always improve human decision-making. Additionally, the effectiveness of explanations can also depend on the specific explanation techniques. Silva, Schrum, Hedlund-Botti, Gopalan, & Gombolay [19] researched the impact of several explanation types in a multiple-choice experiment. The study found that the "simplest" or "clearest" explanations (i.e., using natural language) received the highest scores on perceived explainability. This is supported by other research stating that humans typically prefer simpler explanations [20, 19]. The study also found that participants rated the AI they perceived to be more explainable to be more trustworthy, regardless of the explanation method, with participants also performing slightly better when they perceived the AI to be more explainable.

Both transparency and explainability are essential to build the user's trust in the system. However, it is important to find a balance when it comes to trust. Too much trust can lead to complacency, and too little can at best reduce some of benefits gained from having an AI but at worst make it even harder on the user as they now have an increased workload by having to check what the AI is doing on top of performing the task at hand.

3 Method

The study's goal of investigating challenges faced by practitioners and topics currently discussed in research, was pursued by comparing the outcomes of a workshop conducted within the JARVIS project and a report on state-of-the-art research needs related

to human-AI teaming published by the National Academies of Sciences, Engineering, and Medicine [5]. The research work by National Academies of Sciences, Engineering, and Medicine was chosen as a basis for the comparison, due to its extensive nature and detailed approach. Additionally, as this report takes on a military perspective and a relatively practical approach, it already makes a first attempt to bridge the gap between scientific theory and operational practise.

Practitioners' feedback was gathered during a workshop, which was part of a management board meeting of the JARVIS project. Workshop activities on the topic of human-AI teaming lasted 25 minutes and included a pitch introducing the topic of human-AI teaming, two silent brainstorming sessions, one round of the "world café" discussion method within smaller groups [21], and concluded with a collective discussion of the "world café" insights. To motivate participants and accommodate for the short schedule of the workshop, a competitive element was introduced to the "world-café" discussions. The discussion groups were assigned the task to gather as many AI development challenges as possible, and the group with most challenges received a winner's badge.

During the workshop participants documented their answers on paper with a visual signifier of the activity it belongs to. Throughout the workshop, all paper materials produced by participants were gathered and organised into categories based on their corresponding activity. Before their analysis, the results, which consisted of hand-written notes were manually transcribed into a digital format.

The analysis was executed by two of the human factors experts, who facilitated the workshop. Within multiple rounds, the human factors experts analysed the answers independently and assigned them a topic. During the first round, the topics were formulated according to the experts' personal interpretation, without an initial pool of topics. Before each consecutive iteration, the topics were discussed between the two experts and were modified by merging or reformulating topics. After this adjustment, another round of independent topic assignment was conducted, using the previously agreed-upon topics as the pool. After the analysis of the results, the identified topics were compared to topics described in the state-of-the-art report on human-AI teaming by the National Academies of Sciences, Engineering, and Medicine [5].

4 Participants

Participants in the workshop included twenty practitioners, who were assisted by three human factors experts in a facilitatory role. Nineteen participants took part in-person, while one joined online and took part only in the brainstorming sessions. The practitioners comprised all participants of the project's management board meeting including managers, as well as contributors, who were at the time of the workshop actively engaged in the development of AI tools and technologies for either the cockpit, ATC or airport domains. The background of participants varied among AI assurance experts, aviation and system engineers, science dissemination experts, as well as human factors specialists. The various background of participants was valued for the different perspectives it could bring. Relevant was, however, that all practitioners are actively

engaged in the development and communication within teams working on AI solutions, so that they could report on struggles they have observed.

5 Procedure

At the beginning of the workshop, a presentation was given to create a mutual understanding of the human-AI teaming concept. Relevant information on the topic, such as the differences between classical user interface design, and the design of less predictable systems such as AI solutions was presented. Moreover, crucial questions regarding human-AI teaming, such as what makes AI a true team partner, how can an AI solution and a human operator cooperate and collaborate in reality, and should AI mimic human behaviour, or whether other options are more suitable, were addressed. In relation to these questions, the EASA taxonomy of AI levels (see Table 1) was explained, such that practitioners have a better understanding of the classification of their AI solutions.

Table 1. EASA's classification of applications' AI levels.

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

As a conclusion to the presentation, practitioners were reminded that within the JARVIS project a dedicated group of human-AI teaming experts is available to support them in case questions, challenges or the need for an external advice on the topic arise.

Following the presentation, the first silent brainstorming session was introduced. Practitioners were invited to, from their own perspective, list human-AI teaming concerns, particularly in relation to the interaction with the human operator, related to the technology they are developing within the JARVIS project. Participants were given three minutes to individually write down as many ideas as possible, with a minimum requirement of three topics listed. The question used to prompt participants' ideation was "*What are your solution's major questions, unclarities and concerns regarding the teaming between your system and the human operator currently*".

Immediately after the first brainstorming session, participants were asked to spend three more minutes individually listing challenges that their team has reported to them. The question used in this session was "*Write a minimum of three challenges regarding the teaming between your system and the human operator that your team has reported to you"*. Due to time constraints, the results of the brainstorming sessions were not discussed during the workshop. However, practitioners' ideas from the brainstorming

sessions could serve them as a basis for the next phase of the workshop, where they could be further discussed and refined. In the next phase of the workshop, a so-called "world café discussion" was organised, during which the potential challenges of the separate solutions were discussed in teams. Participants were divided into three groups with two groups including seven participants and one group consisting of six participants. Within the described group size, each group was assigned a facilitator, who was one of the human factors experts, facilitating the workshop. The facilitators were tasked with making sure the discussions stayed productive and focused.

Each group was assigned one of the AI solutions being built by the practitioners and discussed this solution's challenges. Group one discussed challenges that the cockpit digital assistant could face, while groups two and three focused on respectively the ATC and airport solutions. The participants were grouped in such a way that each group excluded participants working on the specific solution the group was assigned. This approach was selected with the goal of avoiding the influence of potential hierarchical dynamics in order to gather different perspectives and new ideas on the topic. To encourage creativity and motivate participants, teams were introduced with the competition of gathering as many challenges as possible with the team that accumulated the most challenges winning the competition. For this purpose, one of the participants in each team was assigned a role of "Challenge Collector" and was tasked with noting down all challenges mentioned in the discussion.

The "world café" discussions consisted of one round and lasted ten minutes. After completing the round all practitioners gathered to discuss the results collectively. Each group's Challenge Collector presented their findings, followed by a short discussion with all participants commenting on the gathered challenges.

6 Results

With the aim of identifying topics relevant for human-AI teaming practitioners, the workshop outcomes underwent a two-phase analysis. The first phase focused on identifying the topics present in the materials, aiming to capture them at a detailed level, while documenting the different nuances. In the second phase the topics were refined, rephrased, and where necessary, merged.

The categorisation was carried out in multiple steps by two human factors experts individually, who agreed to aim for precise and narrowed topics, rather than generalizations. Moreover, the decision was made to assign a single topic per note, with the consideration to select two topics, only when necessary.

Within the first step, nineteen categories were derived. For this coding process, Krippendorff's alpha was calculated as a measure to quantify the inter-rater-reliability (IRR). As the initial value of 0.59 indicated a rather poor agreement, the categories were reworked within a discussion between the two human factors experts. As part of the discussion, definitions for the different topics were formulated, in order to standardise the topics and ensure a mutual understanding between the human factors specialists of what the topics describe, and which practitioners' opinions belong to them.

The nineteen categories identified in the first round were found to be rather high, compared to the number of statements given by the practitioners ($N = 103$), thus we aimed at reducing the number of topics. On average, each category was mentioned 4.26 times in the world café sessions, so all categories with occurrences less than 4 ($n = 9$) were revised. We decided to keep the topics SA and Task-Share, as their definition from a human-factors point of view is precise. We subsumed the other categories and derived 13 categories and calculated IRR again. Krippendorff's alpha was 0.79 for these codings, which is regarded as a moderate agreement. To avoid categories which are too broad, we decided to stop the optimization of the coding scheme here and use these thirteen categories for our further analyses and mappings.

Table 2 shows the final list of topics with their descriptions and the number of times they were expressed in the workshop.

Table 2Table 2. Descriptions and number of mentions of the final set of topics with an acceptable inter-rater-reliability with a Krippendorff's alpha of 0.79.

Topic	Mentions	Description
Human-Centred Design	15	Numerous choices may be made when designing how the human and the machine (the AI) interact with each other. Information may, for example, be shared via audio, or via visual displays, trends can be expressed as dynamically developing graphs or, when a certain limit is reached, a dedicated indication for that may be given either visually or auditory. Important is to find a way to design the system such that it is optimised for an operator's task execution.
Trust	11	Refers to the extent to which human operators can rely on the performance of an AI team member. There are a number of reasons why human operators can, or cannot, grow to trust an AI system in their collaboration.
Interaction	10	May be seen as a part of human-AI teaming. While human-AI teaming describes the broader relationship between human, AI and how each contributes to the team, human-machine *interaction* is more about the direct exchange of information between human and AI.
Human Capabilities	9	This refers to the user's capability to interact with the AI-system. The AI should be designed in such a way that it accommodates human cognitive abilities, ensuring that humans can effectively interact with and understand the AI. This also includes training the human operator to interact with the AI.
Measurement	8	Developing effective methods to measure and assess the performance, efficiency, and effectiveness of

		human-AI collaboration, including metrics for task delegation, decision-making, communication, and overall system outcomes.
Legal Aspects	8	The legal aspects considered within this category, include the General Data Protection Regulation (GDPR), which is the regulation about how (personal) data is used. It is about making clear for what purposes, for how long, and what kind of data are being recorded and stored. Further contents categorized under this topic, include considerations regarding responsibility**.** Responsibility for the safe operations in aviation is currently carried by human operators. However, when systems get more advanced and more autonomous the responsibility might no longer be solely carried by human operators.
AI Level	8	Classification of the autonomy level of an AI function. For this, we refer to the EASA categorisation. In Level 1, the AI *assists* the human, in Level 2, we speak of *cooperation* (2A) or *collaboration* (2B) between human and AI, and in level 3 of advanced automation, where the automation also autonomously *takes decision and initiates actions*, that either can (3A) or cannot (3B) be overruled by the human.
Explainability	8	Ensuring users can understand the reasoning behind actions or decisions made by the system.
Technical Capabilities	6	This concerns the technical capabilities of the AI, enabling it to learn, reason, and interact with humans. AI can have many technical capabilities, however, in this paper the following are relevant: **Services**: This refers to the various functionalities provided by the AI system. **Data availability**: High-quality, relevant, non-biased, and sufficient data is necessary to train, validate, and deploy an accurate and reliable AI model. For example, if an AI tool is implemented that can help pilots divert to another airport in case of an emergency, it is necessary for the AI to have all relevant data on all nearby airports. **Data interoperability**: Enabling the integration of data from various sources, formats and systems. **Adaptivity**: This concerns a specific type of AI that can automatically learn and adapt to new situations, data, or environments. **Feedback driven learning**: The ability of the system to learn from the feedback provided by the operator.

Error Manage-ment	5	Refers to the countermeasures preventing error or to make the human-AI team error tolerant. The errors mentioned here pertain to the failure of the AI but also human errors. Typically, certain tasks are performed in teams in order to facilitate error detection, prevention and tolerance. The topic of error management is closely related to (over)reliance.
Task-share	4	Refers to the allocation of tasks between human and AI. With both the task allocation between human and AI by system design, as well as any potential dynamic task re-allocation being included here.
Situation Awareness	4	Situation awareness is a concept that is, amongst others, defined by Endsley [9]. Endsley's model discriminates between three different levels of understanding: perception, comprehension, projection into the future.
Teams	4	There are two different kinds of teams that the current study discriminates between: the smaller team of one human and one AI agent and the bigger teams where different humans and possibly different AI agents have to collaborate. Topics related to the small team are more focused on the interaction between human and AI. In bigger teams the roles and responsibilities between the different team members need to be crystal clear to make sure that no unexpected misunderstandings between members will lead to poor performance.

The topics were mapped to visualize their interdependencies as illustrated in Figure 1.

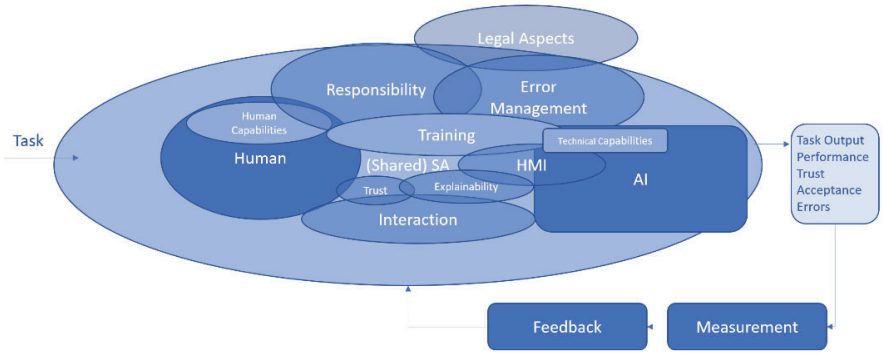

Figure 1. Identified topics and their interdependencies.

As part of our analysis, the identified topics were compared to the categorization of human-AI teaming research needs proposed by National Academies of Sciences, Engineering, and Medicine [5]. The comparison consisted of a six-step process. Within the first step, an exact match between the topics proposed here and the topics as described by National Academies of Sciences, Engineering, and Medicine [5] was searched for using our topics as a search keyword. In the second step, the same approach was applied to topics for which a match had not yet been found but rather by going through the research objectives contained in each category proposed by National Academies of Sciences, Engineering, and Medicine [5]. In the third step, rather than searching for a one-to-one accordance between the wording used by us and those used by National Academies of Sciences, Engineering, and Medicine, the topics identified in state-of-the-art report were read through, interpreted, and compared to our topics. Similarly, in a fourth step the same approach of interpretation was applied to the objectives proposed by the National Academies of Sciences, Engineering, and Medicine [5]. Ultimately, wherever no matches between our topics and those proposed by National Academies of Sciences, Engineering, and Medicine [5] could be found, the descriptions of the objectives were searched through using our topics as keywords. Finally, if no matches were yet found, the objectives' descriptions were interpreted for potential overlaps. Table 3 outlines the six-step process used to match our identified topics to the categorization of human-AI teaming research needs proposed by National Academies of Sciences, Engineering, and Medicine [5], with each row representing a step in the process. The columns Step, Approach, and Matched topics provide an overview of the method used in each step, such as keyword searching or interpretation, and the topics that were successfully matched to National Academies of Sciences, Engineering, and Medicine's [5] categorization at each stage.

Table 3. Matched topics for each of the matching steps

Step	Approach	Matched topic(s)
1	Keyword comparison (topic – topic)	Trust, interaction, explainability
2	Keyword comparison (topic – objective)	Task-share, situation awareness
3	Interpretative comparison (topic – topic)	Human-Centred Design, Measurement
4	Interpretative comparison (topic – objective)	AI Level, Teams
5	Keyword comparison (topic – objective description)	*None*
6	Interpretative comparison (topic – objective description)	Human Capabilities, GDPR, Technical Capabilities, Error Management

7 Discussion

This section examines the challenges encountered by practitioners in designing human-AI teaming systems and compares them to the current state-of-the-art research in

this field. This paper investigates how the challenges faced by practitioners align with existing research gaps, to identify what research is needed to support those designing these AI applications.

Human-centred design was mentioned most often during the workshops. Practitioners seemed to struggle with how to design a system tailored to the human user, especially in a teaming scenario. The main challenge faced by practitioners related to the interface of the system, how to present information to the user such that they benefit from it but in a way that does not become overwhelming. In National Academies of Sciences, Engineering, and Medicine [5] the human-centred design is also quite important but the focus lies less on the interface and more on human-system integration (HSI) in general. In HSI it is common practice to consider the context of use when designing an AI system. National Academies of Sciences, Engineering, and Medicine [5] highlights the importance of considering the context when designing AI systems and states that different disciplines can contribute different insights within the development of HMI designs for human-AI teaming solutions. Teams comprising of system engineers, computer and data scientists, social technical researchers and human factors engineers should collaborate in order to come to an AI system that is designed taking the human operator in mind.

Trust is a topic that is high on the research agenda of National Academies of Sciences, Engineering, and Medicine [5] and was repeatedly brought up in the workshop. Practitioners were wondering how trust can be instilled and how to repair it once damaged as well as how it can be measured in the first place. National Academies of Sciences, Engineering, and Medicine [5] agrees that trust (repair) is an important fundament for an effective team performance stating that trust in a system needs to develop over time. When a system's outputs are consistently reliable, the trust of the human in the system increases but when a system gives an incorrect output trust is suddenly decreased with automation failure having a bigger impact on trust than automation success. National Academies of Sciences, Engineering, and Medicine [5] states that the way trust has been researched and measured needs to change by focussing more on how the human operator and the system interact rather than the operator's reliance on the system.

While human-centred design is very much focussed on the design of the interface, the **interaction** between human and AI may comprises more than just the systems/the hardware. For practitioners, clear guidelines about how dynamic the system should be, and how it should behave exactly would be useful for the concrete design. However, National Academies of Sciences, Engineering, and Medicine [5] indicate that no clear guidelines exist on that yet.

When it came to the subject of **human capabilities**, the practitioners concerned themselves with the capability of humans to interact with the AI system and the need for the AI to be designed taking into account human needs, abilities and limitations. At the same time humans will need to develop new skills and knowledge to effectively work with AI systems, they will need to be trained for new roles and new operating systems and environments. National Academies of Sciences, Engineering, and Medicine [5] state that there is a great deal of knowledge on training human-human teams

but very little on training human-AI teams, meaning research on this topic is highly necessary.

In relation to the topic **measurement**, the practitioners needed more clarity on how the collaboration between human and AI should be tested and validated. National Academies of Sciences, Engineering, and Medicine [5] confirms that there is a need to develop and test methods for analysis, design and evaluation of human-AI team performance.

In relation to **legal aspects** (like GDPR), National Academies of Sciences, Engineering, and Medicine [5] occasionally discuss the need for clarity on how human operator data are being shared, used and stored. Maybe it is due to the fact that National Academies of Sciences, Engineering, and Medicine [5] are more focussed on the scientific approach that this topic did not lead to a substantial chapter or section. However, for building a human-AI teaming system, the collection and application of operator data is unavoidable. As such, guidelines on how that should be done are key. The quality of the interaction between human and AI may be much better if the AI can (pretend to) understand all of the human's behaviour, ranging from inputs made into the system up to frowning or utterances. Guidelines about what human operator data to monitor, use and store, and under what conditions that is or is not allowed, would help the system designers, and prevent them from building features that afterwards will turn out not to be allowed or legal.

Regarding the **AI level**, practitioners asked themselves whether the AI levels as proposed would sufficiently describe the level of teaming, whether Level 2A "collaboration" and 2B "cooperation" are enough to describe these complex matters. Other comments categorised under the topic of AI level dealt with challenges regarding the teaming aspect. Questions along the line of: "*Do we actually want teaming or 'just a tool' that is helpful?*", were asked. The practitioners appeared to, at this moment, prioritize other objectives over establishing a collaborative teaming environment, i.e. a higher level of automation. To them, it appeared that some intermediate steps were necessary or logical to take, that lie in between the current operational concept and an artificial agent that could function as a teammate. National Academies of Sciences, Engineering, and Medicine [5] showed that the effects of higher levels of automation have been researched but indicated that there is a lack of research on working with AI, especially when it comes to maintaining or regaining SA.

Regarding **explainability**, the practitioners seemed to understand the need for the human operator to correctly interpret the recommendations or solutions given by the AI and that the operator will not be comfortable accepting a solution that is difficult to interpret. They feel it is important for the human to be aware and have a mental model of the AI's capabilities and intentions. While only explainability was explicitly mentioned by the practitioners, the latter statement also closely relates to "transparency". The term "transparency" only came up once during the workshop, where it was stated to focus on making operators trust the AI system rather than focusing on transparency. However, transparency can help build trust and help the user to maintain proper SA [5], which practitioners indicated to be important, if done properly. The fact that transparency came up so little during the workshop could also have to do with the fact that transparency and explainability are so closely related and are often used interchangeably. Explainability is also a key mechanism for improving SA, trust and performance

in human-AI teams [5]. However, research is still needed into how to best provide explainability to the user and how to balance SA, trust and performance on the one side, and the attention needed to understand the explanations on the other side. To achieve this balance, it is necessary to research what aspects of the AI need to be made transparent for various types of tasks.

While the concerns regarding the **technical capabilities** of the AI system included the various functionalities that should be provided by the AI system and the idea that the AI might have to be adaptable, they mainly regarded data: How do we ensure that there is sufficient, accurate and relevant data to train the AI models and how do we verify this? How do we address potential bias in our data? Do we allow feedback driven learning? How do we create a robust infrastructure for data collection, storage and management and how do we standardise the data formatting so different systems and stakeholders can also use it? Further challenges involved the establishment of clear guidelines and regulations for the sharing and usage of this data. While very valid concerns, these aspects are more on the technical side and are not directly related to human-AI teaming and human factors. It does however show the struggles practitioners face, and these aspects are highly important to research, especially AI bias as it can cause issues with "fairness". An example of which was given by a practitioner stating that an ATM system that proposes sequences of traffic and structurally would, based on historic data, prioritise one airline over another, would be unfair.

In terms of **error management**, practitioners expressed their concerns regarding failures by the system in combination with the reliance on the system. What would be proper contingencies when the system is wrong or stops working? A relation was also made with "responsibility" i.e., who is responsible for the process outcome if the automation (suddenly) fails? National Academies of Sciences, Engineering, and Medicine [5] focusses more on researching how to detect and mitigate AI bias, as well as how to make the AI aware of its limitations rather than what to do when the AI fails.

Within the topic of **task share**, the practitioners were thinking about how the delegation of tasks between the human operator and the digital assistant could be arranged in a smooth way. The comments made all related to dynamic task re-allocation. According to practitioners it should be quite effortless to delegate tasks, otherwise it does not make sense. The operator could delegate actively, using simple commands or a selection of a function. But the AI can also, e.g. based on data regarding the workload of the operator, decide to offer help that the human could either accept or decline. National Academies of Sciences, Engineering, and Medicine [5] confirms that research is still needed to determine improved methods to support collaboration and to focus more on understanding the ways that people and AI systems can share tasks.

Situation awareness (SA) is a broad construct. For practitioners it comes primarily down to: Do operators know what they need to know in order to execute their tasks, or do their jobs? In the work of National Academies of Sciences, Engineering, and Medicine [5] a first attempt is done to describe how SA may be influenced. National Academies of Sciences, Engineering, and Medicine [5] mention: display interface, the automation-interaction paradigm, the mental model, and trust for developing high levels of SA in demanding and dynamic environments [22]. In fact, SA is a good example of where National Academies of Sciences, Engineering, and Medicine [5] offer information that is at the level that practitioners can use to maximise operator SA. Important

to realise is that the SA interpretation and guidelines of the academies are not solely relevant for SA in relation to higher levels of AI, they are relevant in many more automation related settings.

Practitioners stated that the applications they are working on now are not yet to the point of AI being a member of the **team** but rather "*a clever tool delivering information*". National Academies of Sciences, Engineering, and Medicine [5] also state that human-AI *teaming* is a step beyond human-AI *interaction*, a sentiment shared by the practitioners who feel that there are several steps between the AI tools that they are currently designing and an AI system that functions as a "teammate". This explains why topics such as "team effectiveness" and "team processes" were not mentioned during the workshop whereas they were deemed important by National Academies of Sciences, Engineering, and Medicine [5]. The teaming aspect seemed less important to the practitioners at this stage as they feel it is necessary to first develop and test the in-between steps before going to actual human-AI *teaming*. One practitioner even questioned if the "teaming aspect" is even necessary by stating: "*Do we want an AI teammate or do we want a smart system?*". Taking into account drawbacks that come with "traditional" human-machine interaction and the criticality of the tasks and tooling in aviation, a more conservative approach towards more complex and autonomous systems that work as a team member seems justifiable.

It seems that many concerns that arise in developing AI functions are still very similar to the development of traditional human-machine interaction. This is in line with the idea that the first steps in building higher levels of automation requires the initial development of an automated function that the human can interact with, before it becomes possible to expand the function by feeding it with data it can learn from and start behaving as an intelligent system.

8 Conclusion

The goal of this study was to provide an overview of topics and challenges in designing higher levels of automation in aviation, with a focus on bridging the gap between theory and practice through a comparison with National Academies of Sciences, Engineering, and Medicine's guidance on human-AI teaming. The current study achieves this goal by highlighting key challenges and areas of agreement with existing research, however, but has some limitations. The categorization of practitioner inputs and the use of a singular guidance document, although comprehensive, could introduce bias. Nonetheless, this research contributes significantly to the body of knowledge on human-AI teaming by identifying key challenges, comparing practitioner insights with existing research gaps, and providing practitioner-oriented research insights into human-AI teaming in aviation. The study can also be seen as a guideline regarding the subtopics of human-AI teaming where more research is needed. Overall, this study highlights the need for continued research and collaboration between practitioners, researchers, and policy-makers to address the complex challenges associated with human-AI teaming and to develop effective guidelines and methods for designing and evaluating human-AI teaming systems.

Acknowledgments. JARVIS has received funding from the SESAR Joint Undertaking under the European Union's Horizon Europe research and innovation programme under grant agreement No 101114692. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or SESAR 3 Joint Undertaking. Neither the European Union nor SESAR 3 Joint Undertaking can be held responsible for them.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] A. Joshi, S. P. Miller und M. P. Heimdahl, „Mode Confusion Analysis of a Flight Guidance System Using Formal Methods," in *Digital Avionics Systems Conference (DASC'03)*, Indianapolis, USA, 2003.

[2] D. Kelly und M. Efthymiou, „An analysis of human factors in fifty controlled flight into terrain aviation accidents from 2007 to 2017," *Journal of Safety Research (69),* pp. 155-165, 2019.

[3] C. Collins, D. Dennehy, K. Conboy und P. Mikalef, „Artificial intelligence in information systems research: A systematic literature review and research agenda," *International Journal of Information Management (60),* 2021.

[4] M. Steinberg, „Toward System Theoretical Foundations for Human–Autonomy Teams," in *Systems Engineering and Artificial Intelligence*, Springer Cham, 2021, pp. 77-92.

[5] National Academies of Sciences, Engineering, and Medicine, Human-AI Teaming: State-of-the-Art and Research Needs, Washington, DC: The National Academies Press., 2022.

[6] A. Jacovi, A. Marasović, T. Miller und Y. Goldberg, „Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI," in *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 2021.

[7] European Commission - Directorate-General for Communications Networks, Content and Technology, „Ethics guidelines for trustworthy AI," Publications Office, 2019.

[8] R. Molich und J. Nielsen, „Molich and Nielsen's heuristics (1990)," 1990. [Online]. Available: https://www.dialogdesign.dk/molich-and-nielsens-heuristics-1990/. [Zugriff am 2025].

[9] M. R. Endsley, „Automation and situation awareness," in *Automation and Human Performance*, Mahwah, NJ, Lawrence Erlbaum Associates, Inc., 1996, pp. 163-181.

[10] R. Parasuraman und D. H. Manzey, „Complacency and Bias in Human Use of Automation: An Attentional Integration," *Human Factors,* pp. 381-410, 2010.

[11] M. R. Endlsey und W. M. Jones, „A Model of Inter and Intra-Team Situation Awareness: Implications for Design, Training and Measurement," *New Trends in Cooperative Activities: Understanding System Dynamics in Complex Environments,* 2001.

[12] N. A. Stanton, R. Stewart, D. Harris, R. J. Houghton, C. Baber, R. McMaster, P. Salmon, G. Hoyle, G. Walker und M. S. Young, „Distributed Situation Awareness in Dynamic Systems: Theoretical Development and Application of an Ergonomics Methodology," *Ergonomics,* Bd. 49, Nr. 12-13, pp. 1288-1311, 2006.

[13] J. Jiang, A. J. Karran, C. K. Coursaris, P. Léger und J. Beringer, „A Situation Awareness Perspective on Human-AI Interaction: Tensions and Opportunities," *International Journal of Human-Computer Interaction,* Bd. 39, Nr. 9, pp. 1789-1806, 2023.

[14] E. Kaltenbach und I. Dolgov, „On the Dual Nature of Transparency and Reliability: Rethinking Factors that Shape Trust in Automation," in *Proceedings of the Human Factors and Ergonomics Society 2017 Annual Meeting*, 2017.

[15] J. Y. C. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia und M. Barnes, „Situation Awareness-Based Agent Transparency," Aberdeen Proving U.S. Army Research Laboratory, Ground, MD, 2014.

[16] A. Bhaskara, L. Duong, J. Brooks, R. Li, R. McInerney, M. Skinner, H. Pongracic und S. Loft, „Effect of Automation Transparency in the Management of Multiple Unmanned Vehicles," *Applied Ergonomics,* Bd. 90, 2021.

[17] S. Loft, A. Bhaskara, B. A. Lock, M. Skinner, J. Brooks, R. Li und J. Bell, „The Impact of Transparency and Decision Risk on Human-Automation Teaming Outcomes," *Human Factors,* pp. 846-861, 2023.

[18] A. Rosenfeld und A. Richardson, „Explainability in Human-Agent Systems," *Autonomous Agents and Multi-Agent Systems,* Bd. 33, pp. 673-705, 2019.

[19] A. Silva, M. Schrum, E. Hedlund-Botti, N. Gopalan und M. Gombolay, „Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction," 2022.

[20] T. Lombrozo, „Simplicity and probability in causal explanations," *Cognitive Psychology,* Bd. 10, Nr. 10, pp. 232-257, 2007.

[21] H. Schiele, S. Krummaker, P. Hoffmann und R. Kowalski, „The "research world café" as method of scientific enquiry: Combining rigor with relevance and speed," *Journal of Business Research,* pp. 280-296, 2022.

[22] M. R. Endlsey, „From here to autonomy: Lessons learned from human-automation research.," *Human Factors 59(1),* p. 5–27, 2017.

[23] E. Salas, T. Dickinson, S. Converse und S. Tannenbaum, „Toward An Understanding of Team Performance and," in *Teams: Their Training and Performance*, Norwood, NJ, Ablex, 1992, pp. 3-29.

[24] J. Cannon-Bowers, E. Salas und S. Converse, „Shared Mental Models in Expert Team Decision Making," in *Current Issues in Individual and Group Decision Making*, Hillsdale, NJ, Lawrence Erlbaum, 1993, p. 221–246.

[25] M. R. Endsley, „Toward a theory of situation awareness in dynamic systems," *Human Factors, 37(1),* pp. 32 - 64, 1995.