future transportation

MDPI

Article

The Impact of Artificial Intelligence on Future Aviation Safety Culture

Barry Kirwan

EUROCONTROL, Centre du Bois des Bordes, 91222 Brétigny sur Orge CEDEX, France; barry.kirwan@eurocontrol.int

Abstract: Artificial intelligence is developing at a rapid pace, with examples of machine learning already being used in aviation to improve efficiency. In the coming decade, it is likely that intelligent assistants (IAs) will be deployed to assist aviation personnel in the cockpit, the air traffic control center, and in airports. This will be a game-changer and may herald the way forward for single-pilot operations and AI-based air traffic management. Yet in aviation there is a core underlying tenet that 'people create safety' and keep the skies and passengers safe, based on a robust industry-wide safety culture. Introducing IAs into aviation might therefore undermine aviation's hard-won track record in this area. Three experts in safety culture and human-AI teaming used a validated safety culture tool to explore the potential impacts of introducing IAs into aviation. The results suggest that there are indeed potential negative outcomes, but also possible safety affordances wherein AI could strengthen safety culture. Safeguards and mitigations are suggested for the key risk owners in aviation organizations, from CEOs to middle managers, to safety departments and frontline staff. Such safeguards will help ensure safety remains a priority across the industry.

Keywords: aviation; artificial intelligence; safety culture

Citation: Kirwan, B. The Impact of Artificial Intelligence on Future Aviation Safety Culture. *Future Transp.* **2024**, *4*, 349–379. https://doi.org/10.3390/futuretransp4020018

Academic Editor: Lynnette Dray

Received: 15 January 2024
Revised: 2 April 2024
Accepted: 3 April 2024
Published: 9 April 2024

1. Introduction

1.1. Overview of Paper

Currently, aviation is seen as a very safe mode of transport, and this is in part due to its safety culture. The question raised in this paper is about the potential impact of Artificial Intelligence (AI) on aviation safety culture. Although machine learning has already integrated AI into various aviation sectors, this paper specifically examines the prospects of more advanced AI systems. These systems may include intelligent assistants that have the potential to function semi-autonomously, or even autonomously, in collaboration with human crews and teams.

The paper begins by briefly outlining safety culture in aviation today, including how it is evaluated. The fast-developing area of AI itself is then outlined, focusing on different 'levels' of AI autonomy and the concept of human-AI teaming. This wide-ranging exploration of AI is necessary to envision how human crews and Intelligent Assistants (IAs) might work together in a range of future AI settings (e.g., cockpit, air traffic tower and operations room, airports). The application of an aviation safety culture method is then analyzed in relation to future human-AI teaming scenarios to assess potential safety culture outcomes. The paper concludes by noting the most serious threats to safety culture posed by AI, and how to safeguard against them, as well as suggesting ways forward to harness the potential safety culture benefits from human-AI teaming.

1.2. Safety Culture—An Essential Ingredient of Aviation Safety

In European commercial civil aviation today, safety in terms of low accident rates is very strong, with no major accidents involving EU-registered aircraft in commercial air transport over the past seven years [1], although there have still been fatal air crashes

in general aviation. This level of safety can be attributed to several factors, including technological improvements, maintenance, safety management processes, improved teamworking and safety culture. The first four factors deal with how to keep the aircraft safe, whereas safety culture serves as the driving force for safety, whether in operations, maintenance, or design.

Safety culture as an approach did not emerge out of the blue [2]; prior to safety culture, there was already research on safety climate. Safety climate is a momentary reflection of the present safety culture, based on perceptions and emotions, similar to mood, while safety culture is more enduring, akin to personality, and linked to group activities and organizational histories [3]. As safety culture is more enduring, it is thus harder to change. Safety climate emphasizes managerial prioritization of safety [4], while safety culture focuses on safety-related values and practices within the organization [5]. An early definition of safety culture, from the nuclear power industry, is as follows [6]:

> *"The safety culture of an organization is the product of individual and group values, attitudes, perceptions, competencies, and patterns of behavior that determine the commitment to, and the status and proficiency of, an organization's health and safety management."*

A second, frequently used definition of safety culture and, more generally, organizational culture, is as follows [2]:

> *"Shared values (what is important) and beliefs (how things work) that interact with an organization's structures and control systems to produce behavioral norms (the way we do things around here)."*

The term safety culture was coined following the Chernobyl nuclear power plant accident in 1986 [7]. Chernobyl demonstrated that poor safety culture could overcome all the hardware and software defenses put in place to prevent a nuclear catastrophe. Accidents in space, oil and gas, and rail industries have further demonstrated the importance of this now commonly recognized organizational safety characteristic (see Figure 1). High profile public enquiries into key accidents, such as the Piper Alpha disaster [8] and Clapham Junction rail crash [9], as well as key safety thought leaders at the time [10,11], have continually underscored the importance of safety culture. Such factors, alongside a continuing number of accidents attributed to poor safety culture ever since, have ensured that safety culture remains a focal point in many safety departments across a diverse range of industries.

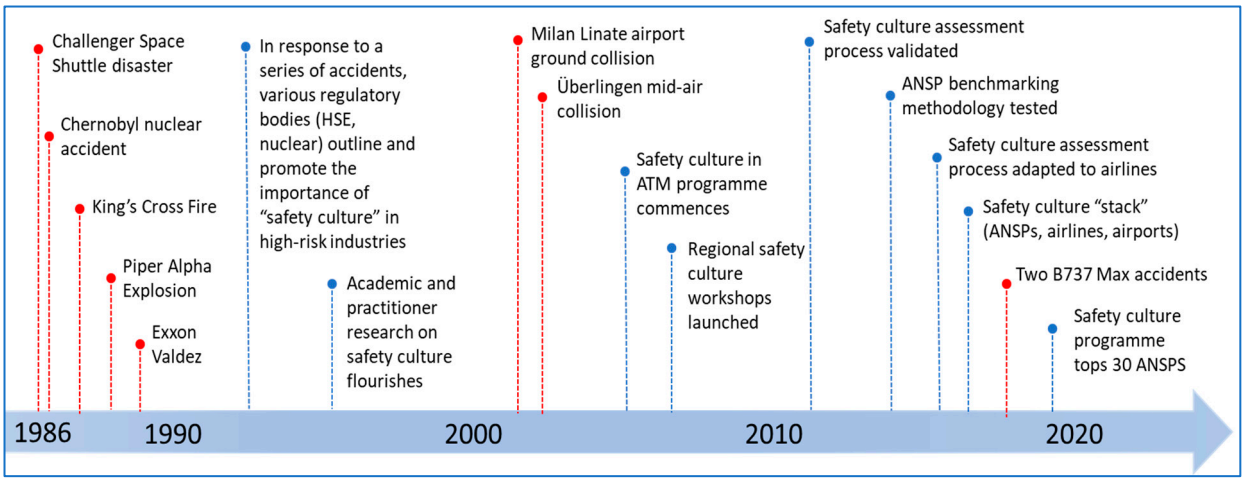

Figure 1. Timeline of events influencing European aviation safety culture.

Initially, safety culture was not considered a major concern in aviation, even after the Kegworth air crash in 1989 [12]. It was thought that measures such as strong training, effective cockpit and air traffic operations room design, robust safety management systems (SMSs) and processes, effective maintenance, and standard operating procedures (SOPs)

were sufficient. In 2002, a pivotal event occurred in Europe with the mid-air collision over Lake Constance in Überlingen [13], following on from the Milan Linate runway collision in the year prior [14]. These two accidents resulted in a profound shift in thinking for Air Navigation Service Providers (ANSPs), revealing the limitations of SMSs and SOPs and emphasizing the criticality of safety culture. While the SMS held all the safety competence including key safety processes, safety culture motivated those processes to achieve safe results.

1.3. The Emergence of a Safety Culture Evaluation Method in Aviation

Given that the two accidents (Uberlingen and Milan) were principally related to air traffic management (ATM), the development of a safety culture assurance method was carried out in that sector of the industry. This was achieved via a combined effort of EUROCONTROL and Aberdeen and LSE universities [5,15], who were able to build upon more than a decade of experience of safety culture evaluation in nuclear power and oil and gas industries. A safety culture questionnaire approach was therefore developed, its results informed by workshops with aviation staff. After an initial pilot test with four European air traffic providers, the approach was rolled out in 2005 and has gradually been applied across Europe. To date, 33 European member states have applied the approach, most of them more than once. The EUROCONTROL Safety Culture Questionnaire has been scientifically validated [5,16] and positively reviewed by the European ANSPs [17]. It has also had more modest applications in airlines and airports [18].

1.4. Measuring Safety Culture

The EUROCONTROL Safety Culture Questionnairecontains 48 questions linked to eight safety culture 'dimensions':

- Management commitment to safety
- Collaboration and involvement
- Just culture and reporting
- Communication and learning
- Colleague commitment to safety
- Risk handling
- Staff and equipment
- Procedures and training

An example of the style of output from such a survey is shown in Figure 2 for three typical safety culture questionnaire items. In this example, the first statement clearly reflects 'positive' safety culture, with a few who are neutral about the issue and a small percentage of dissenters. The second item has a large 'neutral' component; this can mean either that the respondents are not sure, or they do not see how it applies to their work environment, or they prefer not to say. Although the surveys are always anonymous and confidential, some participants remain cautious. The third item has a significant negative component that would be investigated further via confidential workshops with participants. This could include flight and cabin crew, controllers and engineers, airport workers, management, and support staff. The aim of such workshops is to find out what has been happening, and how to establish a better safety culture (or in this specific case, a better just culture). Such workshops are often very useful in seeing 'beneath' the questionnaire results and are helpful in determining practical ways forward.

It is useful to provide a high-level overview when carrying out safety culture surveys, and this is achieved by summarizing results at the 'dimension' level using a spider chart, as shown in Figure 3. The higher the values (the further out from the center), the better the safety culture is perceived to be by the participants. Figure 3 shows the results for five separate organizations at a single airport, with each company represented by a unique color. This is a more recent variant of safety culture evaluation and has the advantage of highlighting where some companies can help others, and seeing if there are any 'best in class' performers and, if so, how their safety approach differs from other companies. This

multiple-party safety culture survey approach is known as the Safety Culture Stack [18]. It is particularly recommended for airports, which have many business partners who rely on each other for collective airport safety.

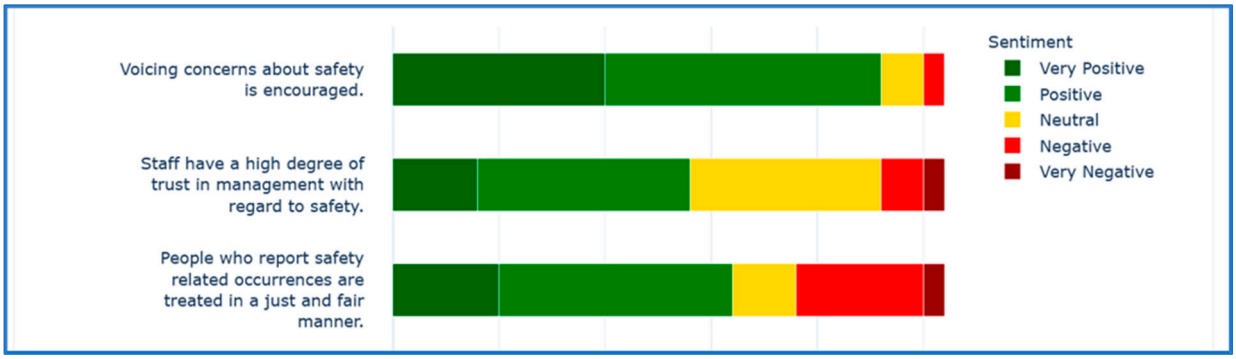

Figure 2. Example summary of responses to a safety culture survey for three questionnaire items.

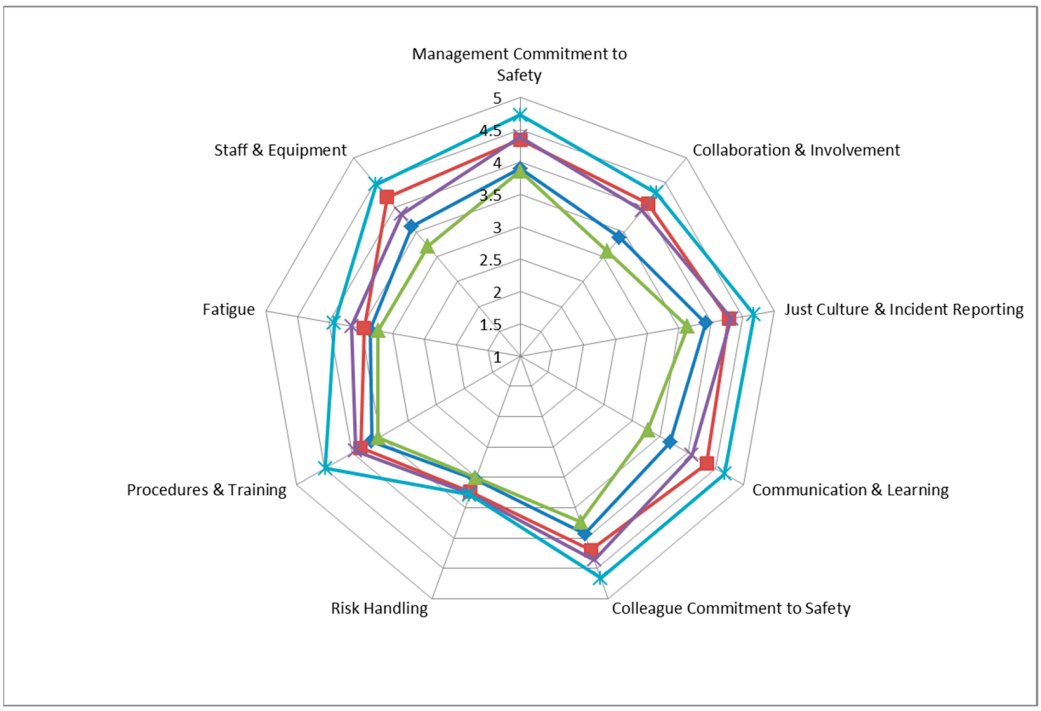

Figure 3. Example of 'spider chart' view of safety culture survey results (Fatigue appears in this diagram, as it is sometimes added to the other dimensions because of its importance as a factor in aviation, though it is not strictly speaking a safety culture dimension and is not used in the ATM-only version, nor in this study with human-AI teaming) (Airport).

Such diagrams give a 'helicopter view' of the survey results and are often appreciated by senior management as they show the safety culture survey 'headlines'.

Safety culture surveys offer the safety culture equivalent of a detailed health check, showing where the organization is healthy and where it needs attention. Such reports include recommendations on how to improve, with the best ideas often arising from the confidential workshops with participants, and many organizations have used such surveys to improve their safety culture [17]. CEOs often find the results of such surveys useful [19] as they form a bridge between them and the front-line employees, so they can see in a relatively unfiltered way what people at the sharp end are concerned about.

The concept of safety culture includes senior and middle management and considers the critical importance of management and designers regarding safety. As the investigation

into the two Boeing 737 Max accidents has shown [20,21], even with the best engineering and a strong track record in safety performance, a compromised safety culture can lead to disaster. Senior managers (CEOs, VPs, Directors) make executive decisions that ripple down through their organisations and can dramatically affect safety culture—they 'set the tone' for the safety culture of their organisation.

1.5. Safety Culture and Future AI—An Unexplored Landscape

For this paper, it is the detailed safety culture questionnaire items, rather than the dimensions per se, that are likely to highlight where AI may affect safety culture. These are returned to following the next section, which explores the development of AI and likely future human-AI teaming scenarios in a range of aviation contexts.

In order to consider how AI might affect safety culture in aviation, it is necessary to see how AI might look in the cockpit, the air traffic operations room, or the airport control center in the coming decade. The next section accordingly builds a preliminary picture of future aviation AI by considering the following points:

1. The origins of AI;
2. AI today;
3. Generative AI;
4. Narrow AI;
5. Visions of future aviation human-AI systems;
6. Trustworthy AI;
7. Accountable AI;
8. AI and just culture;
9. Ethical AI;
10. Maintaining human agency for safety;
11. AI anthropomorphism and emotional AI;
12. AI and safety governance.

The consideration of these issues and perspectives helps narrow the expansive and ever-growing field of AI research to enable a realistic focus on safety culture impacts; in effect, it serves to 'ground' the later analysis in Section 4. It also lays the foundation for determining who should be the risk owners of each of the issues arising from the analysis, whether they are staff at the sharp end, middle managers, or senior management, as further elaborated in Section 5.

2. The Developing Artificial Intelligence Landscape in Aviation

2.1. The Origins of Artificial Intelligence

The simple idea behind artificial intelligence is to go beyond the limitations and capabilities of human thinking. An example of early AI is the 'Bombe' machine [22], used to break German 'enigma' codes used in the second world war. Such codes were unbreakable by humans, and so the 'Bombe' machine did indeed surpass our capabilities. But such machines were not seen as 'thinking'; rather they were running endless calculations or 'running the numbers', hence they were 'computing' rather than thinking. Alan Turing himself was fascinated by the idea of a machine that could one day think, and whose thinking would be indiscernible from that of a human. This led to his famous challenge to the scientific and engineering communities to develop such a machine, to be tested by 'the imitation game' [23]. He predicted such machines would exist by the turn of the century. What is interesting is that many of the questions he posed about artificial intelligence back in 1950 are the same questions we ask today.

In the 1980s, there was another surge in AI interest via (rule-based) expert systems, which ultimately failed to deliver operationally useful tools. This was in part due to their inability to account for the experience-based and highly contextual 'tacit knowledge' that human operators amass, which often far exceeds what is written in procedures. The failure of expert systems led to the so-called 'AI winter', which ended recently as computing power increased dramatically and machine learning finally became possible [24]. This has resulted

in a host of early AI prototypes, products, and services being introduced into European aviation, from automatic speech recognition and passenger support, to optimising safe and expeditious air traffic flow both in normal and hazardous weather [25].

2.2. AI Today

A useful definition of AI is as follows [26]:

> "*. . .the broad suite of technologies that can match or surpass human capabilities, particularly those involving cognition*".

The general aim of artificial intelligence (AI), therefore, is currently seen as supporting human intelligence, and society, by using data science techniques to analyse complex datasets to find new patterns or solutions to problems that are beyond our own intellectual capabilities. In aviation, AI could be used to 'optimise' aviation systems, for example, to help minimise fuel usage across air traffic route networks to reduce aviation's carbon footprint, or to assist flight crew in finding a solution during an emergency. Machine learning can analyse very large, complex and heterogeneous datasets in ways the human mind finds difficult or impossible (e.g., via n-dimensional analysis). So far, such AI tools, though yielding impressive results, do not constitute thinking; they are still computing machines that are 'running the numbers', albeit in very complex and often unfathomable ways. Such AI tools can be seen as 'just more automation' [27], and their impact on safety culture might therefore be expected to be minimal. However, this understanding of AI, as effectively a more powerful automation tool support, shifted dramatically with the release of ChatGPT in 2022 [28], heralding generative AI.

2.3. Generative AI

ChatGPT is effectively a large language model (LLM), using the entire internet as its database, which sits behind a 'chatbot'. This chatbot enables a human user to have a 'conversation' (via the keyboard) on a vast range of issues. It is reminiscent of Turing's 'imitation game' challenge to develop thinking machines. Unlike systems before it, ChatGPT is 'generative' in that it can answer any question. This is because it has used an approach called supervised learning, wherein humans work with the AI to refine the answers it gives, essentially training the AI to give more plausible answers. Whether the answer makes sense is up to the user, and sometimes it produces answers that are inaccurate, or so bizarre they are referred to as hallucinations. It works best when being asked to draw together factual information already in the web, though much of the internet is not fact-checked. This dataset, while truly vast, limits LLMs such as ChatGPT for strict operational usage, e.g., how to land an aircraft in a particular configuration and weather pattern at a specific airport, because such information is not necessarily on the internet. But it has piqued the interest and imagination of millions, as it can write essays for students, compose music, generate business ideas, translate text and produce summaries, etc, and so the notion of generative AI is very much in vogue today. Although an LLM is probably not yet the solution for operational AI in aviation, that there can be generative and realistic dialogue between human and AI may pave the way for 'intelligent assistants' in the cockpit, air traffic tower or ops room.

One important note on models like ChatGPT is that sometimes people think they are interacting with an intelligent entity, when they are not. ChatGPT, despite its impressive outputs, is still 'running the numbers'. When it writes a sonnet, it neither thinks 'I am writing a sonnet', or 'I have written a sonnet', nor reflects on what it has done with any feeling whatsoever, such as pride or disappointment, and both these aspects of reflection and feeling have been hallmarked as requirements of a true thinking machine [23,24].

2.4. Narrow AI

Cognition categories found in AI are typically learning, perception, reasoning, communication, and knowledge representation. Common AI applications include expert systems, machine learning, robotics, natural language processing, machine vision, and speech recog-

nition [29]. However, the same authors note that getting AI applications beyond the end of research and into operational use suffers from the 'valley of death' phenomenon (i.e., many good ideas and prototypes never see industrial usage). This can be for a range of reasons but principally, three stand out:

- A lack of data (most AI systems have vast data appetites. This can also be seen as a scalability issue when moving from research to wider industry). Even though aviation has a lot of data, much of it is not shared for commercial competition reasons, and using 'proxy' or synthetically generated data risks diluting and distorting real operational experience;
- Business leaders lack an even basic understanding of the data science and the technological skills necessary to sustain operational applications of AI;
- A failure to develop the 'social capital' required to foster such a change, leading to users rejecting the AI tool's implementation (for example, because it threatens job losses).

Rather than generative AI, what aviation at least initially requires is narrow AI [30]. Narrow AI can solve specific problems in a domain but cannot generalize as broadly as humans can. Such systems (sometimes called *idiot savants*) can be superhuman at some tasks, and subhuman at others. The advantage of narrow AI is that it can focus on a specific domain or even sub-domain for which there is sufficient data for the AI to work, e.g., hundreds of thousands of aircraft approaches by various aircraft to a particular airport runway. An AI tool can then answer specific questions or find solutions to problems, or simply show how to optimize system performance based on a limited set of parameters for which there are plentiful data. In a field such as aviation, narrow AI is likely to be more fruitful in the short to medium term (i.e., for the next decade).

2.5. Visions of Future Human-AI Teaming Concepts in Aviation

Next, it is useful to consider contemporary visions of future AI concepts, some of which go beyond today's machine learning tools, leading to humans collaborating and negotiating with advanced AI systems. This level of interaction and collaboration between humans and AI, often supposed to involve some kind of dialogue, has led to the term human-AI teaming. The European Union Aviation Safety Agency (EASA—the principal European aviation regulator) has usefully set out a vision of AI and its potential impacts upon aviation operations and practices. EASA's recent guidance on human-AI teaming (HAT) [31] comprises six categories:

1A—Machine learning support (existing today);
1B—Cognitive assistant (equivalent to advanced automation support);
2A—Cooperative agent, able to complete tasks as demanded by the operator;
2B—Collaborative agent—an autonomous agent that works with human colleagues, but which can take initiative and execute tasks, as well as being capable of negotiating with its human counterparts;
3A—AI executive agent—the AI is basically running the show, but there is human oversight, and the human can intervene (sometimes called management by exception);
3B—the AI is running everything, and the human cannot intervene.

To help make some of these categories more concrete, it is useful to consider the EU-funded HAIKU (human-AI knowledge and understanding for aviation safety) Project [32,33], which has six intelligent assistant (IA) aviation human-AI teaming use cases, outlined below:

1. UC1—a cockpit IA to help a single pilot recover from a sudden event that may induce 'startle response' and direct the pilot in terms of which instruments to focus on to resolve the emergency. This cognitive assistant is 1B in EASA's categorization, and the pilot remains in charge throughout;
2. UC2—a cockpit IA used to help flight crew re-route an aircraft to a new airport destination due to deteriorating weather or airport closure, for example. The IA must consider a large number of factors including category of aircraft, runway length,

remaining fuel available and distance to airport, connections possible for individual passenger given their ultimate destinations, etc. The flight crew remain in charge but communicate/negotiate with the AI to derive the optimal solution. This is category 2B;
3. UC3—an IA that monitors and coordinates urban air traffic (drones and sky-taxis). The AI is an executive agent with a human overseer and is handling most of the traffic, with the human intervening only when necessary. This is category 3A;
4. UC4—a digital assistant for remote tower operations, to ease the tower controller's workload by carrying out repetitive tasks. The human monitors the situation and will intervene if there is a deviation from normal (e.g., a go-around situation, or an aircraft that fails to vacate the runway). This is therefore category 2A;
5. UC5—a digital assistant to help airport safety staff deal with difficult incident patterns that are hard to eradicate, using data science techniques to analyze large, heterogeneous datasets. At the moment, this is a retrospective analysis approach, though if effective it could be made to operate in real-time, warning of impending incident occurrence or hotspots. This is currently 1A/1B, but could evolve to become 2A;
6. UC6—a chatbot for passengers and airport staff to warn them in case of an outbreak of an airborne pathogen (such as COVID), telling passengers where to go in the airport to avoid contact with the pathogen. This is 1B.

Early studies on some of these concepts have already raised issues of interest. For use case 1, startle response, the idea is to counter the physiological and cognitive effects of startle as fast as possible, and then to help the pilot understand what is going on and what to do next. Although the startle effect is short-lived, typically 20 s, this is a major hazard if, for example, an aircraft is struck by lightning on final approach to an airport, resulting in instantaneous loss of key electrical systems.

The IA detects startle via a trained neural network that dynamically analyses a host of physiological parameters of the pilot, such as breathing rate, heart rate, skin conductance, etc., to determine whether startle has occurred. Once startle is detected, certain areas of the cockpit dashboard turn green, then slowly fade, then turn green again, then fade, etc. This ambient lighting change is to help the pilot re-establish a calmer breathing pattern (pilots would be pre-trained to breathe in synchrony with the lighting), and lasts for 20 s. Since it is ambient lighting, it does not distract them from any cockpit display or exterior view. Only once it is finished does the directed situation awareness engage. Color coding is used to direct the pilot's attention sequentially to the key instruments, telling the pilot what has happened, the aircraft's current status, and the key flight parameters needing immediate attention.

Preliminary studies with a small group of licensed commercial pilots suggest that the ambient lighting works, including in one case where the pilot did not believe it helped, yet their physiological parameters stabilized more quickly. The pilots found it useful, which may also be because this was a single-pilot scenario, so there was no co-pilot as back-up.

More work is ongoing on the IA's second and more cognitive function, namely the highlighted sequence of displays, as it is difficult to adapt to an individual pilot's speed of accessing and assimilating data. This raises another more general question with human-AI teaming, namely, how far to tailor tools to individual needs and preferences. The operators (i.e., airlines) may need to fine-tune future AI support systems to individual pilots, to ensure smooth and fluent performance in crisis situations.

The work so far does not suggest this system might adversely affect safety culture; the pilots are still very much in control and flying the plane and can switch off the AI at any moment. Rather, it offers a potentially welcome 'breathing space' in an event which, thankfully, most pilots never encounter.

In the tower controller cognitive assistant scenario, an early study with a group of controllers suggests it is perceived as most useful when very busy, or when the work becomes complex. This again raises a more general issue concerning intermittent AI support systems, which are turned on only when needed. In the event of an accident, the question will be whether the AI could have helped avert the accident had it been switched

on. This is particularly an issue for aviation, where safety systems are usually always 'on'. The idea of optional AI systems that can sometimes support safety does not sit well with the current regulatory or certification framework. But again, as with the startle use case, there was no sign this level of AI would degrade the tower controller's safety culture.

One study aiming for more extensive AI autonomy (2B) is attempting a co-design approach, and the pilots are currently more interested in categories 1B and 2A, believing these to be sufficient. This raises a more general potential design friction issue with advanced AI, as humans may be reluctant to cede too much autonomy to an AI, for fear of replacement or other issues. This is not a new phenomenon for aviation, as there has been occasional resistance to improved automation. For example, implementing electronic flight strips has led to less need for flight data assistants, yet has proven itself in terms of aviation system efficiency and effectiveness. Similarly, most commercial aircraft landings are executed by computer, not pilots. This potential reluctance against high-autonomy AI is 'one to watch', as it could affect trustworthiness and safety culture.

The airport safety data case study has offered some surprising results. Initially, it was hoped to use unsupervised learning to analyze the airport's data store, but there was insufficient data, including high 'cardinality' or variegation of particular data threads, and relatively low numbers of incidents, such that initially correlational patterns between a large set of potential contributory factors and incidents were not found. However, when the datasets were presented to the operational stakeholders at a meeting, within a short period of exploration and interrogation of the datasets, a pattern was identified. This was an instance where instead of machine intelligence, there was human–machine intelligence, since neither alone could detect the pattern. The datasets are currently presented in a series of interactive dashboards for operational users to interrogate. This use case seems to enhance human safety culture, as it offers new tools to explore difficult to eradicate incident patterns which otherwise become 'normalized'.

Other advanced AI concepts include [34] digital assistants to help air traffic control provide more efficient and environmentally friendly ('greener') routes, advanced warning in the cockpit of impending flight instability, and digital help for evidence-based training to enhance performance during adverse events.

In a recent EUROCONTROL-FAA (Federal Aviation Administration) debate on aviation human-AI teaming [35], a critical threshold which could challenge the human's 'agency' for safety appeared to be category 2B. At this level, humans and AI collaborate, and each can act independently to a certain extent. This differs from what we have today, and could affect safety culture if safety became the province of the AI, rather than the human. There are currently no AI systems in aviation that autonomously share tasks with humans, or can negotiate, make trade-offs, change priorities, or start and execute tasks under their own initiative. Category 3A could also affect safety culture, as the human may be too far 'outside the loop' to intervene effectively in time. The 'lesser' categories, including 1B and 2A, could impact safety culture, even positively, as they could augment the degree of control the human has over safety. Rather than degrading or eroding safety, AI could therefore possibly enhance safety, offering new safety affordances. This important aspect is returned to in Section 4.

2.6. The Need for Trustworthy AI in Safety-Critical Systems

AI tools will not be used if they are not trusted. A recent model of trustworthy AI [36] comprises eight technical requirements, built on three pillars throughout the entire system life cycle. The three essential pillars are that the AI and its operation are lawful, ethical, and robust. Since law often trails behind innovation, driven by legal cases associated with already-implemented systems, the onus of developing sound and 'humane' policy for AI development and usage will probably fall to requirements associated with ethics and robustness. The eight technical requirements proposed are as follows [36]:

1. Human agency and oversight;
2. Robustness and safety;

3. Privacy and data governance;
4. Transparency;
5. Diversity;
6. Non-discrimination and fairness;
7. Societal and environmental wellbeing;
8. Accountability.

Of these technical requirements, accountability and agency have the most direct links to safety culture.

2.7. Accountability, Certification, and the Double-Bind

Accountability is directly related to safety culture, and Just Culture in particular. Consider the pilot who ignores advice from an AI in favor of their own judgement, and then has an accident, but also the case wherein the pilot follows the AI's advice which turns out to be unsafe, also resulting in an incident or accident. In both cases, it will be easy to blame the human user rather than the AI, yet this could be unfair (such a situation is called a 'double bind' in psychology). Accountability (and justice) would require an adequate means of redress to discern whether the AI 'made a mistake'. As with autonomous car accidents, the question becomes one of whether there is transparency in terms of the equivalent of the AI's algorithms and calculations made, its data—both used and unused—and its trade-offs, if any were made between different priorities, including safety. Such data forensics may prove inconclusive because of the innate complexity and opacity of how advanced AIs work.

There may be a temptation, following an accident involving a human and AI working together, for the AI developer to claim that 'the human remains in charge'. But if the AI is partly taking control or heavily influencing the user, then this is a disingenuous argument. In self-driving cars the driver takes over in case of aberrant behaviour, but this does not appear realistic in situations where things happen suddenly and develop quickly, as may occur in an aircraft or in an air traffic scenario. If an AI tool is useful and meant to help aviation professionals, they will become, to an extent, reliant on it, and such reliance may reduce their own situational awareness. They may also lose skill fluency over time, if not entire skill sets. In the aftermath of an aviation accident, an AI manufacturer may well say that the human should have seen what was happening and taken command. What lines of redress will the aviation professional have in such a circumstance?

There may be an attempt to 'certify' the AI in the cockpit or air traffic ops center or tower, such that once it is certified it is 'fit for purpose', meaning that if anything goes wrong, the invisible finger of judgment swings towards the human. Since much of aviation already has a certification 'mindset', this seems reasonably likely. But, as noted elsewhere [30], "*Certification. . . cannot replace responsibility*". This means that redress is not simply a matter of putting disclaimers here and there. This relates to the aforementioned 'legal pillar', which is not yet written into law.

The degree of safety effort required to certify an AI tool will likely depend on its autonomy. For example, a more autonomous AI system, which could initiate and execute tasks on its own, would have a higher safety certification requirement than a machine learning system simply advising a controller on weather pattern formation, since in the latter the human is more involved and in command. This means that AI developers may be inclined to classify their tools towards the lower end of the classification scheme. A corollary to this is that any such classification scheme linked to certification requirements must be crystal clear so that it cannot be misapplied. Such regulatory approaches should be tested via 'regulatory sandboxes', to see how they would work in practical settings [36]. In relation to this, legal exemptions absolving AI developers of liability should be avoided. Such exemptions could unfairly shift the responsibility from large corporations to smaller actors, users, and communities lacking the resources, access and capabilities to address and alleviate all risks [30]. Such a principle is already being considered in the developing European Act on AI, discussed at the end of this section.

2.8. AI and Just Culture in Aviation

Just Culture is strongly linked to accountability, and is a cornerstone of safety culture in aviation, since it protects safety reporting and therefore enables safety learning. Just Culture is defined as follows [37]:

> "*Just Culture means a culture in which front-line operators or other persons are not punished for actions, omissions or decisions taken by them that are commensurate with their experience and training, but in which gross negligence, willful violations and destructive acts are not tolerated*".

This definition emphasizes that actions, omissions, or decisions taken by aviation professionals should be commensurate with their experience and training. This raises a key question: what formal training on AI/Machine Learning (ML) and its state-of-the-art algorithms should aviation professionals receive? A vast range of potential failure modes exist for ML systems [38], which does not include potential failure modes for future AI systems. Aviation professionals cannot be expected to become data scientists, and how should incident and accident investigators proceed? Will future investigations require data science expertise to mine the data, algorithms, and inputs which lead to a particular AI suggestion, whether right or wrong? Considering such concerns, for air traffic controllers at least, some argue [37] that:

> "*The burden of responsibility gravitates towards the organization to provide sufficient and appropriate training for air traffic controllers. If they are not well trained, it will be hard to blame them for actions, omissions or decisions arising from AI/ML situations. . .*".

It should also be noted that the existing definition of just culture is very human in its language, as it talks of *gross negligence*, *willful violations* and *destructive acts*, all of which signify intent and an understanding of 'right and wrong'. Can any of these terms apply to AI, now or even in the future? This is followed up [39] by considering legal implications, starting from the following vantage point:

> "*The functioning of AI challenges traditional tests of intent and causation, which are used in virtually every field of law*".

From the aviation professional's standpoint, the double-bind scenarios raised in the foregoing section seem particularly risky. Just culture basically states that people rarely go to work in order to cause an accident—quite the reverse—and so should not be punished for 'honest mistakes'. It is easy to have such a viewpoint before an accident, but after an air crash there is a natural—and very human—urge to search for someone to blame. After an accident, hindsight becomes very black and white in terms of 'surely the pilot should have known/done/realized. . .' whereas in reality, prior to the event, usually nothing was so black and white, and other pilots (or controllers, or airport personnel, etc.) may well have chosen exactly the same course of action. This happened shortly after the first B737 Max accident, with a number of pilots publicly stating '*. . .the pilots should have known what to do*'. Only after the second accident was it finally accepted that the design needed to change.

The problem with human-AI teaming and just culture is not simply a moral one. If aviation professionals are concerned about their accountability regarding AI, they will be reluctant to use it, or err on the side of caution, e.g., always agreeing with the AI if the situation is not clear-cut. They may also be less likely to report openly and honestly about their thinking and decision-making prior to the event. For example, stating that '*I did consider that the AI's advice might be correct, but preferred to rely on my own experience*', could lead to problems for the aviation professional in a courtroom situation. If professionals stop reporting incidents, or fail to disclose everything, this would be a retrograde step for aviation, which today has an excellent safety learning system.

According to [40], human-AI teaming is trustworthy by design if the humans and machines can rely on each other, self-organize to take advantage of each other's strengths and mitigate their weaknesses, and can be held accountable for their actions. It is this accountability with future AI systems that remains undefined.

AI categories 2B and 3A, wherein the AI can act autonomously, either in collaboration with the human (2B) or under a human overseer/management-by-exception operational framework (3A), are the 'ones to watch' from a Just Culture perspective. In both cases the AI could hypothetically be considered to have a certain degree of agency. It is then a question of whether the human can detect erroneous AI behavior (or conditions outside the AI's 'competence' or datasets) and intervene in time. In such cases, legal redress would likely fall to the organizations developing/owning the AI, which raises questions concerning the governance of AI systems in industry.

2.9. Ethical AI—Maintaining Meaningful Human Work

Just Culture is linked to the broader field of ethics. As noted above, there is concern that some people may lose their jobs to AI, or that their jobs will be less satisfying, or that they will gain new jobs but receive less remuneration and less favorable employment conditions. In aviation, such concerns are clearly relatable to the concept of single-pilot operations (SPOs) in the cockpit, which could be enabled by AI in the future. It is plausible that diminishing the human role could impact safety culture, because the human crew member may see safety as the AI's job rather than their own, especially if the AI becomes its own autonomous decision-maker. Such issues, essentially about the human's role in work in society, fall into the domain of ethics, which is itself a major issue in the developing AI arena. The European Commission's high-level expert group has outlined preliminary ethical principles for AI (HLEG) [41]:

- Respect for human autonomy: AI systems should not subordinate, coerce, deceive, manipulate, condition or herd humans. AI systems should augment, complement and empower human cognitive, social and cultural skills, leave opportunity for human choice and secure human oversight over work processes, and support the creation of meaningful work;
- Prevention of harm: AI must not cause harm or adversely affect humans, and should protect human dignity, and not be open to malicious use or adverse effects because of information asymmetries or unequal balance of power;
- Fairness: This principle links to solidarity and justice, including redress against decisions made by AI or the companies operating/making them.

Such principles bode well for maintaining human agency and autonomy, which can be critical for safety culture. However, they need to be translated into workable 'good practices' in the industry.

2.10. Human Agency for Safety—Maintaining Safety Citizenship

Earlier it was stated that in aviation, 'people create safety'. What this means is that aviation personnel, whether pilots, cabin crew, air traffic controllers, aeronautical engineers or airport personnel, believe that safety is at the core of their duties. But what if the system, through increasingly effective automation and AI, becomes ultra-safe? There is a concern that if people are effectively 'closed out' from safety, either via automation that excludes human intervention, or because it is simply ultra-safe, then 'safety citizenship'—the innate desire to keep things safe for ourselves and others—may degrade or disappear altogether [42].

Seven factors can erode safety citizenship [42], all of which could be affected by AI taking on a larger share of the safety role, or occupying the 'safety space':

- Safety role ambiguity;
- Safety role conflict;
- Role overload;
- Job insecurity;
- Job characteristics;
- Interpersonal safety conflicts;
- Safety restrictions.

The notion of safety citizenship and being proactive about safety (e.g., speaking up for safety), relies on a sense of self-determination. Whereas an AI can be defined according to its function (what it does) and/or its mechanism (how it achieves it), humans are defined according to agency by what they want to achieve (their goals and motivations), as well as their capabilities and limitations.

People need to feel autonomous (able to self-regulate their actions and experiences according to their interests and values), competent, and related (socially connected) to function in the world. This is self-determination theory. According to [42], this links to people's personal identity (how one feels about oneself) and their social identity (how society thinks about you and the group you belong to), for example, "I am competent as a pilot (or air traffic controller, or engineer, etc.), and pilots are useful in the world". Taking away the human's perception of identity and role can negatively affect self-determination. This may be expected to degrade safety culture, as the human's role in the system's overall 'safety space' (the hypothetical landscape of all safety functions and activities) diminishes.

2.11. AI Anthropomorphism and Emotional AI

Human-AI teaming is itself an anthropomorphic term [27], conveying the notion that the AI is in some sense a team player, devolving human qualities to a machine. This is reminiscent of generative AI systems wherein people sometimes believe they are conversing with a person rather than a program (e.g., ChatGPT). There are two aspects to this issue, one philosophical, the other more practical. The philosophical question is whether an AI, in the distant future, could have sentience. This remains too speculative an issue at this time, and so is left to other authors, as the focus of this paper is on narrow AI. Sentience would only likely become plausible with artificial general intelligence (AGI), which does not yet exist, though it may well do in the next decade [43].

The practical question is whether treating future AI systems as a team member could enhance overall team performance. A key sub-question is whether we can tell the difference between a human and an AI. In a recent study of 'emotional attachment' to AIs as team members [44], most participants could tell the difference between an AI and a human from the interaction, i.e., they guess correctly when it is an AI.

Another study [45] examined human trust in AIs as a function of the perception of the AI's identity. The study found that AI 'teammate' performance matters to HAT performance and trust, whereas AI identity does not. The study authors cautioned against using deceit to pretend an AI is a human. Deception about AI teammate's identity (pretending it is a human) did not improve overall performance, and led to less acceptance of their solutions, whereas knowing it is an AI led to better overall performance. What mattered most was the overall competency and helpfulness (utility) of the AI, which equates to how we learn to trust automation. What the authors also found was that AIs and hybrid-AI teams are better than human-only teams in terms of resource management in crisis management situations, and in a design engineering path-planning exercise.

Taken together, such results suggest that the concept of human-AI teaming does not require anthropomorphism. What will matter to the human members of the team and the executives deciding whether to deploy AI, is the effectiveness of the AI in doing its tasks. Additional critical considerations include how the AI affects the human team members' workload, and whether it has an overall positive impact on the team's performance.

These results are backed up by a further study [46] of attitudes to 'emotional' AI. This study is oriented more towards societal impacts than industrial ones, i.e., generative AI or AGI more so than narrow AI applications. It adds certain potential cultural impact areas into the human-AI teaming landscape, since industries, particularly global ones like aviation, are affected by diverse cultural norms. Key acceptance parameters for emotional AI were found to be loyalty (potential to erode existing social cohesion in the team), fairness, freedom from harm, purity (concerns about 'mental/spiritual contamination' by the AI) and authority (impacts on the status quo). As with the previous study, the authors found that people judge machines by outcomes.

It appears, therefore, that there is as yet no evidence of a performance benefit with emotional AIs. However, there is a reciprocal question concerning whether AIs need to be aware of human emotions. Would it make sense, for example, for AIs supporting humans in an emergency to be aware of stress in the humans' voices as conditions worsen? A recent study [42] found that monitoring people's behavior and emotional activity (speech, gestures, facial expressions, and physiological reactions), even if supposedly for health and wellbeing, can be seen as intrusive. Such monitoring activities can be for stress, fatigue and boredom monitoring, and error avoidance, and of course productivity. Yet, understandably, people may dislike this level of personal intrusion into their behavior, bodies, and personal data. Such data capture can lead to a feeling that the organization does not trust its staff.

Overall, therefore, aviation currently appears to need neither anthropomorphic nor emotional AI. This may sit better with safety culture, as considering the AI component of a human-AI team as an entity with feelings could very well 'muddy the waters' for safety responsibilities and safe interactions with the AI.

2.12. Governance of AI, and Organizational Leadership

In April 2021, the European Commission laid out a proposal for harmonized rules on AI [47]. The primary focus is on generative AI and AGI, where the intention is to have strong risk-based governance on high-risk applications in society. Interestingly, the provisional agreement intends to ban, for example, cognitive behavioral manipulation and emotion recognition in the workplace. However, it also states that there will be an obligation for users of an emotion recognition system to inform natural persons when they are being exposed to such a system. Perhaps, therefore, exceptions will be made for narrow AI applications where there might arguably be a safety advantage, e.g., civil and/or military aviation. Although the EU Act is mainly focused on generative AI and AGI, it is likely that its edicts, when published and written into European law, will set the tone for governance and regulation of narrow AI across a range of industries.

Three aspects from the EU Act on AI likely to bleed over into the industrial arena, including aviation, are notable [47]. The first is that AI systems must be sufficiently transparent to enable users to interpret the system's output and use it appropriately. Second, they must be resilient regarding errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular because of their interaction with natural persons or other systems. Third, human oversight must prevent or minimize safety risks that can emerge both when a high-risk AI system is used under its intended purpose or under conditions of reasonably foreseeable misuse.

Does high-level governance influence safety culture? Governance at this highest level (e.g., EU), along with global bodies such as ICAO [the International Civil Aviation Organization] and organizations such as the Federal Aviation Administration (FAA) in the US, and EUROCONTROL and EASA in Europe, can set the tone for the perception of AI's role in the aviation industry, and the tone matters. If the tone is that AI design must be human-centric and not negatively affect human wellbeing, nor displace the human workforce, this affects how business leaders and CEOs of key organizations—both operational and manufacturing—consider AI and its role, including that of safety. However, if for example, AI's capability is overestimated, such that human error is perceived as the problem and AI the solution, then the industry may work towards reducing human control inside the 'safety space', putting the safety of passengers and crews in the metaphorical hands of AI systems.

This is risky, as already identified in the maritime industry [48], since there can be 'tail effects', wherein low probability events are impractical to train AIs on, so that when they occur the AI cannot handle them. The maritime study suggests the need for active back-up control for autonomous ships (largely controlled from onshore control centers). In this scenario, human control is not decreasing. As AI autonomy goes up, passive back-up is likely to be ineffective, in part because AI can lead to *'increasing invisible interactions'*. In such a case, the humans miss what is going on in terms of the system and sub-system interactions

and relationships and cannot understand the complexity and gain a holistic picture. The maritime study authors also point out that in maritime operations managing VHF comms are easy for humans and hard for AI: part of the '*easy things are hard*' paradox in AI. Their conclusion runs counter to the way aviation (which is arguably more 'techno-centric') is currently heading:

> "*It seems counter-intuitive, then, to categorize the level of automation by degree of autonomous control gained over human control lost, when in practice both are needed to ensure safety/*" [48]

Again, looking for a moment outside the aviation domain, the UK Ministry of Defense has published its own AI Strategy [49]. The strategy does indeed set the tone from the outset with the following statement:

> "*Machines are good at doing things right; humans are good at doing the right thing*."

Such a statement clearly shows that human judgement will continue to be valued in future AI-enhanced defense platforms and scenarios. The paper asks whether the defense industry has the right culture, leadership, policies, and skills in place to make the best use of AI, which it considers it must develop to counter significant foreign threats now and in the future. The defense industry, therefore, is already considering how to approach AI and its potential autonomy from an organizational perspective, including a focus on training middle management concerning AI. The Defense AI Strategy, significantly, also poses a set of questions around when to use AI, and when not to, which sometimes appear missing in the current rush to 'try out AI' in a myriad of projects in several industries, including aviation:

- Where is AI the right solution?
- Do we have the right data?
- Do we have the right computing power?
- Do we have fit-for-purpose models?

A further relevant recent paper on organizational safety and autonomy [50] considered two models of how safety works in large organizations. Safety can be seen as a centralized, hierarchical, rule-based and compliance-based system, or decentralized, responding to local problems in an agile way, through 'loose couplings'. In the former, what the CEO says matters as it will be cascaded down through middle management to the rest of the workforce, including those in design and development, validation and testing, procurement, human resources and training. But even in a decentralized organizational arrangement, people will still respond to what top management says about AI and its role in the organization's strategy and operations. This has been evidenced by the importance of 'management commitment to safety' in many models of safety culture [5].

What top management says, however, needs to be borne from a well-informed understanding of AI and its realistic capabilities and limitations. In the current 'hype' around AI, the former is exaggerated and the latter often under-specified, ignored, or unknown. This may mean that those aviation organizations 'buying into AI' need to recruit serious AI expertise in-house, so that they can make balanced judgements at board level. Here, it is perhaps worth noting that following the two B737 Max accidents [20,21], Boeing invited someone new to their board who had aviation operational experience, since beforehand corporate goals—perhaps under-informed by operational insight—had unwittingly contributed to safety vulnerabilities emerging in the B737 Max design. This could be a salutary lesson for the top management of organizations considering using AI to transform their operations, that the key (AI/data science) expertise should not be buried too low in the organization, or simply outsourced. In a similar vein, following the UK Nimrod accident, the official accident report [51] stated that:

> "*Failures in leadership and organizational safety culture led to the Nimrod incident where the aircraft developed serious technical failures, preceded by deficiencies in safety case and a lack of proper documentation and communication between the relevant organizations*".

> On p. 474: "*The ownership of risk is fragmented and dispersed, and there is a lack of clear understanding or guidance what levels of risk can be owned/managed/mitigated and by whom*".
>
> And, p. 403: "*These organizational failures were both failure of leadership and collective failures to keep safety and airworthiness at the top of the agenda despite the seas of change during the period*".

At the outset of this paper, it was noted that organizations need both an SMS (safety management system) and safety culture. There is a very real danger that the potential safety impact of integration of future AI 'tech' into operational aviation systems is underestimated, if it is believed current SMSs can handle such a transition. This would effectively be the 'old wine in new bottles' approach and could lead to significant safety vulnerabilities emerging in future aviation systems.

Perhaps one thing CEOs need to know is that AIs cannot value things in the way humans can, especially safety, as currently it is not known how to program human values [30]. This may interest CEOs as they are often concerned with value alignment in the organization. As far as safety is concerned, humans can experience a range of emotions including fear and concern for lives under threat, and loss and grief in the event of a fatality, all of which can underpin a strong, even passionate value of safety. An AI cannot experience any of these, and while various reward schemes and supervised learning could in theory reinforce safety in the machine's workings, it will still be 'running the numbers', and if it gets them wrong, will experience neither remorse nor regret. Whilst AIs can mimic human behavior and even have a built-in 'persona', this remains mimicry; they are still machines, or simply 'just more automation' [27]. A CEO might therefore wish to maintain a human eye on the screens, and a human hand within reach of the joystick and, in military aviation where many more lives may be at stake, a human finger on the trigger.

3. Safety Culture Evaluation of Future Human–AI Teaming in Aviation

The foregoing section has outlined the multi-faceted challenges posed by AI in future aviation systems. The remainder of the paper reviews the prospect of intelligent assistants in aviation through the lens of safety culture measurement. Although such measurement tools were not developed with AI in mind, the questionnaire items and dimensions can be analyzed to see where and how intelligent assistants might affect an aviation respondent's answers.

3.1. Materials and Method

The validated EUROCONTROL Safety Culture Questionnaire was used for this study. This questionnaire has been applied in over 30 European member states during the past two decades, and variants of the questionnaire have been applied to a number of European airlines and airports.

Safety culture surveys normally proceed via many operational staff completing the survey anonymously, and then collating the results. Since the kinds of intelligent assistant (IA) of concern (i.e., with levels of autonomy 2B or higher) do not yet exist in aviation, such an approach was not possible. Instead, three aviation safety culture practitioners who have carried out multiple surveys involving airlines, airports, and air traffic controllers, and who are also currently working in the aviation human-AI teaming research area, participated in the study. All three experts are currently involved in multiple aviation human-AI teaming research projects, funded both by the Single European Sky ATM Research (SESAR) programme (https://www.sesarju.eu/ accessed on 8 April 2024) and the EU's flagship Horizon Europe research funding program, as well as reviewing the EASA guidance on human-AI teaming capabilities.

The most experienced expert, who has been involved in over 30 surveys over the past 20 years, carried out the first principal assessment. Each of the 48 questionnaire items from the EUROCONTROL questionnaire was considered in the context of a future intelligent assistant, e.g., 'commitment to safety' might suffer if the intelligent assistant appeared

to handle safety flawlessly. This could affect, for example, flight crew focus on safety, or managers running aviation organizations.

The results of this first analysis were then reviewed independently by the other two practitioners, with alternatives/queries raised. An effort was made by all three practitioners to consider not only negative impacts, but also potential positive ones. The three experts then met to resolve and complete the assessment, culminating in a table of key considerations for each safety culture item.

3.2. Methodological Shortcomings and Countermeasures

It is recognized that three experts is a small sample, and this renders the results speculative. It is a pragmatic approach at this early stage in the development of human-AI teaming systems, in order to gain a first approximate view of potential impacts, so that safety organizations and flight crew and controller bodies can consider how AI may impact their future. In this sense, one aim of this paper is to start a critical conversation on aviation AI and safety culture.

However, the intention is to carry out a second survey in 12–18 months' time, when the six HAIKU prototype AI systems and use cases have matured. At this stage, around 50 flight crew and air traffic controllers will take part in simulations working with their AI counterparts. Following these simulations, the participants will be given the safety culture questionnaire, as well as being given the option of taking part in interviews or focus groups, to gain a more informed assessment of the safety culture impacts they foresee.

Similarly, as the human-AI teaming concepts mature, a safety case approach will be adopted, to consider what safety issues might arise and what early mitigations could to be put in place. This has already begun with two of the use cases. The same operational personnel involved in the simulations and hazard studies will be the ones used in the final safety culture survey for each use case.

4. Results

Table 1 shows how each safety culture questionnaire item might be affected by an intelligent assistant supporting a human team. Each row shows the questionnaire item, the safety culture dimension it relates to, the assessed impact because of IA presence and whether the impact is judged likely to be high, medium, or low.

Table 1. Prospective analysis of the impact of AI on aviation safety culture.

Questionnaire Item	Dimension	IA Impact	H/M/L
B01 My colleagues are committed to safety.	Colleague commitment to safety	The IA would effectively be a digital colleague. The IA's commitment to safety would likely be judged according to the IA's performance. Human-supervised training, using domain experts with the IA, would help engender trust. The concern is that humans might 'delegate' some of their responsibility to the IA. A key issue here is to what extent the IA sticks rigidly to 'golden rules', such as aircraft separation minima (5NM lateral separation and 1000 feet vertical separation) or is slightly flexible about them as controllers may (albeit rarely) need to be. The designer needs to decide whether to 'hard code' some of these rules or allow a little leeway (within limits); this determines whether the IA behaves like 'one of the guys' or never, ever breaks rules.	High

Table 1. *Cont.*

Questionnaire Item	Dimension	IA Impact	H/M/L
B04 Everyone I work with in this organization feels that safety is their personal responsibility.	Colleague commitment to safety	Since an IA cannot effectively take responsibility, someone else may be held accountable for an IA's 'actions'. If a supervisor fails to see an IA's 'mistake', who will be blamed? HAIKU use cases may shed light on this, if there can be scenarios where the IA gives 'poor' or incorrect advice. If an IA is fully autonomous, this may affect the human team's collective sense of responsibility, since in effect they can no longer be held responsible.	High
B07 I have confidence in the people that I interact with in my normal working situation.	Colleague commitment to safety	As for B01, this will be judged according to performance. Simulator training with IAs should help pilots and others 'calibrate' their confidence in the IA. This may overlap significantly with B01.	High
B02 Voicing concerns about safety is encouraged.	Just culture and reporting	The IA could 'speak up' if a key safety concern is not being discussed or has been missed. This could be integrated into crew resource management (CRM) and threat and error management (TEM) practices, and team resources management (TRM) in air traffic management. However, then the IA may be considered a 'snitch', a tool of management to check up on staff. This could also be a two-way street, so that the crew could report on the IA's performance.	High
B08 People who report safety related occurrences are treated in a just and fair manner.	Just culture and reporting	The IA could monitor and record all events and interactions in real time and would be akin to a 'living' black box recorder. This could affect how humans behave and speak around the IA, if AI 'testimony' via data forensics was ever used against a controller in a disciplinary or legal prosecution case.	High
B12 We get timely feedback on the safety issues we raise.	Just culture and reporting	The IA could significantly increase reporting rates, depending on how its reporting threshold is set, and record and track how often a safety issue is raised.	Medium
B14 If I see an unsafe behavior by a colleague I would talk to them about it.	Just culture and reporting	[See also B02] The IA can 'query' behavior or decisions that may be unsafe. Rather than 'policing' the human team, the IA could possibly bring the risk to the human's attention more sensitively, as a query.	High
B16 I would speak to my manager if I had safety concerns about the way that we work.	Just culture and reporting	If managers have full access to IA records, the IA might become a 'snitch' for management. This would most likely be a deal-breaker for honest team-working.	Low
C01 Incidents or occurrences that could affect safety are properly investigated.	Just culture and reporting	As for B08, the IA's record of events could shed light on the human colleagues' states of mind and decision-making. There needs to be safeguards around such use, however, so that it is only used for safety learning.	High
C06 I am satisfied with the level of confidentiality of the reporting and investigation process.	Just culture and reporting	As for B16, the use of IA recordings as information or even evidence during investigations needs to be considered. Just culture policies will need to adapt/evolve to the use of IAs in operational contexts.	High

Table 1. *Cont.*

Questionnaire Item	Dimension	IA Impact	H/M/L
C09 A staff member prosecuted for an incident involving a genuine error or mistake would be supported by the management of this organization.	Just culture and reporting	This largely concerns management attitudes to staff and provision of support. However, the term 'genuine error or mistake' needs to encompass the human choice between following IA advice which turns out to be wrong, and ignoring such advice which turns out to be right, since in either case there was no human intention to cause harm. This can be enshrined in just culture policies, but judiciaries (and the travelling public) may take an alternative viewpoint. In the event of a fatal accident, black-and-white judgements sharpened by hindsight may be made, which do not reflect the complexity of IA's and human-AI teams' operating characteristics and the local rationality at the time.	High
C13 Incident or occurrence reporting leads to safety improvement in this organization.	Just culture and reporting	This is partly administrative and depends on financial costs of safety recommendations. Nevertheless, the IA may be seen as adding dispassionate evidence and more balanced assessment of severity, and how close an event came to being an accident (e.g., via Bayesian and other statistical analysis techniques). It will be interesting to see if the credence given to the IA by management is higher than that given to its human counterparts.	High
C17 A staff member who regularly took unacceptable risks would be disciplined or corrected in this organization.	Just culture and reporting	As for C09, an IA may know an individual who takes more risks than others. However, there is a secondary aspect, linked to B07, that the IA may be trained by humans, and may be biased by their own level of risk tolerance and safety–productivity trade-offs. If an IA is offering solutions judged too risky, or conversely 'too safe', nullifying operational efficiency, the IA will need 're-training' or re-coding.	High
B03 We have sufficient staff to do our work safely.	Staff and equipment	Despite many assurances that AI will not replace humans, many see strong commercial imperatives for doing exactly that (e.g., a shortage of commercial pilots and impending shortage of air traffic controllers, post-COVID low return-to-work rate at airports, etc.).	High
B23 We have support from safety specialists.	Staff and equipment	The IA could serve as a 'safety encyclopedia' for its team, with all safety rules, incidents and risk models stored in its knowledge base.	Medium
C02 We have the equipment needed to do our work safely.	Staff and equipment	The perceived safety value of IAs will depend on how useful the IA is for safety and will be a major question for the HAIKU use cases. One 'wrong call' could have a big impact on trust.	High
B05 My manager is committed to safety.	Management commitment to safety	The advent of IAs needs to be discussed with senior management, to understand if it affects their perception of who/what is keeping their organization safe. They may come to see the IA as a more manageable asset than people, one that can be 'turned up or down' with respect to safety.	High
B06 Staff have high trust in management regarding safety.	Management commitment to safety	Conversely, operational managers may simply be reluctant to allow the introduction of IAs into the system, because of both safety and operational concerns.	Medium
B10 My manager acts on the safety issues we raise.	Management commitment to safety	See C13 above.	Low

Table 1. *Cont.*

Questionnaire Item	Dimension	IA Impact	H/M/L
B19 Safety is taken seriously in this organization.	Management commitment to safety	Depends on how much the IA focuses on safety. The human team will watch the IA's 'behavior' closely and judge for themselves whether the IA is there for safety or for other purposes. These could include profitability, but also a focus on environment issues. Ensuring competing AI priorities do not conflict may be challenging.	Medium
B22 My manager would always support me if I had a concern about safety.	Management commitment to safety	See B16, C09, C17. If the IA incorporates a dynamically updated risk model, concerns about safety could be rapidly assessed and addressed according to their risk importance (this is the long-term intent of Use Case 5 in HAIKU).	Low
B28 Senior management takes appropriate action on the safety issues that we raise.	Management commitment to safety	See B12. A further aspect is whether (and how quickly) the management supports getting the IA 'fixed' if its human teammates think it is not behaving safely.	Low
B09 People in this organization share safety related information.	Communication	The IA could become a source of safety information sharing, but this would still depend on the organization in terms of how the information would be shared and with whom. The IA could however share important day-to-day operational observations, e.g., by flight crew who can pass on their insights to the next crew flying the same route, for example, or by ground crew at an airport. Some airports already use a 'community app' for rapid sharing of such information.	Medium
B11 Information about safety related changes within this organization is clearly communicated to staff.	Communication	The IA could again be an outlet for information sharing, e.g., notices could be uploaded instantly, and the IA could 'brief' colleagues or inject new details as they become relevant during operations. The IA could also upload daily NOTAMs (Notices to Airmen) and safety briefings for controllers, and could distill the key safety points, or remind the team if they forget something from procedures/NOTAMs/ briefings notes.	Medium
B17 There is good communication up and down this organization about safety.	Communication	An IA could reduce the reporting burden of operational staff if there could be an IA function to transmit details of concerns and safety observations directly to safety departments (though the 'narrative' should still be written by humans). An IA 'network' or hub could be useful for safety departments to assess safety issues rapidly and prepare messages to be cascaded down by senior/middle management.	Medium
B21 We learn lessons from safety-related incident or occurrence investigations.	Communication	The IA could provide useful and objective input for safety investigations, including inferences on causal and contributory factors. Use of Bayesian inference and other similar statistical approaches could avoid some typical human statistical biases, to help ensure the right lessons are learned and are considered proportionately to their level of risk. Alternatively, if information is biased or counterfactual evidence is not considered, the way the IA judges risk may be incorrect, leading to a lack of trust by operational people. It could also leave managers focusing on the wrong issues.	High

Table 1. *Cont.*

Questionnaire Item	Dimension	IA Impact	H/M/L
B24 I have good access to information regarding safety incidents or occurrences within the organization.	Communication	IAs or other AI-informed safety intelligence units could store a good deal of information on incidents and accidents, with live updates, possibly structured around risk models, and capture more contextual factors than are currently reported (this is the aim of HAIKU Use Case 5). Information can then be disseminated via an app or via the IA itself to various crews/staff.	High
B26 I know what the future plans are for the development of the services we provide.	Communication	The implementation and deployment of IAs into real operational systems needs careful and sensitive introduction, as there will be many concerns and practical questions. Failure to address such concerns may lead to very limited uptake of the IA.	Medium
C03 I read reports of incidents or occurrences that apply to our work.	Communication	The IA could store incidents, but this would require nothing so sophisticated as an IA. However, if the IA is used to provide concurrent (in situ) training, it could bring up past incidents related to the current operating conditions.	Low
C12 We are sufficiently involved in safety risk assessments.	Communication	Working with an IA might give the team a better appreciation of underlying risk assessments and their relevance to current operations.	Low
C15 We are sufficiently involved in changes to procedures.	Communication	The IA could build up evidence of procedures that regularly require workarounds or are no longer fit for purpose. The IA could highlight gaps between 'work as designed' and 'work as done'.	Medium
C16 We openly discuss incidents or occurrences to learn from them.	Communication	[See C03] Unless this becomes an added function of the IA, it has low relevance. However, if a group learning review, or threat and error management is used in the cockpit following an event, the IA could provide a dispassionate and detailed account of the sequence of events and interactions.	Low
C18 Operational staff are sufficiently involved in system changes.	Communication	There is a risk that if the IA is a very good information collector, people at the sharp end might be gradually excluded in updates to system changes, as the system's developers will consult data from the IA instead.	Medium
B13 My involvement in safety activities is sufficient.	Collaboration	As for C15 and C18.	Low
B15r People who raise safety issues are seen as troublemakers.	Collaboration	It needs to be seen whether an IA could itself be perceived as a troublemaker if it continually questions its human teammates' decisions and actions.	Medium
B20 My team works well with the other teams within the organization.	Collaboration	The way different teams 'do' safety in the same job may vary (both inside companies, and between companies). The IA might need to be tailored to each team, or be able to vary/nuance its responses accordingly. If people move from one team or department to another, they may need to learn 'the way the IA does things around here'.	Medium
B25r There are people who I do not want to work with because of their negative attitude to safety.	Collaboration	There could conceivably be a clash between an IA and a team member who, for example, was taking significant risks or continually overriding/ignoring safety advice, or an IA that was giving poor advice. If the IA is a continual learning system, its behavior may evolve over time, and diverge from optimum, even if it starts off safe when first implemented.	High

Table 1. *Cont.*

Questionnaire Item	Dimension	IA Impact	H/M/L
B27 Other people in this organization understand how my job contributes to safety.	Collaboration	The implementation of an IA in a particular work area (e.g., a cockpit, an air traffic ops room, an airport/airline operational control center) itself suggests safety criticality of human tasks in those areas. If an IA becomes an assimilator of all safety relevant information and activities, it may become clearer how different roles contribute to safety.	Medium
C05 Good communication exists between Operations and Engineering/Maintenance to ensure safety.	Collaboration	If engineering/maintenance 'own' the IA, i.e., are responsible for its maintenance and upgrades, then there will need to be good communication between these departments and ops/safety. A secondary aspect here is that IAs used in ops could transmit information to other departments concerning engineering and maintenance needs observed during operations.	Medium
C10 Maintenance always consults Operations about plans to maintain operational equipment	Collaboration	It needs to be determined who can upgrade an IA's system and performance characteristics, e.g., if a manual change is made to the IA to better account for an operational circumstance that has caused safety issues, who makes this change and who needs to be informed?	Medium
B18 Changes to the organization, systems and procedures are properly assessed for safety risk.	Risk handling	The IA could have a model of how things work and how safety is maintained, so any changes will need to be incorporated into that model, which may identify safety issues that may have been overlooked or played down. This is like current use of AIs for continuous validation and verification of operating systems, looking for bugs or omissions. Conversely, the IA may give advice that makes little sense to the human team, or the organization yet be unable to explain its rationale. Humans may find it difficult to adhere to such advice.	High
C07r We often have to deviate from procedures.	Risk handling	The IA will observe (and perhaps be party to) procedural deviation and can record associated reasons and frequencies (highlighting common 'workarounds'). Such data could identify procedures that are no longer fit for purpose, or else inform retraining requirements if the procedures are in fact still fit for purpose.	High
C14r I often have to take risks that make me feel uncomfortable about safety.	Risk handling	The IA will likely be unaware of any discomfort on the human's part (unless emotion detection is employed), but the human can probably use the IA's advice to err on the side of caution. Conversely, a risk-taker, or someone who puts productivity first, may consult an IA until it gets around the rules (human ingenuity can be used for the wrong reasons).	High
C04 The procedures describe how I actually do my job.	Procedures and training	People know how to 'fill in the gaps' when procedures do not really fit the situation, and it is not clear how an IA will do this. [This was in part why the earlier expert systems movement failed to deliver, leading to the infamous 'AI winter']. Also, the IA could record 'work as done' and contrast it to 'work as imagined' (the procedures). This would, over time, create an evidence base on procedural adequacy (see also C07r).	High

Table 1. *Cont.*

Questionnaire Item	Dimension	IA Impact	H/M/L
C08 I receive sufficient safety-related refresher training.	Procedures and training	The IA could note human fluency with the procedures and how much support it has to give, thus gaining a picture of whether more refresher training might be beneficial.	Medium
C11 Adequate training is provided when new systems and procedures are introduced.	Procedures and training	As for C08.	Medium
C19 The procedures associated with my work are appropriate.	Procedures and training	When humans find themselves outside the procedures, e.g., in a flight upset situation in the cockpit, an IA could rapidly examine all sensor information and supply a course of action for the flight crew.	High
C20 I have sufficient training to understand the procedures associated with my work.	Procedures and training	As for C08 and C11.	Medium

The analysis in Table 1 suggests a broad categorization of the IA's impact on the various safety culture dimensions, from high to low, as follows:

- High impact: Colleague commitment to safety, just culture and reporting, risk handling;
- Medium impact: Staff and equipment, procedures and training, communication and learning, collaboration and involvement
- Low impact: Management commitment to safety

Each of these can be considered either a concern about negative impacts on safety culture that needs to be managed, or alternatively a safety 'affordance', wherein the IA could help support and possibly enhance current safety culture, safety management processes, and operational safety practices. Since several insights in the rows in Table 1 overlap or point to a single central issue, the full results in Table 1 were further refined to distil the key insights from the analysis, in terms of safety culture concerns, and safety culture affordances. These are shown in Table 2.

Table 2. Safety culture concerns and affordances related to future human-AI teaming in aviation.

Safety Culture Concerns	Safety Culture Affordances
Humans may become less concerned with safety if the IA is seen as handling safety aspects. This is an extension of the 'complacency' issue with automation and may be expected to increase as the IA's autonomy increases.	The IA could 'speak up' if it assesses a human course of action as unsafe.
Humans may perceive a double-bind; if they follow 'bad' IA advice or fail to follow 'good' advice, and there are adverse consequences, they might find themselves being prosecuted. This will lead to lack of trust in the IA.	The IA could be integrated into crew resource management practices, helping decision-making and post-event review in the cockpit or air traffic ops room.
If the IA reports on human error or human risk-taking or other 'non-nominal behavior' it could be considered a 'snitch' for management and may not be trusted.	The IA could serve as a living black box recorder, recording more decision-making rationales than is the case today.
If IA recordings are used by incident and accident investigators, just culture policies will need to address such usage both for ethical reasons and to the satisfaction of the human teams involved. Fatal accidents in which an IA was a part of the team are likely to raise new challenges for legal institutions.	If the IA can collect and analyze day-to-day safety occurrence information it may be seen as adding objective (dispassionate) evidence and a more balanced assessment of severity, as well as an unbiased evaluation of how close an event came to being an accident (e.g., via Bayesian analysis).

Table 2. *Cont.*

Safety Culture Concerns	Safety Culture Affordances
An IA that is human trained may adopt its human trainers' level of risk tolerance, which may not always be optimal for safety.	The IA could significantly increase reporting rates, depending on how its reporting threshold is set, and could also record and track how often a safety-related issue is raised.
Introducing intelligent assistants may inexorably lead to less human staff. Although there are various ways to 'sugar-coat' this, e.g., current and predicted shortfalls in staffing across the aviation workforce, it may lead to resentment against IAs. This factor will likely be influenced by how society gets on with advanced AI and IAs.	The IA could serve as a safety encyclopedia, or oracle, able to give instant information on safety rules, risk assessments, hazards, etc.
If the IA queries humans too often, it may be perceived as policing them, or as a troublemaker.	The IA can upload all NOTAMs and briefings etc., so as to keep the human team current, or to advise them if they have missed something.
If the IA makes unsafe suggestions, trust will be eroded rapidly.	If the IA makes one notable 'save', its perceived utility and trustworthiness will increase.
The IA may have multiple priorities (e.g., safety, environment, efficiency/profit). This may lead to advice that humans find conflicted or confusing.	The IA could share important day-to-day operational observations, e.g., by flight crew, controllers, or ground crew, who can pass on their insights to the incoming crew.
Management may come to see the IA as a more manageable safety asset than people, one where they can either 'turn up' or 'tone down' the accent on safety.	The IA could reduce the reporting 'burden' of operational staff by transmitting details of human concerns and safety observations directly to safety departments. An IA 'network' or hub would allow safety departments to assess safety issues rapidly and prepare messages to be cascaded down by senior/middle management.
Operational managers may simply be reluctant to allow the introduction of IAs into the system, because of both safety and operational reservations.	The IA could provide objective input for safety investigations, including inferences on causal and contributory factors. Use of Bayesian inference and other similar statistical approaches could help avoid typical human statistical biases, ensuring the right lessons are learned and are considered proportionately to their level of risk.
If information is biased or counterfactual evidence is not considered, the way the IA judges risk may be incorrect, leading to a lack of trust by operational people. It could also have managers focusing on the wrong issues.	IAs could store information on incidents and associated (correlated) contextual factors, with live updates structured around risk models, and disseminate warnings of potential hazards on the day via an app or via the IA itself communicating with crews/staff.
There is a risk that if the IA is a very good information collector, that people at the sharp end are gradually excluded in updates to system changes, as the systems developers will consult data from the IA instead.	The IA might serve as a bridge between the way operational people and safety analysts think about risks, via considering more contextual factors not normally encoded in risk assessments.
There could conceivably be a clash between an IA and a team member who, for example, was taking significant risks or continually over-riding/ignoring safety advice, or, conversely, an IA that was giving bad advice.	The IA could build up evidence of procedures that regularly require workarounds or are no longer fit for purpose. The IA could highlight gaps between 'work as designed', and 'work as done'.
IAs may need regular maintenance and fine-tuning, which may affect the perceived 'stability' of the IA by ops people, resulting in loss of trust or 'rapport'.	IAs used in ops could transmit information to other departments concerning engineering and maintenance needs observed during operations.
The IA may give (good) advice that makes little sense to the human team or the organization, yet it cannot explain its rationale. Managers and operational staff may find it difficult to adhere to such advice.	The IA could have a model of how things work and how safety is maintained, so that any changes will need to be incorporated into the model, which may identify safety issues that have been overlooked or 'played down'. This is like current use of AIs for continuous validation and verification of operating systems, looking for bugs or omissions.

Table 2. *Cont.*

Safety Culture Concerns	Safety Culture Affordances
A human risk-taker or someone who puts productivity first, may consult ('game') an IA until it gets around the rules.	The human can use the IA's safety advice to err on the side of caution, if she or he feels pressured to cut safety corners either because of self, peer, or management pressure.
People know how to fill in the gaps when procedures don't really fit the situation, and it is not clear how an IA will do this. The AI's advice might not be so helpful unless it is human-supervisory trained.	When humans find themselves outside the procedures, e.g., in a flight upset situation in the cockpit, an IA could rapidly examine all sensor information and supply a course of action for the flight crew.

The results in Table 2 suggest broad equivalence between potential positive and negative impacts of AI on safety culture, though the consequences of loss of safety culture could be much more dramatic in terms of aviation accidents. The next section discusses how to mitigate the negative impacts whilst bolstering the positive ones, allocating issues to risk owners in organisations.

5. Discussion of Results

5.1. Safeguards and Organizational Risk Owners

The analysis above has raised a number of potential threats to safety culture, and a more or less equivalent number of 'safety culture affordances' wherein safety culture could be enhanced. In this sense, the overall impact of AI and human-AI teaming on safety culture will depend on how it is researched, designed, developed, deployed, and managed in actual operational environments. The issues identified, whether positive or negative, can lead to safeguards to prevent safety from being diminished because of the introduction of advanced AI systems into aviation. However, for safeguards to be effective, those who can enact them also need to be identified.

The various impacts are diversely spread across different human 'levels' in organizations; some relating to front-line staff, some to middle management, and some to senior or executive levels. Safety culture always works best in aviation when those at the top—CEOs, VPs and Executive Boards, firmly believe in and support safety as a priority. There needs to be continued safety stewardship by senior executives to maintain the human as the key safety agent, which can then be translated by middle management throughout the organization into satisfactory actionable outcomes. A critical risk owner will be the safety department, typically the hub of safety learning in an organization, as well as the key working interface with external regulators. Next are the front-line and support staff, who are the lifeblood of safety culture in any operational organization. A four-layer model is shown in Figure 4, with the identified safeguards inserted at corresponding 'risk owner' levels.

At the bottom layer are the principal safeguards related to front-line and support staff. First is the need to maintain human agency for safety, i.e., a valid safety role. Second is the fact that the IA can act as a second pair of eyes, whether aiding in an emergency, or noting a safety issue or deviation or risky course of action by the human operator. This leads to a third useful aspect of an IA, that it can be a ready-to-hand safety oracle that the human team can consult at any point when considering the best course of action and the safety risks it might entail. The IA could also be programmed to *'speak up'* for safety if warranted, and this can be embedded into human CRM and TRM practices and training. The IA could be a useful aid for safety reporting, able to rapidly capture events, their precursors, signals and actions, to which the human could then add a narrative. NOTAMs (Notices to Airmen) could be automatically uploaded into the IA, which could remind human crews if they have forgotten or overlooked any relevant aspects during operations. Similarly, the IA could be useful as a day-to-day briefing tool, letting the oncoming shift know of anything unusual, e.g. changes to procedures, or the status of ongoing maintenance, etc. that has happened on previous shifts. Taken together, these eight safeguards could keep safety

at the human's fingertips, eyes, or ears, whilst guarding against simple omissions and reckless acts.

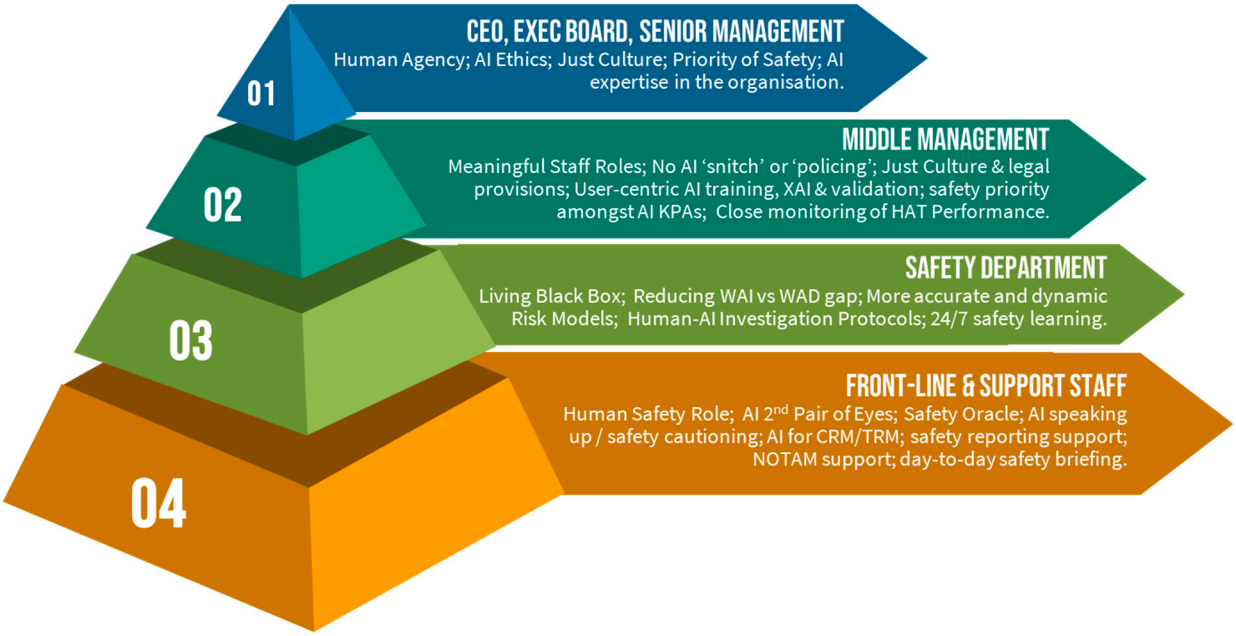

Figure 4. Safety culture safeguards and organizational risk-owners.

At the next layer is the safety department. A first safeguard is the notion of the IA serving as a living black box, such that after an event the IA could reproduce the detailed flow of events, signals, interactions, decisions made and even the thinking underpinning those decisions, prior to and during the incident. This could paint a much more detailed picture than investigators currently have. Such an 'annotated timeline' could also be very useful in safety learning and training. In parallel, investigators will undoubtedly need to develop new human-AI investigation protocols to deal with human-AI teaming events, particularly when relating to the double-bind type of scenarios raised earlier. These protocols should be informed by just culture principles adopted at higher levels in the organization and hopefully enshrined in European law.

The IA could also compare ways of working (what is actually done) against procedures and rules, not to police, but as a way of defining the gap between real operational practices and the official rules and procedures. If the gaps are unsafe, then this can lead to more training. But in many cases it is likely that the official rules are either inefficient or unworkable in real operational conditions, or else need updating as operations and technology has moved on. Similarly, risk models in aviation are often seen as not reflecting operational reality or being at too high a level of abstraction. The IA could record interactions in both safety-related events as well as 'when things go right'. Such information could be fed into risk models to render them more operationally relevant, giving them a more detailed level of description. If this can be achieved, then such models can become useful to operations departments, and not just seen as being for safety departments and regulators. The goal here would be that day-to-day operations are feeding dynamic risk models and safety dashboards, so that live safety performance can be seen, including when things may be drifting towards danger, or when new hazards are emerging. This could pave the way for true 24/7 safety monitoring and real-time learning.

At the next level up in the pyramid model is middle management, who have the challenge of exercising senior management aspirations within real world operational and resource constraints. Part of their mission regarding safety culture and AI is to ensure that staff have meaningful jobs, that IAs do not act as '*snitches*' on staff or police them, and that just culture ideals can be translated into effective and trusted principles and practices

enacted at lower levels, in agreement with social partners (unions etc.). A key role will be the overseeing of introducing autonomous AI systems into the organization, ensuring that they are *user centric.* This is especially the case for '*explainablility*' of the AI's advice or decisions (XAI) to the human, as well as for human-supervised-learning, user validation, and human-AI team training prior to operational deployment. Since 'user centeredness' is likely to be a common safeguarding theme, the discipline of human factors is likely to be a critical determinant of success in these activities.

If the IA, as is likely, has more key performance areas (KPAs) than safety alone (e.g., productivity, green-ness, etc.), then middle management must ensure that safety retains priority when the IA is making trade-offs. Last, middle management must closely monitor the IA's performance as it is deployed, as it will evolve both in its dealings with humans and other IAs.

The top level in the pyramid is senior management, including CEOs, VPs, Directors and Executive Boards. Here is where there needs to be an authentic message that safety is the priority and that 'people (still) create safety', albeit backed up and supported by AI. There should also ideally be a code of ethics related to the use of AI in the organization, as well as a just culture policy and framework, which deals with AI accountability in the case of an accident, to the satisfaction of social partners. It is also at this level that decisions need to be made on having internal AI expertise in the organization, so that organizational leaders can maintain a basic understanding and realistic expectations of their AI 'assets' and be prepared to face and answer the media when things go wrong.

5.2. Further Research Needed

The above section considers specific safeguards and allocates them to risk owners to facilitate their development into operational practices. Taken together, they comprise a preliminary future safety culture strategy for aviation organizations, i.e., a vision of how safety culture could look in the coming decade as AI autonomy rises and intelligent assistants enter the workplace.

However, organizations rarely operate unilaterally. They are subject to industry and regulatory standards and best practices, as well as external laws and edicts such as the forthcoming European Act on AI. Therefore, certain cornerstones of safety need to 'raise their game' in preparation for more advanced AI systems, so that when such systems arrive, organizations will have the right theory, tools, and regulatory landscape to put effective safeguards into place. This is particularly the case since, despite the EU showing foresight via its provisional Act on AI and the EASA providing early guidance on human-AI teaming regulation, the big AI innovations and revolutions themselves are likely to occur outside of Europe.

Five broad research areas are accordingly listed below, aimed at bolstering cross-industry pillars of safety that could both support and leverage organizations in their efforts to secure a stable foundation for safety and safety culture in future AI-assisted aviation:

1. Just culture—if just culture is to be preserved, rationales and arguments need to be developed that will stand up in courts of law. These must protect crew and workers who made an honest (i.e., a priori reasonable) judgement about whether to follow AI advice, and whether to intervene, contravening AI autonomous actions seen as potentially dangerous. Such development of just culture argumentation and supporting principles regarding AI and human-AI teaming should include simulated test cases being run in 'legal sandboxes'.
2. Safety management systems (SMS)—the key counterpart of safety culture in aviation—the SMS—will also need to adapt to higher levels of AI autonomy, as is already being suggested in [31,33]. This will probably require new thinking and new approaches, for example with respect to the treatment of human-AI teaming in risk models, rather than simply producing 'old wine in new bottles.' SMS maturity models, such as those that are used in air traffic organizations around the globe [52], will also need to adapt to address advanced AI integration into operations.

3. Human factors have a key role to play in the development of human-AI teaming [53], especially if such systems are truly intended to be human-centric. This will require co-development work between human factors practitioners/researchers and data scientists/AI developers, so that the human—who after all literally has 'skin in the game'—is always fully represented in the determination of optimal solutions for explainability, team-working, shared situation awareness, supervised learning, human-AI interaction means and devices, and training strategies. Several new research projects are paving the way forward. This applied research focus needs to be sustained, and a clear method developed for assuring usable and trustworthy human-AI teaming arrangements.
4. There are currently several human-AI teaming options on the table, e.g., from EASA's 1B to 3A; see also [54]), with 2A, 2B and 3A offering the most challenges to the human's agency for safety, and hence the most potential impacts on safety culture. Yet, these are the levels of AI autonomy that could also bring significant safety advantages. It would be useful, therefore, to explore the actual relative safety advantages and concomitant risks of these and other AI autonomy levels, via risk evaluations of aviation safety-related use cases. Such analysis could lead to common and coordinated design philosophies and practices for aviation system manufacturers.
5. Inter-sector collaboration will be beneficial, whether between human-AI teaming developments in different transport modalities (e.g., road, sea, and rail) or different industry sectors, including the military, who are likely to be most challenged with both ethical dilemmas and high intensity, high-risk human-AI team-working. This paper has already highlighted learning points from maritime and military domains for aviation, so closer collaboration is likely to be beneficial. At the least, collaboration between the transport domains makes sense, given that in the foreseeable future, AIs from different transport modes will likely be interacting with each other.

6. Limitations of the Study and Further Work

This paper and its analysis are speculative, given that the focus is high-autonomy intelligent assistance in aviation, which does not yet exist. Such speculation is arguably warranted, however, to forestall potential negative impacts on the very fabric of safety culture in aviation, as well as to 'get AI right the first time' by capitalizing on potential safety affordances. As noted earlier, there is some breathing space, but AI is a fast-developing and potentially disruptive technology and could well become a 'game-changer' for the industry. It is intended to follow up the safety culture exercise in this paper with another one towards the end of the HAIKU project in 2025, when more operational and research personnel have been exposed to realistic simulations with AI counterparts in a broad range of use cases. This may also lead to the development of a safety culture questionnaire focusing solely on human-AI teaming.

7. Conclusions

This paper has reviewed the current positive state of safety culture in aviation, and the future AI possibilities and challenges for the industry, focusing on human-AI teaming envisaged for the 2030+ timeframe, wherein an intelligent assistant could have a moderate or high degree of autonomy in an operational environment. The results of a preliminary analysis of the potential impacts of future advanced AI on aviation safety culture suggest there are both significant threats but also potential benefits, depending on how the AI is designed and implemented, and whether the AI is 'human-centric' or not.

Given the importance of safety culture to the aviation safety industry, adopting a 'wait and see' attitude is not advisable. Accordingly, a range of safeguards with associated organizational risk owners have been proposed, along with more fundamental research avenues to help aviation navigate a safe course through the development and deployment of advanced AI support systems. Such safeguards will help ensure that aviation's hard-won level of safety culture and safety are maintained, if not improved. A cornerstone to all these

strategies is that the human must maintain a strong safety role in aviation, whatever the AI's role. It may be that in the far future, e.g., 2050, AI will have proven itself to be more reliable than humans, but until such time, people will continue to create safety, and should remain at the heart of safety in aviation operations.

One way forward for the aviation industry to show leadership and keep humans at the center of safety would be a pan-industry charter on aviation human-AI teaming. Such a charter could set out the principles that are to be adopted in AI conceptualization, interaction design, training, deployment, and post-operational system performance monitoring. Such a charter could also highlight key development and research avenues, as well as 'guard-rails' for all those working towards the improvement of the aviation system, steering a course to ensure aviation maintains its reputation as the safest mode of transportation.

Funding: This publication is based on work performed in the HAIKU Project which has received funding from the European Union's Horizon Europe research and innovation program, under Grant Agreement no 101075332. Any dissemination reflects the authors' view only and the European Commission is not responsible for any use that may be made of information it contains.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Acknowledgments: The author would like to acknowledge the support of Andrew Kilner and Beatrice Bettignies-Thiebaux from EUROCONTROL, for early comments on the paper and the concerns/affordances. Thanks also to the three reviewers for their useful and insightful comments.

Conflicts of Interest: The author declares no conflict of interest.

References

1. EUROSTAT. Air Safety Statistics in the EU. 2023. Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Air_safety_statistics_in_the_EU&oldid=587873 (accessed on 27 March 2024).
2. Guldenmund, F. Understanding safety culture through models and metaphors: Taking stock and moving forward. In *Safety Cultures, Safety Models*; Gilbert, C., Journé, B., Laroche, H., Bieder, C., Eds.; Springer Open: Cham, Switzerland, 2018.
3. Cox, S.; Flin, R. Safety Culture: Philosopher's Stone or Man of Straw? *Work Stress* **1998**, *12*, 189–201. [CrossRef]
4. Zohar, D. Thirty years of safety climate research: Reflections and future directions. *Accid. Anal. Prev.* **2010**, *42*, 1517–1522. [CrossRef] [PubMed]
5. Reader, T.W.; Noort, M.C.; Kirwan, B.; Shorrock, S. Safety sans frontieres: An international safety culture model. *Risk Anal.* **2015**, *35*, 770–789. [CrossRef] [PubMed]
6. Advisory Committee on the Safety of Nuclear Installations (ACSNI) Study Group. *Third Report: Organizing for Safety*; H.M. Stationery Office: Sheffield, UK, 1993.
7. IAEA. *Safety Culture*; Safety Series No. 75-INSAG-4; International Atomic Energy Agency: Vienna, Austria, 1991.
8. Cullen, D. *The Public Enquiry into the Piper Alpha Disaster*; HMSO: London, UK, 1990.
9. Hidden, A. *Investigation into the Clapham Junction Railway Accident*; HMSO: London, UK, 1989.
10. Turner, R.; Pidgeon, N. *Man-Made Disasters*, 2nd ed.; Butterworth-Heineman: Oxford, UK, 1997.
11. Reason, J.T. *Managing the Risks of Organizational Accidents*; Ashgate: Aldershot, UK, 1997.
12. AAIB. *Report No: 4/1990. Report on the Accident to Boeing 737-400, G-OBME, near Kegworth, Leicestershire on 8 January 1989*; Air Accident Investigation Board, Dept of Transport: Hampshire, UK, 1990. Available online: https://www.gov.uk/aaib-reports/4-1990-boeing-737-400-g-obme-8-january-1989 (accessed on 27 March 2024).
13. Nunes, A.; Laursen, T. Identifying the factors that led to the Uberlingen mid-air collision: Implications for overall system safety. In Proceedings of the 48th Annual Chapter Meeting of the Human Factors and Ergonomics Society, New Orleans, LA, USA, 20–24 September 2004.
14. ANSV. Accident Report 20A-1-04, Milan Linate Airport 8 October 2001. Agenzia Nazionale Per La Sicurezza Del Volo, 00156 Rome. 20 January 2004. Available online: https://skybrary.aero/bookshelf/ansv-accident-report-20a-1-04-milan-linate-ri (accessed on 27 March 2024).
15. Mearns, K.; Kirwan, B.; Reader, T.W.; Jackson, J.; Kennedy, R.; Gordon, R. Understanding Safety Culture in Air Traffic ManagementDevelopment of a methodology for understanding and enhancing safety culture in Air Traffic Management. *Saf. Sci.* **2011**, *53*, 123–133. [CrossRef]
16. Noort, M.; Reader, T.W.; Shorrock, S.; Kirwan, B. The relationship between national culture and safety culture: Implications for international safety culture assessments. *J. Occup. Organ. Psychol.* **2016**, *89*, 515–538. [CrossRef] [PubMed]

17. Kirwan, B.; Shorrock, S.T.; Reader, T. *The Future of Safety Culture in European ATM—A White Paper*; EUROCONTROL: Brussels, Belgium, 2021. Available online: https://skybrary.aero/bookshelf/future-safety-culture-european-air-traffic-management-white-paper (accessed on 27 March 2024).
18. Kirwan, B.; Reader, T.W.; Parand, A.; Kennedy, R.; Bieder, C.; Stroeve, S.; Balk, A. *Learning Curve: Interpreting the Results of Four Years of Safety Culture Surveys*; Aerosafety World, Flight Safety Foundation: Alexandria, VA, USA, 2019.
19. Kirwan, B. *CEOs on Safety Culture*; A EUROCONTROL-FAA Action Plan 15 White Paper. October; EUROCONTROL: Brussels, Belgium, 2015. [CrossRef]
20. Zweifel, T.D.; Vyal, V. Crash: BOEING and the power of culture. *J. Intercult. Manag. Ethics Issue* **2021**, *4*, 13–26. [CrossRef]
21. Dias, M.; Teles, A.; Lopes, R. Could Boeing 737 Max crashes be avoided? Factors that undermined project safety. *Glob. Sci. J.* **2020**, *8*, 187–196.
22. Turing, A.M.; Copeland, B.J. *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life Plus the Secrets of Enigma*; Oxford University Press: Oxford, UK, 2004.
23. Turing, A.M. Computing machinery and intelligence. *Mind* **1950**, *49*, 433–460. [CrossRef]
24. Pearle, J.; Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*; Penguin: London, UK, 2018.
25. European Commission. CORDIS Results Pack on AI in Air Traffic Management: A Thematic Collection of Innovative EU-Funded Research results. October 2022. Available online: https://www.sesarju.eu/node/4254 (accessed on 27 March 2024).
26. DeCanio, S. Robots and Humans—Complements or substitutes? *J. Macroecon.* **2016**, *49*, 280–291. [CrossRef]
27. Kaliardos, W. *Enough Fluff: Returning to Meaningful Perspectives on Automation*; FAA, US Department of Transportation: Washington, DC, USA, 2023. Available online: https://rosap.ntl.bts.gov/view/dot/64829 (accessed on 27 March 2024).
28. Wikipedia on ChatGPT. 2022. Available online: https://en.wikipedia.org/wiki/ChatGPT (accessed on 27 March 2024).
29. Uren, V.; Edwards, J.S. Technology readiness and the organizational journey towards AI adoption: An empirical study. *Int. J. Inf. Manag.* **2023**, *68*, 102588. [CrossRef]
30. Defoe, A. AI Governance—A Research Agenda. Future of Humanity Institute. 2017. Available online: https://www.fhi.ox.ac.uk/ai-governance/#1511260561363-c0e7ee5f-a482 (accessed on 27 March 2024).
31. EASA. EASA Concept Paper: First Usable Guidance for Level 1 & 2 Machine Learning Applications. February 2023. Available online: https://www.easa.europa.eu/en/newsroom-and-events/news/easa-artificial-intelligence-roadmap-20-published (accessed on 27 March 2024).
32. EU Project Description for HAIKU. Available online: https://cordis.europa.eu/project/id/101075332 (accessed on 27 March 2024).
33. HAIKU Website. Available online: https://haikuproject.eu/ (accessed on 27 March 2024).
34. SAFETEAM EU Project. 2023. Available online: https://safeteamproject.eu/ (accessed on 27 March 2024).
35. Eurocontrol. *Technical Interchange Meeting (TIM) on Human-Systems Integration*; Eurocontrol Innovation Hub: Bretigny sur Orge, France, 2023. Available online: https://www.eurocontrol.int/event/technical-interchange-meeting-tim-human-systems-integration (accessed on 27 March 2024).
36. Diaz-Rodriguez, N.; Ser, J.D.; Coeckelbergh, M.; de Pardo, M.L.; Herrera-Viedma, E.; Herrera, F. Connecting the dots in trustworthy AI: From AI principles, ethics and key requirements to responsible AI systems and Regulation. *Inf. Fusion* **2023**, *99*, 101896. [CrossRef]
37. MARC Baumgartner & Stathis Malakis. *Just Culture and Artificial Intelligence: Do We Need to Expand the Just Culture Playbook?* Hindsight 35, November; EUROCONTROL: Brussels, Belgium, 2023; pp. 43–45. Available online: https://skybrary.aero/articles/hindsight-35 (accessed on 27 March 2024).
38. Kumar, R.S.S.; Snover, J.; O'Brien, D.; Albert, K.; Viljoen, S. *Failure Modes in Machine Learning*; Microsoft Corporation & Berkman Klein Center for Internet and Society at Harvard University: Cambridge, MA, USA, 2019.
39. Franchina, F. *Artificial Intelligence and the Just Culture Principle*; Hindsight 35, November; EUROCONTROL: Brussels, Belgium, 2023; pp. 39–42. Available online: https://skybrary.aero/articles/hindsight-35 (accessed on 27 March 2024).
40. Ramchum, S.D.; Stein, S.; Jennings, N.R. Trustworthy human-AI partnerships. *IScience* **2021**, *24*, 102891. [CrossRef] [PubMed]
41. European Commission. Ethics Guidelined for Trustworthy AI. *High Level Expert Group (HLEG) on Ethics and AI.* 2019. Available online: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed on 27 March 2024).
42. Lees, M.J.; Johnstone, M.C. Implementing safety features of Industry 4.0 without compromising safety culture. *Int. Fed. Autom. Control (IFAC) Pap. Online* **2021**, *54*, 680–685.
43. Macey-Dare, R. How Soon Is Now? Predicting the Expected Arrival Date of AGI-Artificial General Intelligence. 2023. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4496418 (accessed on 27 March 2024).
44. Schecter, A.; Hohenstein, J.; Larson, L.; Harris, A.; Hou, T.; Lee, W.; Lauharatanahirun, N.; DeChurch, L.; Contractor, N.; Jung, M. Vero: An accessible method for studying human-AI teamwork. *Comput. Hum. Behav.* **2023**, *141*, 107606. [CrossRef]
45. Zhang, G.; Chong, L.; Kotovsky, K.; Cagan, J. Trust in an AI versus a Human teammate: The effects of teammate identity and performance on Human-AI cooperation. *Comput. Hum. Behav.* **2023**, *139*, 107536. [CrossRef]
46. Ho, M.-T.; Mantello, P.; Ho, M.-T. An analytical framework for studying attitude towards emotional AI: The three-pronged approach. *MethodsX* **2023**, *10*, 102149. [CrossRef] [PubMed]
47. European Commission. Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence. 21 April 2021. Available online: https://data.consilium.europa.eu/doc/document/ST-8115-2021-INIT/en/pdf (accessed on 27 March 2024).

48. Veitch, E.; Alsos, O.A. A systematic review of human-AI interaction in autonomous ship design. *Saf. Sci.* **2022**, *152*, 105778. [CrossRef]
49. UK Ministry of Defence. Defense Artificial Intelligence Strategy. 2022. Available online: https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy (accessed on 27 March 2024).
50. Grote, G. Safety and autonomy—A contradiction forever? *Saf. Sci.* **2020**, *127*, 104709. [CrossRef]
51. Haddon-Cave, C. *An Independent Review into the Broader Issues Surrounding the Loss of the RAF Nimrod MR2 Aircraft XV230 in Afghanisatan in 2006*; HMSO: London, UK, 2009; ISBN 9780102962659.
52. CANSO. CANSO (Civil Air Navigation Services Organisation) Standard of Excellence in Safety Management Systems. 2023. Available online: https://canso.org/publication/canso-standard-of-excellence-in-safety-management-systems (accessed on 27 March 2024).
53. CIEHF. *The Human Dimension in Tomorrow's Aviation System*; Chartered Institute for Ergonomics and Human Factors (CIEHF): Loughborough, UK, 2020. Available online: https://ergonomics.org.uk/resource/tomorrows-aviation-system.html (accessed on 27 March 2024).
54. Dubey, A.; Abhinav, K.; Jain, S.; Arora, V.; Puttaveerana, A. HACO: A framework for developing Human-AI Teaming. In Proceedings of the 13th Innovations in Software Engineering Conference (ISEC), Jabalpur, India, 27–29 February 2020; pp. 1–9.