Article

The Impact of Artificial Intelligence on Future Aviation Safety Culture

Barry Kirwan *

Posted Date: 23 January 2024

doi: 10.20944/preprints202401.1624.v1

Keywords: Aviation; Artificial Intelligence; Safety Culture

Article

The Impact of Artificial Intelligence on Future Aviation Safety Culture

Barry Kirwan [1]

[1] EUROCONTROL, Bretigny, France, barry.kirwan@eurocontrol.int

Abstract: Artificial Intelligence is developing at a rapid place, with examples of Machine Learning already utilized in aviation to improve efficiency. In the coming decade it is likely that Intelligent Assistants (IAs) or 'Digital Colleagues' will be developed and deployed to assist aviation personnel in the cockpit, the air traffic control centre and airports. This will be a game-changer and may herald the way forward for single pilot operations, AI-based management of Urban Mobility including fleets of drones and sky-taxis, and more carbon-efficient transport across global commercial transport networks. Yet in aviation there is a core underlying tenet that 'people make safety' and keep the skies and passengers safe, based on a robust industry-wide safety culture. The introduction of IAs into aviation could potentially undermine aviation's hard-won track record in this area. This paper applies a scientifically validated safety culture tool to explore the potential impacts of the introduction of 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 organisations, to ensure safety remains a priority in the industry.

Keywords: aviation; artificial intelligence; safety culture

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. This paper therefore poses the question of how future Artificial Intelligence (AI) may impact aviation safety culture. Although AI is already in use in numerous aviation sectors via Machine Learning, this paper focuses on the more advanced AI systems likely to appear in the next decade, including Intelligent Assistants that could work semi-autonomously or even autonomously and collaboratively with human crews and teams.

The paper begins by briefly outlining safety culture today in aviation, including how it is evaluated. The fast-developing area of AI itself is then outlined, in particular 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 IAs might work together in a range of future AI settings (e.g. cockpit, air traffic Tower and Ops Room; airports). One of the principal aviation safety culture methods is then applied in detail to future Human AI Teaming scenarios to estimate whether there could be a net safety culture detriment or benefit via Human AI Teaming. 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 accident rates is generally seen as very strong, with zero fatal accidents since 2016 [1], although there are still fatal accidents in general aviation. This level of safety is in part attributed to the level of safety culture in aviation, hard won due to various accidents during the early years (1950s onwards) of commercial aviation, and despite accident 'spikes', e.g., following the introduction of 'glass cockpits' in the 1980's [2]. Since commercial aviation accidents in flight are often fatal, there is both strong public awareness of air crashes (they usually make the headlines) and appropriately demanding regulation on safety across the aviation industry spectrum [3]. Hence, safety culture has generally been a hallmark of commercial aviation in aircraft manufacturing, air traffic control, airport operations and in the airlines themselves.

Safety culture as an approach did not emerge out of the blue [4]; prior to safety culture, there was already research on *safety climate*. These two terms are often used interchangeably, but there are key differences. Most commonly, safety climate is considered to be a temporary snapshot of the current safety culture made up of perceptions and feelings at the time (and is likened to *mood*), whereas safety culture is more stable (and compared to *personality*), and related to group activities and organisational histories [5]. Safety culture is more enduring, and thus harder to change; it takes time. Whereas safety climate focuses on managerial prioritisation of safety [6] and has its roots in social psychology, safety culture has its roots in anthropology, and is more concerned with the safety-related values and practices that permeate the organisation [7]. An early definition of safety culture, coming from the nuclear power industry, is the following [8]:

> *"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 organisational culture, is as follows [4]:

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

The origins of Safety Culture are usually traced back to the Chernobyl nuclear power plant accident in 1986 [9]. Just as the Three Mile Island nuclear power plant accident in 1979 demonstrated unequivocally the importance of Human Factors in the design of human-machine interfaces in high systems, Chernobyl showed that the prevailing operational culture could catastrophically trump established safety procedures and processes. A spate of high-profile accidents from different domains (see Figure 1), including space, oil and gas, and rail – all seen as 'safety culture accidents' – only served to emphasise the enduring importance of this newly identified organisational trait. Certain high profile public enquiries into key accidents such as the Piper Alpha disaster [10] and Clapham Junction rail crash [11], as well as key safety thought leaders at the time [12, 13], and a number of accidents at least partly attributed to safety culture ever since, have ensured that safety culture has endured as a critical and non-negotiable attribute for any high risk organisation.

Figure 1 highlights the fact that safety culture wasn't initially seen as being of too much concern for aviation. This was despite the Kegworth air crash in 1989 [14], which had certain safety culture aspects. The thinking at the time was that the strong training and (Human Factors-based) design in the cockpit and air traffic operations rooms, as well as Safety Management Systems (SMS) and Standard Operating Procedures (SOPs), were sufficient. In Europe, this notion was shattered with the mid-air collision over Lake Constance in Überlingen in 2002 [15], following shortly after the Milan Linate runway collision a year earlier [16]. As Chernobyl did for nuclear, these accidents triggered a rethinking that SMS and SOPs were not enough, and that safety culture was crucial. SMS was seen as safety *competence*, comprising the key safety *processes*, whereas safety culture was the *motivation* to energise such processes for safe outcomes.

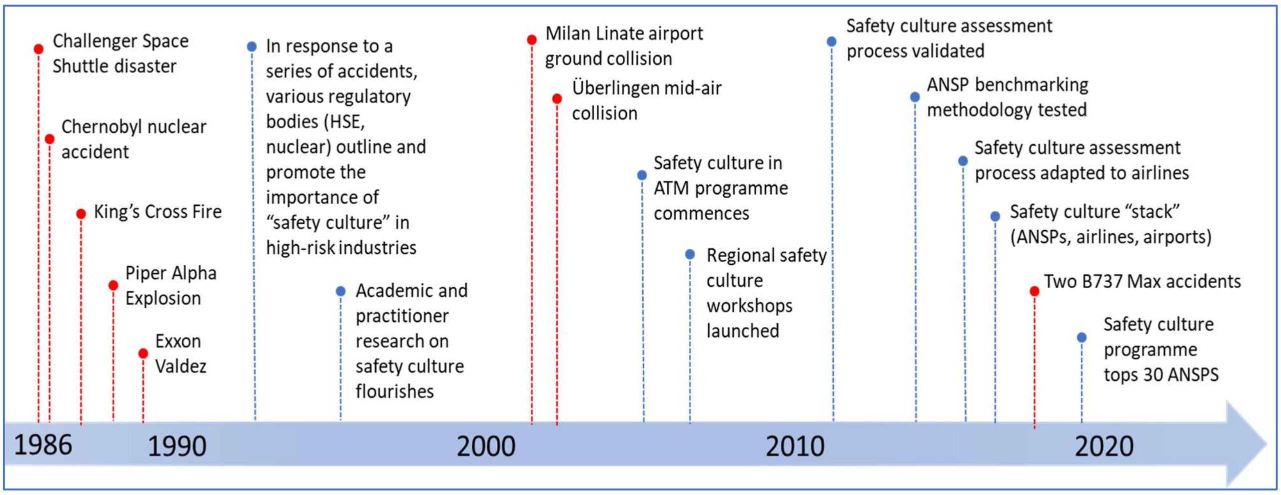
Challenger Space Shuttle disaster
Chernobyl nuclear accident
King's Cross Fire
Piper Alpha Explosion
Exxon Valdez
In response to a series of accidents, various regulatory bodies (HSE, nuclear) outline and promote the importance of "safety culture" in high-risk industries
Academic and practitioner research on safety culture flourishes
Milan Linate airport ground collision
Überlingen mid-air collision
Safety culture in ATM programme commences
Regional safety culture workshops launched
Safety culture assessment process validated
ANSP benchmarking methodology tested
Safety culture assessment process adapted to airlines
Safety culture "stack" (ANSPs, airlines, airports)
Two B737 Max accidents
Safety culture programme tops 30 ANSPS
1986
1990
2000
2010
2020

Figure 1. Safety Culture Timeline for European Aviation.

1.3. The emergence of a safety culture evaluation methodology in aviation

Since the two accidents (Uberlingen and Milan) were primarily related to air traffic management (ATM), the safety culture assurance method development was carried out in that sector of the industry, via a combined effort of EUROCONTROL and Aberdeen and LSE universities [17, 18], and was able to build upon more than a decade of experience of safety culture evaluation in other industry sectors such as nuclear power and oil and gas. A safety culture questionnaire survey 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, almost all European Air Navigation Service Providers (ANSPs) have engaged in at least one of several independent safety culture evaluations of their organisation. A total of 33 European member States have applied the method (see Figure 2), most more than once [19]. The EUROCONTROL Safety Culture Questionnaire has been scientifically validated [7, 20] and positively reviewed by the European ANSPs [21]. It has also had a more modest application in airlines [22, 23] and airports [24, 25].

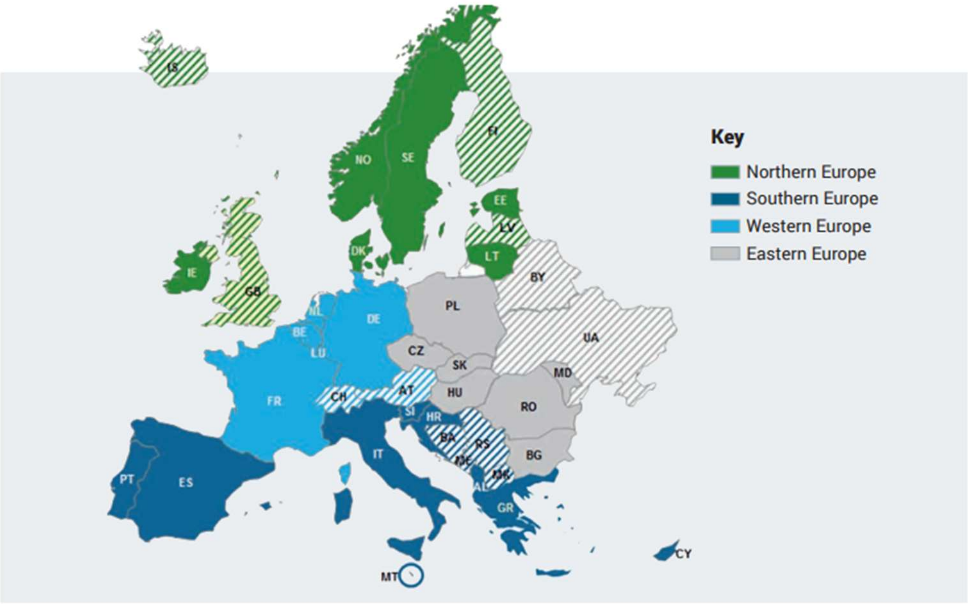

Figure 2. EUROCONTROL Safety Culture Survey Application to European ANSPs.

1.4. Measuring Safety Culture

The EUROCONTROL Safety Culture questionnaire[1] contains 48 questions linked to eight safety culture 'dimensions':

- Management Commitment to Safety
- Collaboration & Involvement
- Just Culture & Reporting
- Communication & Learning
- Colleague Commitment to Safety
- Risk Handling
- Staff and Equipment
- Procedures & Training

Each of these dimensions is supported via a set of carefully worded questions, tailored to different segments of the aviation industry (ANSP, airline, airport, and airframe manufacturer), e.g. '*My colleagues are committed to safety*,' and '*If I see unsafe behaviour by one of my colleagues, I would talk to them about it*.' The answers to these questions build a picture of the safety culture of an organisation, including an understanding of how it may differ in an organisation's various sub-cultures.

An example of the output from such a survey is shown in Figure 3 for three typical safety culture questionnaire items. In this example, the first statement is clearly reflecting '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, which means either that the respondents are not sure, don't see how it applies to their work environment, or prefer not to say (although the surveys are always anonymous and confidential, some participants are cautious). The third item has a significant negative component that would be investigated further via confidential workshops with participants (e.g. flight and cabin crew, controllers and engineers, airport workers, management and support staff), in order 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 ways forward.

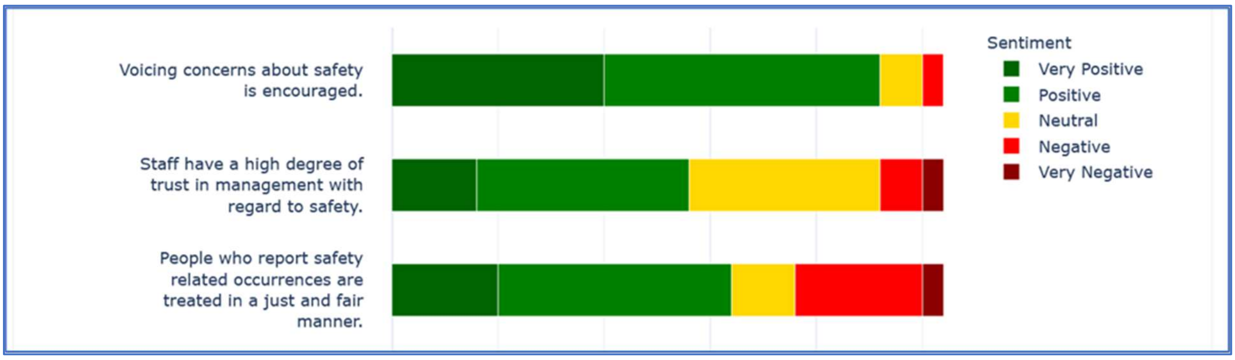

Figure 3. Example summary of responses to a safety culture survey for 3 questionnaire items.

It is useful to provide a high-level overview when carrying out safety culture surveys, and this is achieved by summarising results at the 'dimension' level using a spider-chart (because it resembles a spiderweb, e.g. as shown in Figure 4[2]. The higher the values (the further out from the centre), the better the safety culture.

[1] This particular variant is for all aviation sectors (ATM, airlines and airports); the 'official' EUROCONTROL questionnaire has slightly different wording in some items and is for ATM organisations only.

[2] 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.

doi:10.20944/preprints202401.1624.v1

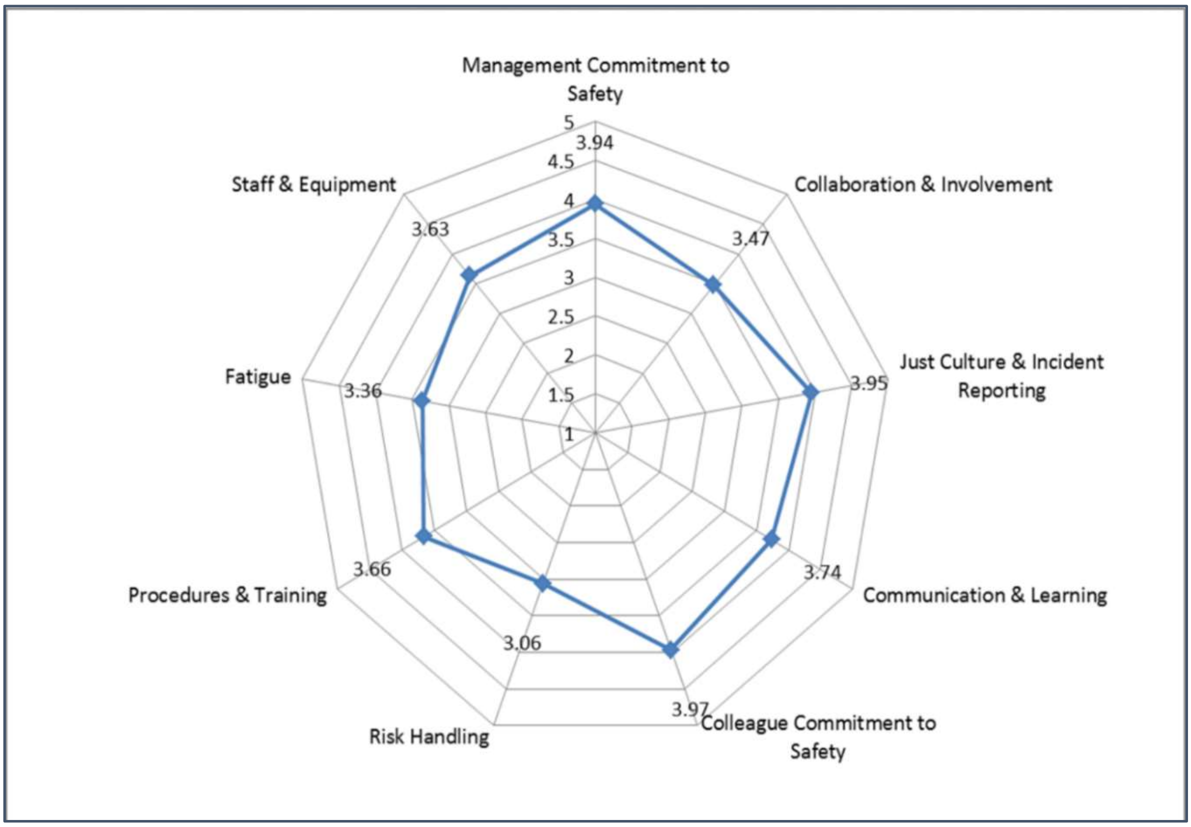

Figure 4. Example of 'Spider-Chart' view of Safety Culture Survey Results (note for this survey in the airport domain an extra dimension was added: 'fatigue').

Here it is easier to see which dimensions are the ones to worry about. In this case Risk Handling, Collaboration and Fatigue are the three dimensions that need more urgent attention. Such diagrams give a 'helicopter view' of the survey results, and are often appreciated by senior management as showing the safety culture survey 'headlines'.

Safety culture surveys offer the safety culture equivalent of a detailed health check, showing where the organisation is healthy and where it needs attention. Such reports include recommendations on how to improve, often the best ideas arising from the confidential workshops with participants, and many organisations have used such surveys to improve their safety culture [26]. CEOs often find the results of such surveys useful, 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. Indeed, the concept of safety culture necessarily includes senior and middle management, and considers the critical importance of management and designers regarding safety, based on their collective values, beliefs and behaviour, as well as the safe behaviour and attitudes of operational personnel, engineering and other support staff at 'the sharp end'. As the wide-ranging investigation into the two Boeing 737 Max accidents has shown [27, 28], 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 also 'set the tone' for the safety culture of their organisation.

1.5. Safety Culture & Future AI – an unexplored landscape

For the purposes of this paper, it is the detailed safety culture questionnaire items, rather than the dimensions *per se* that are likely to highlight where AI may impact safety culture. These are returned to following the next section, which expores the development of AI and likely future Human-AI Teaming scenarios in a range of aviation contexts. Given the above-mentioned importance of senior management, this review of AI necessarily includes aspects of AI regulation and governance.

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

i. The origins of AI
ii. AI today
iii. Generative AI
iv. Narrow AI
v. Visions of Future Aviation Human-AI Systems
vi. Trustworthy AI
vii. Accountable AI
viii. AI and Just Culture
ix. Ethical AI
x. Maintaining Human Agency for Safety
xi. AI Anthropomorphism and Emotional AI
xii. AI and Safety Governance

The consideration of these issues and perspectives helps to narrow the expansive and ever-growing field of AI research to enable a realistic focus on safety culture impacts, and in effect serves to 'ground' the analysis in Section 3. 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 discussed in Section 4.

2. The Developing Artificial Intelligence Landscape in Aviation

There is currently much 'hype' about AI, and it can be difficult to discern fact from fiction, or operational reality from the 'promise' of AI. This section therefore starts by making the key distinction between Generative AI and Narrow AI, the latter being the AI most likely to be implemented in aviation in the mid-term (now until at least 2030). Since the level of AI of interest in this paper is advanced but foreseeable for the 2030+ timeframe, there are currently many design 'degrees of freedom' being discussed, often at supra-national levels, for example with respect to ethics or whether Intelligent Assistants (IAs) should mimic emotions, etc., as well as the current preliminary regulatory perspective on AI and AI autonomy for aviation. This section then homes in on key aspects of Human-AI Teaming – including basic principles for trustworthy, accountable and ethical AI – and their potential impact on, e.g., safety citizenship. Just Culture and AI is also reviewed, as potential issues here are already identified by the aviation community, before considering the higher level issues of governance and organisational (safety) leadership.

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 [29], used to break German 'Enigma' codes used in the second world war. Such codes were seen as 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 – *running the numbers*, as it is known – and hence were 'computing' rather than thinking. 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, leading to his famous challenge to the scientific and engineering communities to develop such a machine, to be tested by 'the imitation game' [30]. He predicted that such machines would exist by the turn of the century. What is interesting is that many of the questions he posed about such artificial intelligence back in 1950 are the same questions we find ourselves posing today.

In the 1980s there was another surge in AI interest via (rule-based) expert systems, which ultimately failed to deliver operationally useful tools, in part due to their inability to account for the

doi:10.20944/preprints202401.1624.v1

experience-based and highly contextual 'tacit knowledge' that human operators possess, which usually goes far beyond what is written in procedures. The failure of expert systems arguably led to the so-called 'AI winter', which ended recently as computing power increased dramatically and Machine Learning became finally possible [31]. 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 conditions [32].

2.2. AI Today

A useful definition of AI is as follows [33]:

> "…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 in general – 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 in air traffic route networks in order to reduce aviation's carbon footprint, or to rapidly assist flight crew in finding a solution during an emergency (treading in the footsteps of the former expert systems approach, but with more speed and accuracy). Machine Learning can generally 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 sometimes unfathomable ways. Such AI tools can be seen as 'just more automation' [34], and their impact on safety culture might therefore be expected to be minimal. However, this understanding of AI – as effectively more powerful automation tool support – shifted dramatically with the release of ChatGPT in 2022 [35], heralding the advent of Generative AI.

2.3. Generative AI

ChatGPT is 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, and 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 in large part because it has used an approach called supervised learning, wherein humans work with the AI to refine the types of answers it gives, essentially training the AI to give better answers. Whether the answer makes sense is up to the user, and sometimes it produces answers that are inaccurate, wrong, or so bizarre they are referred to as hallucinations. It works best when being asked to draw together factual information that is already residing in the web, although much of the internet is not itself 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. And although a LLM is probably not the solution for operational AI in aviation, the fact that there can be generative and realistic dialogue between human and AI may pave the way one day for 'Intelligent Assistants' in the cockpit or 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. They are not. ChatGPT, despite its sometimes 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 [30, 31].

2.4. Narrow AI

Generally, 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 recognition [36]. 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) – and even though aviation has a lot of data, much of it is not shared for commercial reasons, and using 'proxy' or synthetically generated data runs the risk of 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; and
- a failure to develop the '**social capital'** required to foster such a change, such that users reject the AI tool's implementation (for example, because it threatens job losses).

Rather than generative AI, what aviation at least initially requires is what is known as **Narrow AI** [37]. 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 be focused on a specific domain or even sub-domain for which there is sufficient data for the AI to work, e.g. tens 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

At this juncture, 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 negtoiating with advanced AI systems. The European Union Aviation Safety Authority (EASA – the principal European aviation regulator) has set out a vision of AI and its potential impacts upon aviation operations and practices. EASA's recent guidance on Human-AI Teaming (HAT) [38] suggests six categories of future Human-AI partnerships:

1A – Machine learning support (already 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.

In order 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 Sagety) Project [39, 40, 41], which utilises six aviation Human-AI Teaming safety-related use cases, outlined below.

1. UC1 – a cockpit AI to help a single pilot recover from a sudden event that may induce 'startle response', and directs the pilot in terms of which instruments to focus on to resolve the emergency situation. This cognitive assistant is 1B in EASA's categorisation, and the pilot remains in charge throughout.
2. UC2 – a cockpit AI to help flight crew re-route an aircraft to a new airport destination due to deteriorating weather or airport closure, for example, taking into account a large number of factors (e.g. category of aircraft and runway length; remaining fuel available and distance to airport; connections possible for 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 2A but could be 2B.

3. UC3 – an AI that monitors and coordinates urban air traffic (drones and sky-taxis). The AI is an executive agent with a human overseer, and is actually handling most of the traffic, the human intervening only when necessary. This is category 3A.
4. UC4 – a digital assistant for remote tower operations, to alleviate 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 technques to analyse 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 develop 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.

Other advanced AI concepts include [42] 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 assistance for evidence-based training to enhance performance during adverse events.

In a recent debate on the issue of Human-AI Teaming [43], a critical threshold which could challenge the human's 'agency' for safety appeared to be category 2B – where humans and AI collaborate, and each can act to a certain extent independently – since this is different from what we have today, and could impact safety culture if safety was seen as the province of the AI, rather than the human. There are currently no AI systems in aviation today that autonomously share tasks with humans, can negotiate, make trade-offs, change priorities, and initiate 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. However, 'lesser' categories including 1B and 2A, could impact on 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 prospect 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 [44] comprises 7 technical requirements, built on 3 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 seven technical requirements proposed are as follows [44]:

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 favour of their own judgement, and then has an accident, but also the case wherein the pilot follows an 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 both unfair and unjust (such a situation is called a 'double-bind' in psychology). Accountability (and justice) would require 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 of trade-offs if any were made between different priorities, including safety. Such *data forensics* may prove inconclusive due to 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 try 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 is meant to take over in case of aberrant behaviour, but this does not appear realistic in situations where things happen suddenly and evolve quickly, as may also 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 may reduce their own situation awareness accordingly, not to mention losing certain 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 Centre or tower, such that once it is certified it is seen as 'fit for purpose', meaning that if anything goes wrong, the invisible finger of judgment will swing towards the human. Since much of aviation already has a certification 'mindset', this seems reasonably likely. But as noted elsewhere [37], ***"Certification… cannot replace responsibility."*** This means that redress is not simply a matter of putting disclaimers here and there. This relates to the afore-mentioned '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 degree of autonomy, e.g. a more autonomous AI system, which could for example 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' [44] to see how they would work in practical settings. In relation to this, but in the context of future General AI systems, ill-advised implementation of legal exemptions to absolve general AI developers of liability should be avoided, as such exemptions could unfairly shift the responsibility from large corporations to smaller actors, users and communities lacking the necessary resources, access and capabilities to effectively address and alleviate all risks [37]. This 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 seen as a cornerstone of safety culture in aviation, since it protects safety reporting and therefore enables safety learning. Just culture is defined as follows [45]:

> *"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, wilful violations and destructive acts are not tolerated."*

This definition emphasises that actions, omissions, or decisions taken by aviation professionals should be commensurate with their experience and training. This raises a key question: what, if any, formal training on AI/ML and its state-of-the-art algorithms, such as Neural Networks, and their limitations, should aviation professionals receive? A vast range of failure modes exist for ML systems [46], 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, in the context of air traffic controllers at least, it has been argued [45] that:

> *"the burden of responsibility gravitates towards the organisation to provide sufficient and appropriate training to 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 can also be noted that the existing definition of Just Culture is very human in its language, as it talks of *gross negligence*, *wilful* violations and *destructive acts*, all of which signify intent. Can any of these terms apply to AI, now, or even in the future? This is followed up in [47] 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."*

Two scenarios, reminiscent of the Double-Bind raised earlier, emphasise the issue [47]:

1. AI suggests a correct action, but the air traffic controller (ATCO) does not follow the suggestion, leading to an occurrence:

- *Is the ATCO liable for breach of duty or professional negligence?*
- *On what basis does AI suggest a 'correct action'?*
- *Does the AI follow a different standard or benchmark than the one followed by the ATCO?*
- *Does the ATCO have a duty to follow the AI's suggestions?*
- *Can AI suggestions be used as evidence?*

2. AI suggests a wrong action, and the ATCO follows the suggestion, leading to an occurrence:

- *Is the ATCO liable for breach of duty or professional negligence?*
- *Does the ATCO have an appropriate mental model about how the AI functions?*

From the aviation professional's standpoint, both scenarios seem risky. Just culture basically states that people do not usually 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 – i.e. very human – search for someone to blame, and hindsight tends to become very black and white in terms of 'the pilot should have known/done/realized…' whereas in reality, prior to the event, nothing was so black and white, and many other pilots (or controllers, or airport personnel, etc.) could have chosen the same course of action.

The problem with Human-AI Teaming and Just Culture is not simply a moral one. If aviation professionals are concerned about their accountability with respect to 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, e.g. stating that 'I did consider that the AI's advice might be incorrect,' could lead to problems for the aviation professional in a courtroom situation. If professionals stop reporting incidents, or not disclosing everything, this would be a retrograde step for aviation and open up new safety vulnerabilities in what is generally seen today as the safest mode of transportation.

According to [48], Human-AI partnerships are trustworthy by design if the humans and machines can rely on each other, self-organise 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 as yet unclear.

Scenario 2 above is exacerbated in future-envisaged situations involving 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). In both cases the AI could hypothetically be considered (at least legally) to have a certain degree of agency, and it is then a question of whether the human can detect erroneous AI behaviour (or conditions outside the AI's 'competence') 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 favourable employment conditions. In aviation, such concerns are clearly relatable to the concept of single pilot operations (SPO) in the cockpit, which could be enabled by AI at some point in the future. It is plausible that diminishing the human role could impact safety culture, because the human crew member may begin to see safety as the AI's job rather than their own, especially if the AI becomes its own 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. Preliminary ethical principles have been outlined by the European Commission's High Level Expert Group on AI (HLEG) [49]:

- *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 due to 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 particularly critical for safety culture. However, they need to be translated into workable 'good practices' in industry.

2.10. Human Agency for Safety – Maintaining Safety Citizenship

Earlier it was stated that in aviation, 'people make safety.' What this means is that aviation personnel, whether pilots, cabin crew, air traffic controllers, aeronautical engineers or airport personnel, believe that safety is core to 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 [50].

Seven factors are cited which can erode safety citizenship, all of which could be affected by AI taking on a larger share of the safety role or 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 more generally being proactive about safety (e.g. speaking up for safety), relies on a sense of self-determination [50]. 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, in an interdependent sense) in order to function in the world (self-determination theory). According to [50] this links to people's personal identity (how you feel about yourself), and their social identity (how society thinks about you and the group you belong to. e.g. "*I'm 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, and may also be therefore expected to degrade safety culture, as the human's role in the system's overall safety space diminishes.

2.11. Ex Machina – AI Anthropomorphism, and Emotional AI

Human-AI Teaming is itself an anthropomorphic term [34], 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 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, at some time 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 so in the next decade [51].

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 AI as team members [52], most participants could tell the difference between an AI and a human from the interaction, i.e. they generally guess correctly when it is an AI.

Another study [53] examined human trust in AIs as a function of the perception of the AI's identity. The authors found that AI 'teammate' *performance* matters to HAT performance and trust, whereas AI *identity* does not. The authors also 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, both 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, as well as how it 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 [54] of attitudes to **'emotional' AI**. This study is oriented more to societal impacts than industrial ones, i.e. general AI more than narrow AI applications. Nevertheless, it adds certain potential cultural impact areas into the HAT 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 no evidence of a performance benefit with emotional AIs. However, there is another question concerning whether AIs need to be aware of human emotions. Would it make sense, for example, for AI's supporting humans in an emergency to be aware of stress in the humans' voices as conditions worsen? A recent study [50] found that monitoring people's behaviour 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, but understandably people may not like this level of personal intrusion of their behaviour, bodies and data, and can lead to a feeling that the organization does not trust them.

Overall, therefore, it appears that aviation needs neither anthromorphic or emotional AI. This may sit better with safety culture, as considering the AI component of a Human-AI Team as an entity with feelings may 'muddy the waters' when it comes to 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 harmonised rules on AI [55]. The primary focus is on General AI where it intends to have strong risk-based governance on high-risk applications in society. Interestingly, the provisional agreement intends to ban, for example, cognitive behavioural manipulation, and emotion recognition in the workplace, although 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 General AI, 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 due to their interaction with natural persons or other systems. Third, ***human oversight*** must be employed to prevent or minimise safety risks that can emerge both when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse.

But 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 Organisation) and organisations such as the 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. The tone matters. If, for example, 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 organisations – both operational and manufacturing – consider AI and its role, including that of safety. If, for example, AI's capability is overestimated, such that human error is seen as the problem and AI the solution, then the industry may work towards reducing human control of 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 [56], since there can be 'Tail Effects', wherein low probability events are impractical to train AIs on, but when they occur the AI will not be able to handle them. The Maritime study suggests the need for *active* back-up control for autonomous ships (largely controlled from onshore control centres). In this scenario, human control is not necessarily decreasing. As AI autonomy goes up, passive back-up is likely to be ineffective, in part because AI can lead to 'increasing invisible interactions' such that the humans miss what is going on, in terms of the system and sub-system interactions and relationships, and are unable to understand the complexity and gain a holistic picture. The 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 technocentric) is headed:

> "*It seems counter-intuitive, then, to categorise the level of automation by degree of autonomous control gained over human control lost, when in practice both are needed to ensure safety*." [56]

Again, looking for a moment outside the aviation domain, the UK Ministry of Defence has published its own AI Strategy [57]. The strategy indeed sets 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 makes the point that human judgement will continue to be valued in future AI-enhanced defense platforms and scenarios. The paper goes on to ask 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 defence industry, therefore, is already considering how to approach AI and its potential autonomy from an organizational perspective. It is also interesting to note that they intend to focus on training middle management concerning AI. The Defence AI Strategy, significantly, also poses a set of questions around when to use AI, and when not to, which sometimes appears missing in the current rush to 'try out AI' in myriad projects in a number of 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 recent paper on organizational safety and autonomy [58] considered two models of how safety works in large organisations: safety can be seen as a centralized and hierarchical rule-based, 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 say about AI and its role in the organisation's strategy and operations, as evidenced by the importance of 'Management Commitment to Safety' in most models of safety culture.

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 tend to be exaggerated and the latter underspecified, ignored or unknown. This may mean that those aviation organisations '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 [27, 28], 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 organisations 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 report [59] 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 organisations.*"

Furthermore, on p.474: "*The ownership of risk is fragmented and dispersed, and there is a lack of clear understanding or guidance as to what levels of risk can be owned/managed/mitigated and by whom.*"

And, p.403: "*These organisational 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 organisations 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, to the extent that it is believed current SMSs can handle it. This would effectively be the 'old wine in new bottles' approach, and could lead to significant safety vulnerabilities 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 values [37]. 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 value for 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 not experience remorse or regret. Whilst AIs can mimic human behaviour and even have a built-in 'persona', this remains mimicry; they are still machines, or simply 'just more automation.' [34] A CEO might therefore wish to have 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.

In summary, governance[3] of AI, both at supra-national and corporate level, could be a critical pillar in terms of human-AI partnerships that retain aviation's strong safety record.

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

The foregoing introduction has elucidated the multi-faceted challenges posed by AI in fuure aviation systems. The remainder of the paper reviews the prospect of Intelligent Assistants in aviation through the lens of safety culure measurement, because this is how safety culture is evaluated in organisations to determine where it is working well, and where improvements are needed. Although such measurement tools – typically questionnaires – were not developed with AI in mind, the questionnaire items and dimensions – especially those relating to teamwork – can be analysed to see where and how Intelligent Assistants might affect a respondent's answers.

3.1. Materials and Method

The EUROCONTROL Safety Culture Questionnaire was utilised for this study, in particular a version that has been adapted to also accommodate airlines and airports [23-25]. This questionnaire includes the fatigue dimension which is seen as a key operational factor in airline and airport surveys in particular, but the fatigue-related questions were excluded from the current study as fatigue is not strictly speaking a safety culture dimension and, put simply, AI's do not get tired, and it is too speculative at this juncture to consider how their future use may or may not affect aviation crew and staff fatigue.

Safety culture surveys normally proceed via a large number of operational staff completing the survey anonymously, and then collating the results. Since the kinds of AI or IAs 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. The most experienced expert, who has been involved in more than 30 surveys over the past 20 years, carried out the first principal assessment. Each of the remaining 48 questionnaire items from the EUROCONTROL questionnaire was considered in the context of a future Intelligent Assistant, e.g. *commitment to safety* might reduce if the Intelligent Assistant was judged to be handling safety well and robustly, and this could affect both human operators at the 'sharp end' as well as managers running the organisation.

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 finalise the assessment, culminating in a table of key considerations for each safety culture item.

4. Results

Table 1 shows how each safety culture questionnaire item might be impacted by the presence of an Intelligent Assistant supporting a human team. Each row shows the questionnaire item, the safety

[3] For a recent general survey on potential industrial safety and security governance of AI systems, see [60].

culture dimension it relates to, the assessed impact due to 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
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 and Threat and Error Management practices, and CRM's corollary, Team Resources `Management in ATM. However, then	High

		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.	
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 also record and track how often a safety issue is raised.	Medium
B14 If I see an unsafe behaviour by a colleague I would talk to them about it.	Just culture and reporting	[See also B02] The IA can 'query' behaviour 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 could potentially become a 'snitch' for management. This would most likely be a deal-breaker for honest teamworking.	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 need 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	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

reporting and investigation process.			
C09 A staff member prosecuted for an incident involving a genuine error or mistake would be supported by the management of this organisation.	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, nor the over-riding benefits to the industry.	High
C13 Incident or occurrence reporting leads to safety improvement in this organisation.	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 actually 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 organisation.	Just culture and reporting	As for C09, an IA may be aware of 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 seen as offering solutions judged too risky, or conversely 'too safe', nullifying operational efficiency, the will need 're-training' or re-coding in some way.	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 appropriate support from safety specialists.	Staff and equipment	The IA could serve as a 'safety encyclopaedia' 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 organisation 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 a high degree of trust in management with regard to safety.	Management commitment to safety	Conversely, operational managers may simply be reluctant to allow the introduction of IAs into the system, due to both safety and operational concerns.	Medium
B10 My manager takes action on the safety issues we raise.	Management commitment to safety	See C13 above.	Low
B19 Safety is taken seriously in this organization.	Management commitment to safety	Depends on how much the IA is designed to focus on safety. The human team will watch the IA's 'behaviour' 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 priorities do not conflict may be challenging.	Medium
B22 My manager would always support me if I had a	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

concern about safety.			
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 organisation 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 organisation 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 organisation 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 quickly assess safety issues, and prepare messages to be cascaded down by senior/middle management.	Medium
B21 We learn lessons from safety-related incident or occurrence	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	High

investigations.		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.	
B24 I have good access to information regarding safety incidents or occurrences within the organisation.	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 capturing 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 are relevant to our work.	Communication	The IA could be used to store incidents, but this would not require anything 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 in an attempt 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 [70], or Threat and Error Management is used in the cockpit following an event, the AI 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 systems 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 trouble-maker if it continually questions its human team-mates' 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 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 behaviour may evolve over time, and diverge from optimum, even if it starts off safe when first implemented.	High

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 centre) 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 adjustment is made to the IA to better account for an operational circumstance that has caused safety issues, who makes this adjustment and who needs to be informed?	Medium
B18 Changes to the organisation, 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 similar to current use of AIs for continuous validation and verification of operating systems, looking for bugs or omissions. Conversely, the IA may give advice that does not make sense to the human team or the organisation, 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 as well as frequencies (highlighting common 'workarounds'). Such data could be used to identify procedures that are no	High

		longer fit for purpose, or else inform retraining requirements if the procedures are in fact still fit for purpose.	
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 emotional AI is employed), but the human can probably uilise 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 finds a way to get around the rules (human ingenuity can be used for the wrong reasons).	High
C04 The procedures describe the way in which I actually do my job.	Procedures and training	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. [This was in part why the earlier Expert Systems movement failed to deliver, leading to the infamous 'AI winter']. Also, the IA could conceivably 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
C08 I receive sufficient safety-related refresher training.	Procedures and training	The IA could take note of 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 categorisation 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 & Reporting; Risk Handling*
- **Medium Impact:** *Staff and Equipment; Procedures & Training; Communication & Learning; Collaboration & Involvement*
- **Low Impact:** *Management Commitment to Safety*

Each of these can be either considered a concern about negative impacts on safety culture that needs to be managed, or a 'safety affordance', wherein the IA could help support and possibly enhance current safety culture, safety management processes and operational safety practices. Since a number of the insights in the rows in Table 1 overlap or point to a single central issue, they were further analysed 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 Advanced AI 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 behaviour' it could be considered a 'snitch' for	The IA could serve as a living black box recorder, recording more of decision-making strategies than is the case today.

management, and may not be trusted.	
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 is able to collect and analyse day-to-day safety occurrence information it may be seen as adding objective (dispassionate) evidence and a more balanced sassessment of severity, as well as an unbiased evaluation of how close an event came to being an accident (e.g. via Bayesian analysis).
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.
The introduction of Intelligent Assistants may inexorably lead to less human staff. Although there are various ways to 'sugar-coat' this, e.g. current 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 more generally 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 trouble-maker.	The IA can upload all NOTAMs and briefings etc. so as to be able 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 really good '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 quickly assess safety issues 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, due to both safety and operational concerns.	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, thereby 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	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'.

advice, or, conversely, an IA that was giving bad advice.	
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 advice that does not make sense to the human team or the organisation, yet be unable to 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 similar to current use of AIs for continuous validation and verification of operating systems, looking for bugs or omissions.
A human risk-taker or someone who puts productivity first, may consult ('game') an IA until it finds a way to get around the rules.	The human can uilise the IA's safety advice to err on the side of caution, if she or he feels pressured to cut safety corners either due to 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 them to risk owners in organisations.

4. Discussion of Results

4.1. Safeguards and Organisational 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. These issues, whether positive or negative, can lead to safeguards to prevent safety from being diminished due to the introduction of advanced AI systems into aviation. However, for safeguards to be effective, those who can enact them need to be identified.

The various impacts are notably diversely spread across different human 'levels' in organisations; 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 useful, if simple hierarchical model of organizational safety was proposed in a project considering safety culture and safety management in aviation organisations [61], and is illustrated in Figure 5 below. This model has been used to highlight the key – but often under-explored – role of the middle management layer in safety [62].

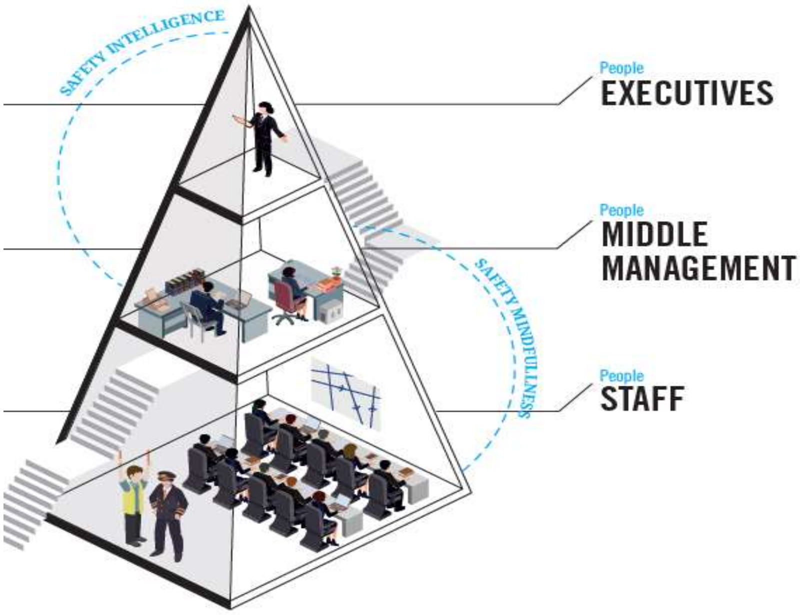

Figure 5. A Simplified Model of Safety Management [75]**.**

To this three-level model must be added an extra level, highlighting the key role of the safety department in an aviation organization, including risk modelers and incident investigators, as this is typically the hub of safety learning (local safety knowledge) for the organization, and the key working interface with external regulators. This new, four-layer model, is shown in Figure 6, with the key safety culture concerns and affordances inserted at appropriate 'risk owner' levels.

At the bottom layer are the impacts on 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 in time when considering the best course of action and any 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, quickly capture events, their precursors, signals and actions, to which the human could then add a narrative. All *NOTAM*s (Notices to Airmen – these are likely to be digitized in the near future) could be automatically uploaded into the IA, which could remind human crews if they have forgotten any aspects during operations. Similarly, the IA could be useful as a *day-to-day briefing* tool, letting the oncoming shift know of anything unusual, or changes to procedures, or the status of ongoing maintenance, etc. that has happened on previous shifts. Taken together, these nine elements could keep safety at the human's fingertips, eyes or ears, whilst guarding against simple omissions as well as reckless acts.

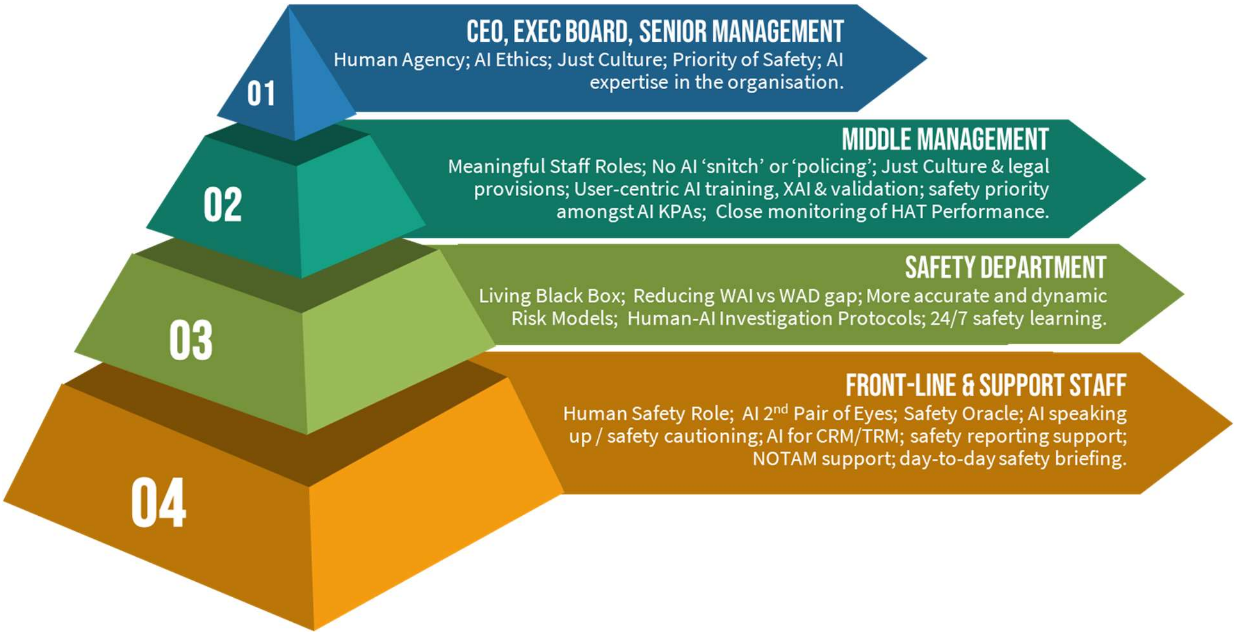
01
CEO, EXEC BOARD, SENIOR MANAGEMENT
Human Agency; AI Ethics; Just Culture; Priority of Safety; AI expertise in the organisation.
02
MIDDLE MANAGEMENT
Meaningful Staff Roles; No AI 'snitch' or 'policing'; Just Culture & legal provisions; User-centric AI training, XAI & validation; safety priority amongst AI KPAs; Close monitoring of HAT Performance.
03
SAFETY DEPARTMENT
Living Black Box; Reducing WAI vs WAD gap; More accurate and dynamic Risk Models; Human-AI Investigation Protocols; 24/7 safety learning.
04
FRONT-LINE & SUPPORT STAFF
Human Safety Role; AI 2nd Pair of Eyes; Safety Oracle; AI speaking up / safety cautioning; AI for CRM/TRM; safety reporting support; NOTAM support; day-to-day safety briefing.

Figure 6. Safety Culture Safeguards & Organisational Risk-Owners.

At the next layer is the safety department. A first safety affordance 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 be used to compare ways of working (what is actually done) against procedures and rules, not as a means of policing, 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 even unworkable in real operational conditions, or else are in need of updating as operations and technology may have moved on. Similarly, *risk models* in aviation are often seen as not really reflecting operational reality, or being at too high a level of description. The IA could record interactions in both safety-related events and 'when things go right', with such information feeding into risk models to make them more operationally relevant, and 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 for safety departments and regulators. The ultimate goal here would be that day-to-day operations are feeding *dynamic risk models* so that 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 24/7 safety monitoring and 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 with respect to safety culture & 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 roled will be the overseeing of the introduction of autonomous AI systems into the organization, ensuring that they are *user-centric* especially where '*explainablility*' of the AI's advice or decisions (XAI) to the human is concerned, and undergo *human-supervised-learning* followed by *user validation* and *Human-AI Team training* prior to operational deployment. 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, greenness, etc.), then middle management must ensure that safety retains priority when the IA is making trade-offs. Lastly, middle management must closely monitor the IA's introduction as it will evolve both in its dealings with humans and also 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 make 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 if things go wrong.

4.2. 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 future safety culture 'Conops' (Concept of Operation) for aviation organisations, i.e. a vision of how safety culture could look in the coming decade as AI autonomy rises and Intelligent Assistants enter the workspace.

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 the advent of more advanced AI systems, so that when such systems arrive, organisations will have the right theory, tools and regulatory landscape to put effective safeguards into place. Five broad research areas are accordingly listed below, aimed at bolstering cross-industry pillars of safety that could both support, and leverage organisations in their efforts to secure a stable foundation for safety and safety culture in future AI-assisted aviation:

i. **Just Culture** – if Just Culture is to be preserved, rationales and arguments need to be developed that will stand up in courts of law, that will protect crew and workers who made not only an honest mistake, but an honest (i.e. a priori reasonable) judgement about whether or not to follow AI advice, and whether or not to intervene, thus contravening AI autonomous actions seen as potentially dangerous. Such development of Just Culture argumentation and supporting principles with respect to AI and Human AI Teaming should include simulated test cases being run in 'legal sandboxes'.

ii. **Safety Management Systems (SMS)** – the key counterpart of safety culture in aviation – will also need to adapt to higher levels of AI autonomy, as is already being suggested in [38; 41]. This will probably need new thinking and new approaches, for example with respect to the treatment of HAT in risk models, rather than simply producing 'old wine in new bottles'. SMS maturity models, such as are used in air traffic organisations around the globe [63], will also need to adapt to address advanced AI integration into operations.

iii. **Human Factors** has a key role to play in the development of HAT [64], 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, literally has 'skin in the game' – is always fully represented in the determination of optimal solutions for explainability, teamworking, shared situation awareness, supervised learning, human-AI interaction means and devices, and HAT training strategies. A number of new research projects are beginning to pave the way forward. This applied research focus needs to be sustained, and a clear methodology developed for assuring usable and trustworthy Human-AI Teaming arrangements.

iv. There are currently a number of **Human -AI Teaming options** on the table, e.g. from EASA's 1B to 3A; see also [65]), with 2A, 2B and 3A offering most challenge to the human's agency for safety, and hence with most potential impact 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 result in common **design philosophies** and practices for aviation system manufacturers.

v. **Inter-Sector Collaboration** will be beneficial, whether between HAT 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 when it comes to both ethical dilemmas and high intensity, high-risk human-AI teamworking. This paper has already highlighted learning points from maritime and military domains for aviation, so closer collaboration is likely to be beneficial. **Collaboration between the transport domains** in particular makes sense, given that in the foreseeable future AIs from different transport modes will likely be interacting with each other.

7. Limitations of the Study and Further Work

This paper and its analysis are speculative in nature, 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, potentially disruptive and could well become a 'game-changer' for the industry. Nevertheless, it is intended to follow up the safety culture exercise in this paper with another one towards the end of the HAIKU project (2025), when more operational and research personnel will have been exposed to realistic simulations of future HAT systems in a broad range of use cases. This may also lead to the development of a safety culture 'mini-questionnaire' focusing solely on AI and HAT.

8. Conclusion

This paper has reviewed the current positive state of safety culture in aviation, and the future AI possibilities and challenges for the industry, in particular focusing on Human-AI Teaming envisaged for the 2030+ timeframe, wherein an Intelligent Assistant could have a moderate or high degree of autonomy. The results of a preliminary analysis of the potential impacts of future advanced AI on aviation safety culture suggests there are both significant threats and 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 aviation safety, adopting a 'wait and see' attitude is not advisable. Accordingly, a range of safeguards with associated organizational risk owners has been identified, along with more fundamental research avenues to help aviation steer a safe course through the development and deployment of advanced AI support systems. These will help ensure that aviation's hard-won level of both safety culture and safety are maintained, if not improved. A cornerstone to all of these strategies is that the human must maintain a strong safety role in aviation, whatever the AI role. It may be that at some time in the future, e.g. 2050, AI will have proven itself to be more reliable than humans, but until such time, people will continue to make safety, and should remain at the heart of safety of all aviation operations.

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 programme, 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.

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.

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

References

1. Cheimariotis, I., Stepniak, M., Gkoumas, K., Lodi, C., Marques Dos Santos, F., Grosso, M. and Marotta, A., Research and Innovation in Transport Safety and Resilience in Europe, Publications Office of the European Union, Luxembourg, 2023, doi:10.2760/951717, JRC134936.

2. Billings, C. (1996) Aviation Automation: the Search for a Human-Centred Approach. CRC Press: New York.
3. EASA (2022) Annual Safety Report. European Union Aviation Safety Agency, D-50668, Cologne, Germany. https://www.easa.europa.eu/en/newsroom-and-events/news/easa-annual-safety-review-2022-published
4. Guldenmund, F. (2018) Understanding safety culture through models and metaphors: taking stock and moving forward. In Safety Cultures, Safety Models, Gilbert, C. Journé, B., Laroche, H. and Bieder, C. (Eds). Springer Open, Switzerland.
5. Cox, S. and Flin, R. (1998) Safety Culture: Philosopher's Stone or Man of Straw? Work and Stress, 12, 189-201.
6. Zohar, D. (2010). "Thirty years of safety climate research: reflections and future directions." Accid Anal Prev 42(5): 1517-1522
7. Reader, T. W., Noort, M. C., Kirwan, B., & Shorrock, S. (2015). Safety sans frontieres: An international safety culture model. Risk Analysis, 35, 770-789.
8. Advisory Committee on the Safety of Nuclear Installations (ACSNI) Study Group. Third Report: Organizing for Safety. Sheffield: H.M. Stationery Office; 1993.
9. IAEA (1991) Safety Culture. Safety Series No.75-INSAG-4. International Atomic Energy Agency, Vienna.
10. Cullen, D. (1990) The public enquiry into the Piper Alpha Disaster. London: HMSO.
11. Hidden, A. (1989) Investigation into the Clapham Junction Railway Accident. London: HMSO.
12. Turner, R. and Pidgeon, N. (1997) Man-made disasters (2nd edition) Oxford: Butterworth-Heineman.
13. Reason, J.T. (1997) Managing the risks of organisational accidents. Aldershot: Ashgate.
14. AAIB (1990) 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, UK. https://www.gov.uk/aaib-reports/4-1990-boeing-737-400-g-obme-8-january-1989
15. Nunes, A. & Laursen, T. (2004) Identifying the factors that led to the Uberlingen mid-air collision: implications for overall system safety. Proceedings of the 48th Annual Chapter Meeting of the Human Factors and Ergonomics Society, September 20 - 24, 2004, New Orleans, LA, USA.
16. ANSV (2004) Accident Report 20A-1-04, Milan Linate Airport October 8, 2001. Agenzia Nazionale Per La Sicurezza Del Volo, 00156 Rome, January 20. https://skybrary.aero/bookshelf/ansv-accident-report-20a-1-04-milan-linate-ri
17. Mearns, K., Kirwan, B., Reader, T.W., Jackson, J., Kennedy, R., and Gordon, R. (2011) Understanding Safety Culture in Air Traffic Management. Safety Science. Development of a methodology for understanding and enhancing safety culture in Air Traffic Management. Safety Science, 53, 123-133.
18. Noort, M., Reader, T.W., Shorrock, S. and Kirwan, B. (2016) The relationship between national culture and safety culture: Implications for international safety culture assessments. Journal of Occupational and Organizational Psychology, 89, 515–538.
19. Kirwan, B. and Shorrock, S.T. (2015) A view from elsewhere: safety culture in European air traffic management. In Waterson, P. (Ed.) Patient Safety Culture, Ashgate, Aldershot, pp. 349-370.
20. Noort, M., Reader, T.W., Shorrock, S. and Kirwan, B. (2016) The relationship between national culture and safety culture: Implications for international safety culture assessments. Journal of Occupational and Organizational Psychology, 89, 515–538.
21. Kirwan, B. Shorrock, S.T. and Reader, T. (2021) The future of safety culture in European ATM – a White Paper. EUROCONTROL, Skybrary: https://skybrary.aero/bookshelf/future-safety-culture-european-air-traffic-management-white-paper
22. Reader, T., Parand, A. and Kirwan, B. (2016) European pilot's perceptions of safety culture in European aviation. Future Sky Safety Report D5.4, November. DOI: 10.13140/RG.2.2.14285.51686
23. Kirwan, B., Reader, T.W., Parand, A., Kennedy, R., Bieder, C., Stroeve, S., and Balk, A. (2019) Learning curve: interpreting the results of four years of safety culture surveys. Aerosafety World, Flight Safety Foundation, January.
24. Kirwan, B., Reader, T.W., and Parand, A. (2019) The safety culture stack – the next evolution of safety culture? Safety and Reliability, 38, 3, pp. 200-217. DOI: 10.1080/09617353.2018.1556505
25. Ogica, M., Kirwan, B., Bettignies-Thiebaux, B., Reader, T.W. and Harris, D. (2020) Harnessing inter-dependencies at an airport for safety – The Safety Stack approach. ESREL 2020 / PSAM 15, November 1-6.
26. Kirwan, B. et al (2015) CEOs on Safety Culture. A EUROCONTROL-FAA Action Plan 15 White Paper. October. DOI: 10.13140/RG.2.2.12126.10562
27. Zweifel, T.D. and Vyal, V. (2021) Crash: BOEING and the power of culture. Journal of Intercultural Management and Ethics Issue No. 4, 13-26.
28. Dias, M., Teles, A, and Lopes, R. (2020) Could Boeing 737 Max crashes be avoided? Factors that undermined project safety. Global Scientific Journals: Volume 8, Issue 4, April, Online: ISSN 2320-9186.
29. Turing, A.M. and Copeland, B.J. (2004) The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life plus The Secrets of Enigma. Oxford University Press: Oxford.

30. Turing, A.M. (1950) Computing machinery and intelligence. Mind, 49, 433-460. https://doi.org/10.1093/mind/LIX.236.433
31. Pearle, J. and Mackenzie, D. (2018) The Book of Why: the new science of cause and effect. Penguin: London.
32. European Commission (2022) CORDIS Results Pack on AI in air traffic management: A thematic collection of innovative EU-funded research results. October 2022. https://www.sesarju.eu/node/4254
33. DeCanio, S. (2016) Robots and Humans – complements or substitutes? Journal of Macroeconomics, 49, 280-291.
34. Kaliardos, W. (2023) Enough Fluff: Returning to Meaningful Perspectives on Automation. FAA, US Department of Transportation, Washington DC. https://rosap.ntl.bts.gov/view/dot/64829
35. Wikipedia on ChatGPT (2022) https://en.wikipedia.org/wiki/ChatGPT
36. Uren, V and Edwards, J.S. (2023) Technology readiness and the organizational journey towards AI adoption: an empirical study. Int J of Information Management, 68, 102588.
37. Defoe, A. (2017) AI Governance » A Research Agenda. Future of Humanity Institute. https://www.fhi.ox.ac.uk/ai-governance/#1511260561363-c0e7ee5f-a482
38. EASA (2023) EASA Concept Paper: first usable guidance for level 1 & 2 machine learning applications. February. https://www.easa.europa.eu/en/newsroom-and-events/news/easa-artificial-intelligence-roadmap-20-published
39. Kirwan, B. (2023) The Future Impact of Digital Assistants on Aviation Safety Culture. In Human Interaction and Emerging Technologies (IHIET-AI 2023): Artificial Intelligence and Future Applications. Ahram, T., and Taiar, R. (Eds). Volume 70, pp. 77-87. AHFE International, Lausanne, Switzerland, April 13-15. http://doi.org/10.54941/ahfe1002922
40. https://cordis.europa.eu/project/id/101075332 EU Project Description for HAIKU.
41. https://haikuproject.eu/ HAIKU Website.
42. SAFETEAM EU Project (2023) https://safeteamproject.eu/
43. https://www.eurocontrol.int/event/technical-interchange-meeting-tim-human-systems-integration [Day 2]
44. Diaz-Rodriguez, N., Ser, J.D., Coeckelbergh, M., de Pardo, M.L., Herrera-Viedma, E., and Herrera, F. (2023) Connecting the dots in trustworthy AI: from AI principles, ethics and key requirements to responsible AI systems and Regulation. Information Fusion, 99, 101896.
45. MARC Baumgartner & Stathis Malakis (2023) Just Culture and Artificial Intelligence: do we need to expand the Just Culture playbook? Hindsight 35, November, pp43-45. https://skybrary.aero/articles/hindsight-35
46. Kumar, R.S.S, Snover, J, O'Brien, D., Albert, K. and Viljoen, S. (2019) Failure modes in machine learning. Microsoft Corporation & Berkman Klein Center for Internet and Society at Harvard University. November.
47. Franchina, F. (2023) Artificial Intelligence and the Just Culture Principle. Hindsight 35, November, pp.39-42. https://skybrary.aero/articles/hindsight-35
48. Ramchum, S.D., Stein, S. and Jennings, N.R. (2021) Trustworthy human-AI partnerships. iScience, 24, 102891, August. CelPress.
49. European Commission (2019) Ethics Guidelined for Trustworthy AI. High Level Expert Group (HLEG) on Ethics and AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
50. Lees, M.J. and Johnstone, M.C. (2021) Implementing safety features of Industry 4.0 without compromising safety culture. International Federation of Automation Control (IFAC) Papers Online, 54-13, 680-685.
51. Macey-Dare, R. (2023) How Soon is Now? Predicting the Expected Arrival Date of AGI- Artificial General Intelligence. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4496418
52. Schecter, A., Hohenstein, J., Larson, L., Harris, A., Hou, T., Lee, W., Lauharatanahirun, N., DeChurch, L, Contractor, N. and Jung, M. (2023) Vero: an accessible method for studying human-AI teamwork.
53. Zhang, G., Chong, L., Kotovsky, K. and Cagan, J. (2023) Trust in an AI versus a Human teammate: the effects of teammate identity and performance on Human-AI cooperation. Computers in Human Behaviour, 139, 107536.

54. Ho, Manh-Tung, Mantello, P. and Ho, Manh-Toan (2023) An analytical framework for studying attitude towards emotional AI: The three-pronged approach. In MethodsX, 10, 102149.
55. European Commission (2021, April 21). Proposal for a regulation laying down harmonised rules on artificial intelligence. https://data.consilium.europa.eu/doc/document/ST-8115-2021-INIT/en/pdf
56. Veitch, E. and Alsos, O.A. (2022) A systematic review of human-AI interaction in autonomous ship design. Safety Science, 152, 105778.
57. UK Ministry of Defence (2022) Defence Artificial Intelligence Strategy. https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy
58. Grote, G. (2020) Safety and autonomy – a contradiction forever? Safety Science, 127, 104709.
59. Haddon-Cave, C. (2009) An Independent Review into the Broader Issues Surrounding the Loss of the RAF Nimrod MR2 Aircraft XV230 in Afghanisatan in 2006. HMSO: London. ISBN: 9780102962659.
60. https://www.governance.ai/post/broad-expert-consensus-for-many-agi-safety-and-governance-best-practices

61. Stroeve, S. Smeltink, J.; Kirwan, B. (2022) Assessing and Advancing Safety Management in Aviation. Safety, 8, 20. https://doi.org/10.3390/safety8020020
62. Callari, T.C., Bieder, C. and Kirwan, B. (2019) What is it like for a Middle Manager to take Safety into account? Practices and Challenges. Safety Science, 113, 19-29.
63. CANSO (2023) CANSO Standard of Excellence in Safety Management Systems https://canso.org/publication/canso-standard-of-excellence-in-safety-management-systems
64. Kirwan, B., Charles, R., Jones, K, Wen-Chin, L., Page, J., Tutton, W., and Bettignies-Thiebaux, B. (2020) The human dimension in tomorrow's aviation system. CIEHF, September 23. https://ergonomics.org.uk/resource/tomorrows-aviation-system.html
65. Dubey, A., Abhinav, K., Jain, S., Arora, V., and Puttaveerana, A. (2020) HACO: A framework for developing Human-AI Teaming. Proceedings of the 13th Innovations in Software Engineering Conference (ISEC), Article 10, pages 1-9, February.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.