Article

Not peer-reviewed version

Human Factors Requirements for Human-AI Teaming in Aviation

Barry Kirwan *

Posted Date: 13 January 2025

doi: 10.20944/preprints202501.0974.v1

Keywords: Aviation; Artificial Intelligence; Aviation, Interface Design.; Human Factors; Human-AI Teaming

 doi:10.20944/preprints202501.0974.v1

Article

Human Factors Requirements for Human-AI Teaming in Aviation

Barry Kirwan

EUROCONTROL, Bretigny, France, barry.kirwan@eurocontrol.int

Abstract: The advent of Artificial Intelligence in the cockpit and the air traffic control centre in the coming decade could mark a step change improvement in aviation safety, or else could usher in a flush of 'AI-induced' accidents. Given that contemporary AI has well-known weaknesses, from data biases and edge or corner effects, to outright 'hallucinations', in the mid-term AI will almost certainly be partnered with human expertise, its outputs monitored and tempered by human judgement. This is already enshrined in the EU Act on AI, with adherence to principles of human agency and oversight required in safety-critical domains such as aviation. However, such sound policies and principles are unlikely to be enough. Human interactions with current automation, let alone future AI, already require extensive requirements, methods and validations to ensure a robust (accident-free) partnership. Since AI will inevitably push the boundaries of traditional human-automation interaction, there is a need to revisit Human Factors to meet the challenges of future human-AI interaction design. This paper briefly reviews the origins of AI and its current 'landscape', together with the evolution of Human Factors, to identify the critical areas where Human Factors can aid future human-AI system performance and safety. The result is a detailed requirements set organised into eight Human Factors areas, from Human-Centred Design to Organisational Readiness. The requirements set is scalable to different design maturity levels and different levels of AI autonomy. The application of the requirements is illustrated via a cockpit Human-AI Teaming use case.

Keywords: aviation; artificial intelligence; aviation; interface design

1. Introduction

1.1. Overview of Paper

The first part of this paper is principally a review of Artificial Intelligence (AI), from its origins to its status today, and its current and likely mid-term usage in aviation settings. The second part tracks the evolutionary trajectory of Human Factors as it developed to focus both on cognition and automation in aviation, aiming to form a safe and effective human-automation partnership in workplaces such as cockpits and air traffic control centres and towers. The third part of this paper derives a detailed set of Human Factors Requirements for future AI-based systems in which there will be a working partnership – task and goal-sharing – between humans and AI. The fourth part illustrates the application of the requirements set to an aviation research use case. The fifth and final section concludes on the utility of the approach and the most urgent Human Factors research needed to prepare for more advanced AI systems yet to come.

1.2. AI – the new normal or déjà vu?

Artificial Intelligence (AI) may feel new to many, but it has a relatively long history as illustrated in Figure 1, dating back to the 1950s [1, 2] and arguably even back to the 19th Century via Charles Babbage's counting machines [3]. AI initially had the goal of replicating certain human cognitive abilities, albeit with more reliability. Hence the fundamental notion of computation, of 'crunching the numbers', which humans could do given enough time, but rarely error-free. As computing power grew, it soon surpassed what humans could do even given infinite time, and with the advent of deep

 doi:10.20944/preprints202501.0974.v1

learning [4], it could sometimes derive novel solutions to problems that we would never have thought of. Thus, the goal of AI shifted from replicating human cognitive capabilities to surpassing them.

The AI timeline in Figure 1 shows that many inventions or initiatives believed generally to be relatively recent, have in fact been around for a long time, including robots, natural language processing, neural networks, computer vision and even self-driving cars. However, many of these were prototypes, and did not go 'mainstream' until recently (robot assembly for car manufacture being a notable exception) [5].

As Figure 1 also shows, AI has already experienced two 'AI winters', where its promise vastly exceeded its delivery, leading to an investment and technological cliff edge, so that AI largely disappeared for a while from the public eye [6, 7]. But as computing power increased by orders of magnitude, and as Machine Learning approaches began to show their worth and earn their keep, AI once again caught the public's attention, as well as attracting both brilliant minds and eye-watering levels of investment. The public release of ChatGPT [8] and other Generative AI (hereafter called GenAI) systems in 2022 transformed AI from being a technophile, jargon-imbued subject to being a workplace and even household commodity. Even if most people have little idea of how AI works, they know that it can do things for them, whether helping their research, finishing or finessing a report, or providing a diagnosis of their health symptoms. Though many have experienced first-hand the errors and biases of GenAI systems, they accept the trade-off between its accessibility and instant power to answer questions and requests, against the occasional inaccuracies or plausible fabrications called hallucinations [9].

Will this current AI phase result in yet another AI winter, or will the increasingly ubiquitous usage of AI become the new normal? There are already signs that GenAI is reaching its limits and not going as far as had been hoped, resulting in a minor decline in the massive investment seen in the past couple of years, But for many applications it is likely to be 'good enough', and recent research suggests that Large Language Models (LLMs) are performing better in certain areas, e.g. ChatGPT 4.0 vs ChatGPT 3.5 in the medical domain [10]. Doubtlessly, new checks and balances need to be put in place to counter the inevitable quality gaps in its output, including knowing how to ask questions to get robust answers [11], and how to integrate AI more effectively into 'messy' organisational settings such as hospitals [12], but its immediacy as a generalised resource will be too appealing for many to give up. Although the long term societal and cultural ramifications of this shift to an AI-assisted society are yet to emerge, AI is therefore likely here to stay.

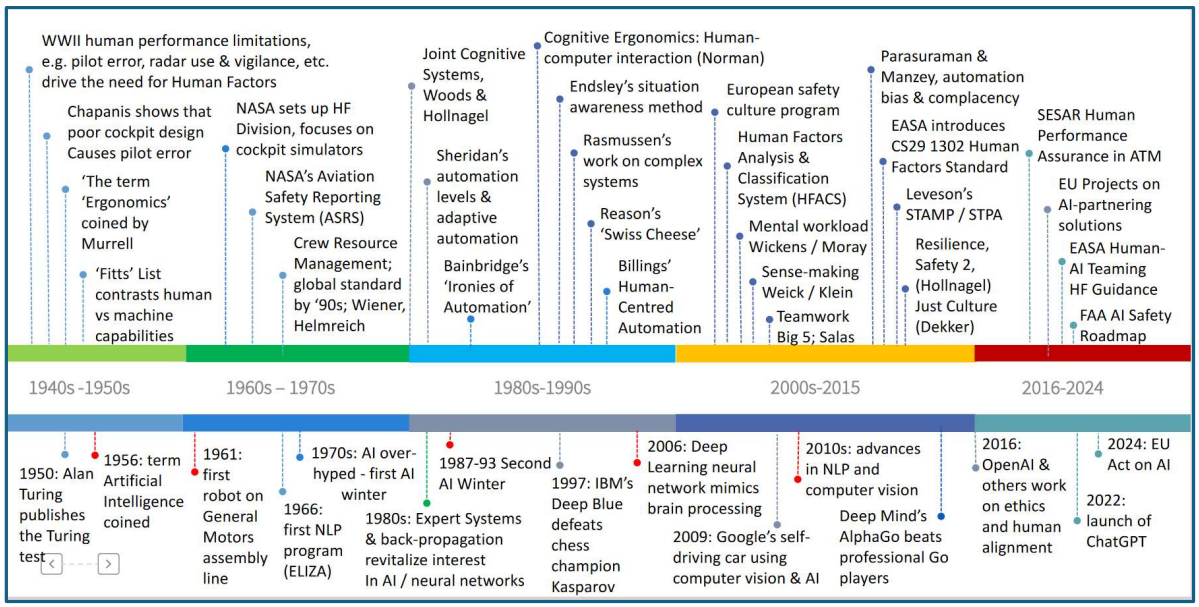

Figure 1. Timeline of Parallel Evolution of AI (lower section) and Human Factors (upper section).

 doi:10.20944/preprints202501.0974.v1

1.3. The Current AI Landscape

A definition of Artificial Intelligence pertinent to this paper is as follows [13]:

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

Most AI systems today represent what is known as Narrow AI [14], namely AI-based systems and services focused on a specific domain such as aviation, typically using Machine Learning (ML). In 'normal' programming, e.g. for automation, a machine is programmed exactly what to do and, given stable inputs, that is exactly what it will do. The code may be complex, but is completely explainable, at least to a software analyst or data scientist.

Machine Learning is different, because it learns autonomously, e.g. via statistics and mathematical programming [4] to analyse data sets often called 'training data', where the data have been 'labelled' by humans ('supervised learning'). Some ML approaches do not require labelled data (unsupervised learning), and these may be even more opaque to human scrutiny. In essence, the ML system learns through experience with 'training data' (and/or human supervision) and gradually becomes better at whatever task it has been set to do. The 'code', or the way the ML works, is still explainable, but probably only to a data scientist. ML is used to extract knowledge or insights from very large amounts of data, often to then be used for operational purposes or for forecasting events e.g. aircraft diversion requirements around thunderstorms [15].

Deep Learning [4] goes a step further, because the AI finds its own statistical or algorithmic 'reasoning' to develop a solution, based on a lot more data, and data that are unlabelled. Deep Learning therefore does not require human supervised learning. The more levels there are in Deep Learning, or hidden layers' in neural networks, the more unfathomable such systems become.

Machine Learning can develop models and predictions that perhaps, given enough time, humans could do. Deep Learning is different and can come up with solutions humans likely would never think of. That said, Deep Learning typically uses artificial neural networks, themselves inspired by the way human brains work. Deep Learning is used for some of the more complex human cognitive processes we take for granted, such as natural language processing [16] and image recognition [17], and also for tackling complex problems such as finding cures for intractable diseases [18] (see [19] for a general summary of contemporary AI application areas).

Machine learning and Deep Learning are both seen as **'Narrow AI'** [14], as their focus is narrowly defined in a domain or sub-domain. As such, Narrow AI supports humans in their analysis, decision-making and other tasks. In cases where tasks are well-specified and predictable, it can execute its functions without human intervention, and with minimal supervision.

The hallmark of **Generative AI** (GenAI) tools, such as ChatGPT, Google Gemini and DALL-E, are that they can create new content that is often indistinguishable from human-generated content. They utilise deep learning neural networks trained on vast datasets, and natural language processing to render interaction with human users smoother. Large Language Models (LLMs) like ChatGPT can respond instantly to any query. Whether the response is a valid or correct one is another matter.

The point about ML, Deep Learning and GenAI is that these systems are still 'crunching the numbers', and in the case of LLMs they are predicting the next word in a sequence. Nowhere is there understanding, or a mind, or thought. They may be very useful, but they are all essentially 'idiot savants' [14]. The problem is that particularly with advanced LLMs, it can feel to the user as if they are interacting with a person. This can matter in a safety critical environment and is returned to later in this paper.

Artificial General Intelligence (AGI) does not yet exist but is predicted to emerge in the coming decades, e.g. by 2041 [20]. It would effectively comprise a mind capable of independent reasoning and could therefore in theory attain sentience. AGI would be able to set its own goals, and its intelligence could grow very rapidly to eclipse that of human beings [21]. Whilst the step from GenAI to AGI may seem small, given the way LLMs such as ChatGPT respond instantly and can summarise vast swathes of knowledge, in practice the step is huge. AGI would likely require a different

 doi:10.20944/preprints202501.0974.v1

computing architecture, and it is not known how to inculcate a sense of self (putting aside for the moment the question of why we would want to do that), which may be a prerequisite for AGI, since reflexive thought is a hallmark of intelligence. For the rest of this paper, therefore, AGI is ignored, as it may never be realised.

What could exist in a relatively short timeframe are '**Intelligent Assistants'** (IAs) or 'Digital Colleagues', as envisaged by the concept of Human-AI Teaming, also called Human Autonomy Teaming, both using the same acronym HAT [see 22, 23, 24]. The essential nature of HAT, which differentiates it from automation, is the idea of an IA having a degree of autonomy/agency such that it can have intent, form goals and decisions, and execute such decisions, in collaboration with a human agent. Such IAs therefore could, in the coming decade, collaborate with and converse with pilots and air traffic controllers in operational settings. These IAs are still Narrow AI (ML, including Deep Learning) with or without Natural Language Processing capability. They would not be GenAI, at least not now in the current legal framework in Europe given the EU Act on AI [25] and would certainly not constitute AGI. Early examples of HAT prototypes include cockpit support for startle response [26], determining safe alternate airports due to severe weather degradation [22]; delivering air traffic control sector workload prognoses [23]; cockpit management of unstable approaches [27, 28]; single pilot operations [29]; and ATC support for landing/arrival sequencing [30]. For a general overview of review of recent research in AI assisting air traffic operations see [31].

1.4. AI Levels of Autonomy – the European Regulatory Perspective

Given that Intelligent Assistants (IAs) are intended to interact and collaborate with humans in aviation contexts, there is clearly an increase in their autonomy compared to automation simply presenting information and warnings. This shift in degree of autonomy affects the relationship between the human and the AI-based automation in two ways. First, the information or advice, or even executive action, can be based on calculations that are opaque to the end users (e.g. pilots), because the level of complexity and transparency of how AIs derive their answers mean that no amount of theoretical training for pilots will enable them to follow the AI's 'reasoning', unless an additional layer of such 'explainability' is afforded to the pilot by the AI-based automation. The pilot must therefore come to trust the IA, or its advice will be rejected. Second, the role of the pilot is affected, because currently the pilot is always in command and ready (i.e. trained) to take over in case the automation (e.g. automatic landing) fails. The fundamental notion of collaboration suggests an inter-dependence, that such control becomes to a greater or lesser degree shared between human and AI.

It is therefore useful to map the degree of AI autonomy, and autonomy sharing between human and AI. Increasing levels of autonomy can be represented on a scale, and the most influential scale in aviation currently is that provided by the European Union Aviation Safety Agency (EASA). Recent guidance on Human-AI Teaming (HAT) from the EASA [32] envisages six categories of future Human-AI partnerships, illustrated in Figure 2:

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)

 doi:10.20944/preprints202501.0974.v1

3B – the AI is running everything, and the human cannot intervene.

Figure 2. EASA Human-AI Teaming Categories (as interpreted by the HAIKU project).

It has been argued that AI innovation, for all its benefits, is essentially 'just more automation' supporting the human operator [33]. The critical threshold in AI autonomy where this may no longer hold appears to be between category 2A and 2B [34, 35], since this is different from what we have today in civil aviation cockpits. The distinction between 2A and 2B is clarified by EASA as follows, including what each one is and is not [35]:

- Cooperation Level 2A: cooperation is a process in which the AI-based system works to help the end user accomplish his or her own goal. The AI-based system works according to a predefined task allocation pattern with informative feedback to the end user on the decisions and/or actions implementation. The cooperation process follows a directive approach. Cooperation does not imply a shared situation awareness[1] between the end user and the AI-based system. Communication is not a paramount capability for cooperation.
- Collaboration Level 2B: collaboration is a process in which the end user and the AI-based system work together and jointly to achieve a predefined shared goal and solve a problem through a co-constructive approach. Collaboration implies the capability to share situation awareness and to readjust strategies and task allocation in real time. Communication is paramount to share valuable information needed to achieve the goal.

There are currently no [AI-based] civil aviation systems that autonomously share tasks with humans, and can negotiate, make trade-offs, change priorities and initiate and execute tasks under their own initiative. Even for 'lesser' autonomy levels such as 1B to 2A, and also for 3A, there is the opacity issue, wherein the AI's are akin to 'black boxes', and as such may surprise, confound or confuse the end users, because most end users will not be able to follow the computations underpinning AI advice. This leads to the *raison d'etre* of this paper, namely that AI and IAs represent something novel, and as such, conventional automation interface design and Human Factors approaches may not be sufficient to assure the safe use and acceptability of such systems. At the very least, issues such as roles and responsibilities, trust in automation and situation awareness will have

[1] Shared situation awareness involves the ability of both humans and AI systems to gather, process, exchange and interpret information relevant to a particular context or environment, leading to a shared comprehension of the situation at hand [35]

 doi:10.20944/preprints202501.0974.v1

additional significant nuances that may require augmentation of existing approaches. Furthermore, new issues such as operational explainability of AI systems may require completely new approaches or design requirements.

In order to determine what is needed from Human Factors, it is useful to briefly review Human Factors as a whole, over its seventy years history, to see how it has developed into its current capability set, in order to better understand what needs to be added to accommodate new AI-based systems and, in particular, Intelligent Assistants.

2. The Human Factors Landscape

In order to consider what is required from Human Factors for future AI-based systems, it is useful to briefly chart the evolution of Human Factors in aviation, including Human Factors approaches to increasing automation in the cockpit and elsewhere. This helps show where we are now, and where we may have to go. It is also helpful as some of the stages Human Factors has passed through over the past seventy years may be worth revisiting and updating to account for AI interactions.

2.1. The Origins of Human Factors

WWII was arguably the forge from which human performance partnering with technology first came to the fore. It was noted that many aircraft losses were due to 'pilot error', and that the new role of radar operators was susceptible to vigilance failures due to low signal-to-noise ratios. It is important to appreciate that prior to Human Factors, it was generally believed that mistakes and errors could be avoided by training, professionalism, motivation and strength of character. Yet even when these were in abundance, WWII showed resoundingly that 'good men' can make simple mistakes when interacting with technology, that cost them and others their lives. There was a pressing need for a way to avoid high attrition rates due to 'human error'. Drills and drums on the battlefield, when rifles and bayonets were the main weapon, were no longer sufficient when warfare involved complex machinery in the form of aircraft fighters and bombers, warships, submarines, and radar control stations.

During the war Alphonse Chapanis, a lieutenant in the army, showed that much of pilot error related to cockpit design (e.g. locating the landing gear control next to the bomb release control, resulting in a number of tragic mishaps on landing). His work showed how better design could reduce pilot error [36, 37, 38]. Similarly, studies of radar operators revealed that despite the best motivation, signals would be missed if that radar operators worked beyond certain periods. The lessons learned from human performance under sometimes extreme stress were brought together under a new domain known as Ergonomics in the UK [36], and Human Factors or Human Factors Engineering in the US [38], denoting the study of humans at work. The term Ergonomics is derived from the Greek word for work, *ergon*. Ergonomics / Human Factors primarily constitutes a fusion of three disciplines: psychology, physiology and engineering.

2.2. Key Waypoints in Human Factors

Human Factors and Ergonomics have always focused largely on human work and the workplace. As highlighted in Figure 1, Human Factors started out with a focus on physical ergonomics and on the layout of cockpit instruments and controls, for example. By the 1980s and 1990s the focus had shifted to cognitive ergonomics, in parallel with the fact that many work situations involved computing and automation. As work complexity grew, constructs such as situation awareness and mental workload came to the fore, and as automation became the norm, there was a corresponding need to consider complacency and bias, as well as a move from a focus on human error to system resilience. Arguably these developments over time have prepared Human Factors for the next major step change in the guise of the rise of AI.

 doi:10.20944/preprints202501.0974.v1

This section therefore details key milestones and themes in Human Factors over the past seven decades that have helped application domains such as aviation achieve and sustain very high levels of safety. These are the principal focal areas and capabilities of Human Factors that need to be preserved, revisited (or even resurrected), or updated in the context of AI, and will form the basis of a comprehensive Human Factors Requirements set for AI-based systems. Each milestone is outlined below, with one or more key references, along with the implications for Human Factors assurance of future Human-AI Teaming systems in aviation.

2.2.1. Fitts' List

One of the first landmark achievements was the development of a contrasting list of what machines are good at, versus what humans are good at. This was developed by Paul Fitts and is perennially known as Fitts' List [39, 40] (see Figure 3). Notably at the time, most of the cognitive 'heavy lifting', including pattern recognition and interpretation of 'noisy' data was left to the human. A more recent analysis [41] showed that the distinction between man and machine's relative capabilities has shifted or at least blurred. Given the current and intended prowess of AI-based systems, it would be useful to update this list given the advent of AI, and clarify where humans are still preferred, where AI should 'go it alone', and where AI-augmented human performance is the optimal solution.

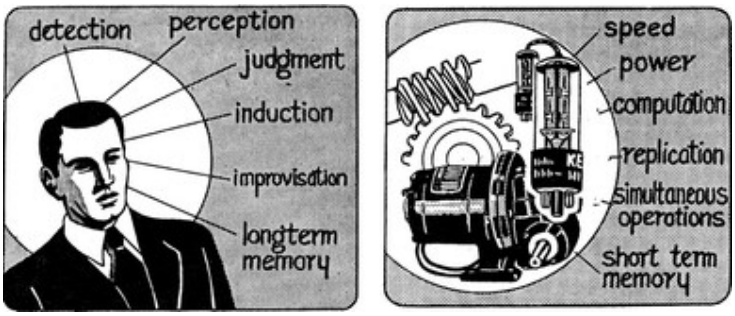

Figure 3. – Image of Fitts' List.

2.2.2. Aviation Safety Reporting System

Human Factors was given a boost when NASA created a Human Factors group to contribute to the Apollo missions. At around that time the Aviation Safety Reporting System (ASRS) was set up by the FAA and NASA (https://asrs.arc.nasa.gov/). ASRS collected data on safety-related events (e.g. mistakes made by pilots), guaranteeing pilots immunity (with a few exceptions) from prosecution if they reported. ASRS was possibly a stroke of genius, as it provided a constant wellspring of viable information about what was not working in the cockpit or on the ground. ASRS created a continuous feedback loop that has run in parallel with the inexorable increase of the role of technology in the cockpit.

If (when) AI begins to be used operationally in the cockpit or ATC Ops Room or Control Tower, the immunity issue (strongly related to Just Culture – see later) may need to be revisited, as AI-based systems may automatically report anomalies or deviations from the norm. This notion is already embedded in incoming law in Europe, via the EU Act on AI, for safety-critical settings. ASRS and equivalent systems (e.g. ECCAIRS in Europe [https://aviationreporting.eu/en]) would also likely need additional categories to capture AI-related characteristics of events, and errors that arise from Human-AI Teaming operations (see HFACS, later).

2.2.3. Crew Resource Management

Crew Resource Management [42] (originally called Cockpit Resource Management), was to become one of the major contributions of Human Factors to aviation safety. CRM gave flight crews

 doi:10.20944/preprints202501.0974.v1

the means to become more resilient against human errors and teamwork problems, and complemented ongoing work to make the cockpit design itself less error prone. The need for CRM grew from the world's worst civil aviation air disaster at Tenerife airport in 1977 but was also linked to United Airlines Flight 173 air crash in 1978 [43]. Both accidents highlighted team and communication errors, and the need for specific training on leadership, decision-making and communication on the flight deck, and with air traffic control (which in Europe later developed its own version of CRM called Team Resource Management or TRM [44]). CRM focuses on proper use of the available human resources in operational teams and has spread to other domains such as maritime via Bridge Resources Management (BRM) or Maritime Resources management (MRM) [45]. CRM remains strong today, and is currently in its 6th generation, and continues to be a mainstay in aviation Human Factors and aviation safety.

The resilience of flight crews and air traffic controllers in the face of abnormal, challenging, time-critical situations is essential to safety in today's complex, high-tempo operations. If AI-based systems begin to play a role in either cockpit or ATC workplaces, or significantly support flight crew/ATCOs, potential impacts on CRM/TRM need to be understood and safeguards put in place. This will be particularly the case with Intelligent Assistants at EASA category 2A or 2B, where flight crew and the IA are sharing tasks and even negotiating over how best to achieve goals.

2.2.4. Human-Centred Design & Human Computer Interaction

The 1980s heralded some of the key paradigms of Human Factors, and is generally considered to be the epoch wherein Human Factors and Ergonomics became more 'cognitive' in their approach. Norman's landmark work on design [46] reflected the ubiquitous rise of computer usage in aviation and other industrial domains, and ushered in a lasting focus on human-computer interaction (HCI) and the benefits of human-centred design (HCD) [47]. It also contributed to the development of the broader and still flourishing field of Usability [48].

Human-AI interaction, whether via keyboard, speech or other media, will be critical for the safe and effective introduction of AI-based systems into operational aviation contexts. A superordinate Human-Centred Design approach to the development and validation of Human-AI Teaming solutions would facilitate the principles of both human agency and human oversight [49]. Many HCI and Usability requirements will probably still apply to AI-based interfaces and interaction, with the proviso that new requirements may well be needed for a human-AI speech interface via natural language processing (NLP) as well as operational explainability (OpXAI) to help the user understand the AI's decision-making processes [50]. Explainability is likely to be a major new area for Human Factors research.

2.2.5. Joint Cognitive Systems

Joint Cognitive Systems (JCS) [51] and Cognitive Systems Engineering (CSE) [52] introduced the concept of a cognitive system as an adaptive system that functions using knowledge about itself and the environment in the planning and modification of actions. CSE sought to side-step the man-vs machine paradigm and consider the cognitive output of both working together. With its focus on the joint cognitive picture emerging from the agents in the system (man/machine), it paved the way for Situation Awareness (see later). Its focus on examining work *in situ*, rather than in the lab, and its focus on more naturalistic examination of decision-making scenarios (aka naturalistic decision-making [53]), also paved the way for Safety II and Resilience movements in Human Factors. JCS was perhaps ahead of its time, but surely has renewed relevance with the advent of AI-based systems and the concept of shared (i.e. between human and AI) situation awareness. Given that JCS / CSE are about coordination, unsurprisingly they also drew together a range of different disciplines, forming a hybrid community comprising Computing/Data Science, Systems Engineering and Systems Thinking, Human Factors, Neuro-Science and Social Sciences (including Psychology). Such a hybrid community approach would probably benefit Human-AI Teaming design and development.

 doi:10.20944/preprints202501.0974.v1

2.2.5. Ironies of Automation

At around the same time, Bainbridge [54] published her seminal 'Ironies of Automation' article, which highlighted some of the key dilemmas of human-automation pairing that still exist today and are relevant to human-AI pairing. As an example, as automation increases, human work can require exhausting monitoring tasks, so that rather than needing less training, operators need to be trained more to be ready for the rare but crucial interventions. In particular, Bainbridge and her colleagues stated the case that automation is often given precedence, while humans are left to do the things that automation can't do, including stepping in when the automation can no longer deal with the current conditions, or simply fails. The importance of the *Ironies* is that they act as checks and balances for system designers, with warnings about certain design philosophies and pathways that tend not to work, leading instead to endemic system performance problems and drawbacks.

This approach warrants updating in the context of AI. Endsley [55] has already begun the process, but it needs updating in the context of LLMs as well as ML systems, and in particular for Human-AI Teaming scenarios.

2.2.6. Levels of Automation & Adaptive Automation

Sheridan was one of the key proponents of 'levels of automation', which can be seen as an alternative or complement to Fitts' List. His ten levels of automation [56, 57] run from fully manual to fully automated:

1. The computer offers no assistance, human must take all decisions and actions
2. The computer offers a complete set of decision/action alternatives, or
3. Narrows the selection down to a few, or
4. Suggests one alternative, and
5. Executes that suggestion if the human approves, or
6. Allows the human a restricted veto time before automatic execution
7. Executes automatically, then necessarily informs the human, and
8. Informs the human only if asked, or
9. Informs the human only if it, the computer, decides to
10. The computer decides everything, acts autonomously, ignores the human

Sheridan's work also contributed to the notion of adaptive automation [58], wherein, for example, the automation could step in when (or ideally, before) the human became overloaded in a work situation. However, there is always the danger that as automation fails, the flight crew once again encounter the primary irony of automation and are left to fend for themselves when they most need support. Adaptive automation / autonomy usefully parsed human-machine interactions into four categories: 1) information acquisition; 2) information analysis; 3) decision and action selection; and 4) action implementation. It proposed that levels of autonomy (from high to low) could exist in each of these categories. This thinking has influenced EASA's six categories, and is inherently useful, for example, in suggesting that machines/automation/AI can advise pilots, but the pilot takes the final action, or makes the final decision (thus maintaining a critical degree of 'agency').

Adaptive automation has never quite realised the success it had hoped for, but with AI it is clearly a prospective area where adaptive automation could work and bring benefits. Both 'halves' of adaptive AI would be needed. The first is the detection of a situation where the demands on the human are excessive, whether due to external demands such as air traffic intensity, or internal factors such as fatigue or startle response. The second is the ability to take over certain tasks in a way that would be acceptable to, and managed/supervised by, the humans in charge. The earlier-mentioned case study on startle response [26] is effectively AI-supported adaptive automation, detecting startle and then directing the pilot's attention to key display components to stabilise the aircraft.

Adaptive automation suggests the need for a number of requirements, including how to protect the role and expertise of the human, how to switch from human to AI and back again as required. Adaptive automation also raises ethical issues in terms of data protection, e.g. if AI components such as neural networks use real-time human performance data (EEG, heart rate, galvanic skin response,

 doi:10.20944/preprints202501.0974.v1

etc.) as inputs to determine when to take over. Pilots may have concerns over the measurement and recording of such data, as it could reflect on their medical fitness for duty, a pilot licensing requirement.

2.2.7. Situation Awareness, Mental Workload & Sense-Making

Prior to Situation Awareness as a Human Factors construct, there were (and still remain) related concepts such as vigilance and attention [59], born from the early study of WWII radar display operators (and later, air traffic controllers) and the ability to detect signals such as incoming missiles or aircraft, given that the signal-to-noise ratio was low, and given both time-on-task and fatigue. Whereas attention and vigilance can be considered to be states related to alertness or arousal, situation awareness tends to be more contextualised, e.g., awareness of elements in the environment, such as aircraft in a sector of airspace.

Situation Awareness (SA) therefore focused less on the state of cognitive arousal, and more on the detail of what the human was aware of and what it meant for the current and future operation. Indeed, the principal SA method, the Situation Awareness Global Assessment Technique (SAGAT, [60]) focused on three time-frames: past, current and future (the latter also known as prospective memory). As complexity rose in human-plus-automation environments such as cockpits and air traffic control centres, the question became how well the human understood what was going on, what was going to happen in the near future, and how to act. The next question became how much the pilot or air traffic controller could assimilate from their controls and displays, in both short timeframes and longer durations. This led to the study of mental workload (MWL), considering the task demands vs. human capabilities and capacities, including problems of overload and underload [59, 61].

SA and Mental Workload signalled a deepening of the focus on cognitive activities, relating them in a measurable fashion to the actual cognitive work and operational context, delivering viable metrics that could be measured in simulations. Both SA and MWL have become mainstays in the design, validation and operation of human-operated aviation systems. They are useful constructs because they help designers determine what the flight crew need to know (and see/hear), and when, and in what sequence and priority order, and when users may be overloaded. However, the perception of what is happening (and going to happen) to an object may be insufficient when the AI is processing its inputs; the human may need to understand what is behind the AI output.

More recently, this delving into mental processes has focused on *sense-making*, which is the way people make sense out of their experience in the world [62]. Sense-making deals with the human need to comprehend, often via the exploration of information. In relation to AI, sense-making helps to test the plausibility of an AI's explanations as well as anomalous outputs or event characteristics. Sense-making is a useful construct particularly in conditions where uncertainty is high and not all signals may be present, or else some signals may be erroneous. Unfortunately, for such situations, as can happen in flight upsets, there is no straightforward associated metric to determine how easily flight crew will be able to make sense of a particular scenario. Instead, realistic simulations are carried out with pilots undergoing non-nominal events, and SA measurements and post-simulation debriefs, as well as safety and aircraft performance measures, are the best way to determine the safety of the cockpit design or air traffic control system.

There is a very real danger that AI systems, which tend to be 'black boxes', can undermine the human crew's situation awareness, both in terms of what is going on, and of what the AI is doing or attempting to do. A critical question therefore becomes how to develop an interface and interaction means so that the AI and the human can remain 'on the same page'. This is not a trivial matter.

Additionally, there will need to be operational explainability (OpXAI), so that the human crews can determine (i.e. make sense of) why a course of action has been recommended (or taken) by the AI. Such explainability needs to be in an operational context in language that crews can follow (as opposed to data analytic explainability, which refers more to how to trace an AI's outputs to its internal architecture, data sources and algorithmic processes).

 doi:10.20944/preprints202501.0974.v1

Workload will become more nuanced with human-AI teaming partnerships, as there may be more periods of underload followed by intense workload periods, especially if the AI is suddenly unable to function.

Realistic and real-time simulations with human crews as occur now, both for flight crew and air traffic controllers, should continue. These will become human-plus-AI simulations. The human crews need to see how AI works in realistic contexts, so they can gain trust in it. They also need to see it when it fails or becomes unavailable, so that they can recover from such scenarios. Low SA, plus a sudden spike in mental workload due to loss of the AI support, and an inability of the AI to explain its recommendations, could well be a recipe for disaster.

2.2.8. Rasmussen & Reason – Complex Systems, Swiss Cheese, and Accident Aetiology

Rasmussen's work on complex systems and safety, underpinned by his Skill, Rule and Knowledge-Based Behaviour hierarchy [63], had a big impact in the 80's and 90's in many high-risk industry domains. It may be worth revisiting this model as there are undoubtedly skill-sets that may be lost (this already happens with automation), and rules (e.g. Standard Operating Procedures [SOPs] in cockpits) may become more fluid as AI-based support systems find ever-new ways of optimising operations in real-time. Perhaps most interesting will be the area of knowledge-based behaviour (KBB), namely having to consider what is going on in a situation based on a fundamental understanding of how the system works and responds to external and internal perturbations. KBB incorporates not only factual knowledge (also called declarative knowledge) but experience amassed over years of operating a system (e.g. an aircraft) in a wide range of conditions.

KBB can be supported by ecological interface design [64], resulting in high level displays to monitor critical functions or safety parameters, as were developed in the nuclear power domain to avoid misdiagnoses of nuclear emergencies. Since AI operational explainability can probably never be completely trustworthy, an interface that affords the pilot or other aviation worker an effective system safety overview, unfiltered by AI, would seem a sensible precaution. Moreover, since diagnosis of an aircraft emergency may become a shared human-AI process, the question is one of how to ensure that no human 'tunnel vision' or misdiagnosis by an AI leads to catastrophe. A deeper question becomes how to visualise AI performance so that it is evident to the human when the AI is outside its own knowledge base or has limited certainty about its prognosis.

Reason's so-called Swiss Cheese model of accident aetiology (see Figure 4) [65], although dated, is still in regular use today. It proposes that accidents occur via vulnerabilities in a succession of barriers (e.g. organisation, preconditions, [un]safe acts and defences). The vulnerabilities are like the holes in Swiss cheese, and the larger or more proliferated they are, the easier it is for them to 'line up' and for an accident to occur. What is interesting is that AI could in theory affect all these layers, either increasing or decreasing the size and quantity of the holes. It could also reduce the independence between each barrier, so that in reality a system has fewer barriers before an accident occurs.

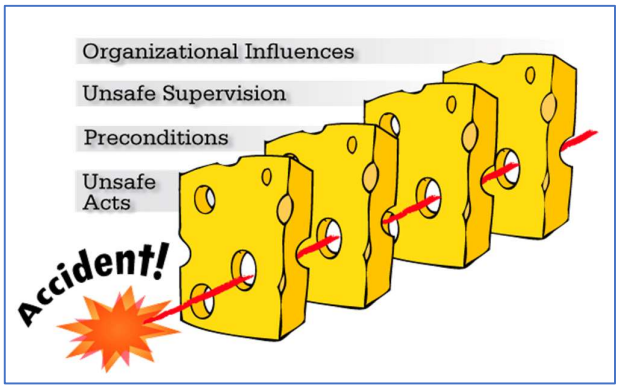

Figure 4. – Reason's Swiss Cheese Safety Metaphor.

 doi:10.20944/preprints202501.0974.v1

It would be useful to consider how AI could affect the Swiss Cheese layers differentially, e.g. use of LLMs at the *organisational* and *preconditions* layers, and AI-based tools at the *unsafe acts* and *defences* layers. Such a layered model also leads to the question of how different AIs will interface with each other. We already talk of Human-AI Teaming, but there will also be Human-AI-AI-Human and Human-AI-Human-AI variants before long, potentially allowing for problems to propagate unchecked across traditional 'defence-in-depth' boundaries.

2.2.9. Human Centred Automation

In the 1990s, after a series of 'automation-assisted aviation accidents' following the introduction of glass cockpits, Billings [66] developed the concept of Human Centred Automation. This tradition has generally persisted in aviation ever since. The nine core principles of HCA are as follows:

i. The human must be in command
ii. To command effectively, the human must be involved
iii. To be involved, the human must be informed
iv. The human must be able to monitor the automated system
v. Automated systems must be predictable
vi. Automated systems must be able to monitor the human
vii. Each element of the system must have knowledge of the others' intent
viii. Functions should be automated only if there is good reason to do so
ix. Automation should be designed to be simple to train, learn, and operate

These are the 'headline' principles, but there are many others in this watershed research carried out by Billings and others, e.g. '*automation should not be allowed to fail silently*.' Such principles (i.e. all of them, not only the top 9) should be revisited for AI, as already several of them are in danger of not being upheld, e.g.: some AI-based systems may not be predictable; reciprocal knowledge of the others' intent may prove difficult to achieve in practice; there is a tendence to 'try AI' everywhere rather than to ask where it is needed and viable; and it may not prove so easy to train / learn / operate AI-based systems. There is urgent research to be done here.

2.2.10. HFACS (& NASA-HFACS and SHIELD)

The Human Factors Analysis and Classification System (HFACS) approach [67] was developed for the US Navy, and has proven popular, often leading to 'variants' in other domains including space (NASA-HFACS [68]) and maritime (the SHIELD taxonomy, [69]. HFACS and equivalent taxonomies of human error and its causes / contributory factors have been in use for decades to determine how and why an error occurred, and how to prevent its recurrence. HFACS-like systems do not address only the surface factors (what happened), but deeper causes, including Human Factors elements, supervisory practices, and organisational and cultural factors. In this respect HFACS embodies a Swiss Cheese approach.

As AI-based systems are introduced to support flight crews and air traffic controllers in their daily operations, taxonomies such as HFACS will likely need updating, adding new terms linked to human-AI interactions. Although some blanket terms already exist, such as complacency and over-trust with respect to automation, these are probably not nuanced enough to capture the full extent of the transactional relationships that will exist between human crews and AI support systems, especially as those systems become more advanced and even executive (i.e. not requiring human oversight).

Additionally, as the AI is seen more as an 'agent', consideration must be given to the ways in which it, too, can fail. Already, as noted above, some general failure modes have been considered, such as data biasing, hallucinations, edge and corner cases, etc., but there are likely many more, some of which may be subtle and hard or even impossible to detect. The scenario in which 'data forensics' is required to understand why an AI suggested something that contributed to an accident, is probably not too far in the future.

 doi:10.20944/preprints202501.0974.v1

2.2.11. Safety Culture

Safety culture, namely the priority given to safety by the organisation, from the CEO down to the front-line and support workers, originated in the nuclear power industry following the Chernobyl disaster [70]. It has been applied to aviation most notably since the Uberlingen midair collision in 2002 and mentioned frequently in other aviation accidents. Today, most European air traffic management organisations have undergone safety culture surveys [71]. Commercial aviation (air traffic control and commercial flight crews) is generally seen as having positive safety culture.

As noted in a companion paper to this one [71], the introduction of AI into operational aviation systems could aid or degrade safety culture in aviation. In particular, the concern is that human personnel may delegate some of their safety responsibility to the AI, especially if the AI is taking more of an executive role. There is therefore a critical need for a new safety culture approach to monitor what is happening to safety culture as AI is introduced into the cockpit and ATC Ops room or Tower.

2.2.12. Teamwork and the Big Five

According to the 'Big Five' theory of Teamwork [72], the core components of teamwork include team leadership, mutual performance monitoring, back-up behaviour, adaptability, and team orientation, which collectively lead to team effectiveness. These components are predicated on shared mental models, mutual trust and closed-loop communication. What is clear from Salas et al's original thorough study of teamwork effectiveness is that it doesn't happen naturally and is something that needs to be continually maintained.

Similar to Crew Resource Management, Teamwork is something that could be significantly disrupted by the imposition of AI 'agents'. However, AI could also act as a diverse and potentially more comprehensive 'mental model' backup system for the pilots, one that is instantly available and adaptable when a sudden unexpected situation occurs. This contrasts with today, where it can take humans a short time to re-adapt, time that can be crucial in an emergency scenario in-flight (one of the principal reasons there are always two pilots in the cockpits of commercial airliners). Probably key for Human-AI Teamwork will be trust and closed-loop communication. The latter will likely entail short, succinct and contextual explainability provided by the AI, whether in procedural or natural language, and/or visually via displays. Team orientation will be an issue for AI-based systems that aim to carry out and execute certain tasks. The very existence of such agents suggests the need for specialised training for leadership of a human-AI team.

2.2.13. Bias and Complacency

Parasuraman [73] considered the potential end states of automation as implemented in aviation systems. He defined four (non-exclusive) categories: use, misuse (over-reliance), disuse (disengagement) and abuse (poor allocation decisions between human and machine). He found the primary factors influencing these end states to be trust, mental workload, risk, automation reliability and consistency, and knowing the state of the automation. Further work in this area focused on complacency with automation. One of the more worrying biases was the withdrawal of attention from cross-checking the automation and considering contradictory evidence, summarised as 'looking but not seeing'. More worrying still is that effects such as complacency (not checking) and automation bias (over-trust) are not easy to fix, e.g. via training, and are apparently prevalent in both experienced or naïve (i.e. new) users [74].

As with several other major study areas of Human Factors, the area of automation bias (especially complacency) needs to be revisited, for several reasons. The first is that cross-checking AI is likely to be more complex, as the way the AI works will itself be more complex and sometimes not open to scrutiny (either non-explainable or unfathomable for humans). Added to this is the fact that GenAI tends to be 'seductive' in that the way things are worded sounds plausible and is couched in natural language. The availability and salience of contradictory evidence is highly pertinent to the

 doi:10.20944/preprints202501.0974.v1

human capacity of acting as a back-up to the AI, given the well-known biases of humans, including representativeness, availability, anchoring and confirmation biases [75, 76] (AIs, especially LLMs, are also not immune to biases). The human may want to know why a particular course of action was suggested and others were ignored. Ways to show such 'alternates' therefore need to be considered, possibly including the trade-offs the AI has made, or data it has ignored as outliers or not relevant. Similarly, knowing the state of the AI will also be important; not simply whether it is 'on' or 'off', but its confidence level given the situation at hand compared to the data it was trained on, for example. There is also the question of prior experience: existing pilots can compare their experience to what the AI is suggesting, whereas new pilots (in the future) who have never known a system without AI support, may never have such 'unfiltered' experience.

2.2.14. SHELL, STAMP / STPA, HAZOP and FRAM

There are a number of means of analysing the risks associated with human-machine systems. A general thematic framework systems approach model is provided by SHELL [77], which considers the software (procedures and rules), hardware, environment, and 'liveware' (humans and teams), and how they can interact to either yield safe or unsafe outcomes. Approaches such as the Systems Theoretic Accident Model and Processes (STAMP) framework and its derivative Systems Theoretic Process Analysis (STPA), both developed by Leveson [78], deliver a powerful and formalised analytic approach for identifying human-related hazards and potential mitigations. The Hazard and Operability Study (HAZOP) approach developed in the chemical industry in the 70s by Kletz [79], though arguably less structured, is still a powerful and agile approach to identifying human-related hazards in complex systems, including those with AI [80]. The Functional Resonance Analysis Method [81] has as core premises that most of the time, most things go right, and that there is often a large gap between 'work as imagined/designed' and 'work as done' in the real operational system (this perspective on risk known colloquially as Safety II [82]). As people are continually making adjustments in the work situation (because most work systems are not static, they evolve), things can go wrong due to the aggregation of small variances in everyday performance, which can aggregate and propagate in an unexpected manner. FRAM can identify how such variability can effectively 'resonate' to yield unsafe outcomes.

No matter how good the design of an AI-based system, neither humans nor AI nor their combination can think of everything; there are likely to be potential hazards that can lead to accidents. As a safety-critical industry, aviation must carry out hazard assessment to identify such potential hazards and develop appropriate mitigations. At the moment, all of the methods above (and others) can be used to identify hazards that can occur in human-AI systems. The problem is that we are missing two inputs. The first is a model of how the AI can fail, or exhibit aberrant behaviour, or suggest inaccurate/biased resolutions or advice. We already know some of the answers, in terms of hallucinations, edge cases, corner cases, biased data usage, etc. [e.g. see 83]. But these are generic. What we need is a way to determine when these and other AI 'failure modes' are likely, given the type of ML / LLM being used, its data, and the operational context it is being applied to. It is likely that a taxonomy of AI failure mechanisms will develop as more experience is gained with AI-based systems, but it would be preferable not to have to learn the hard way.

The second unknown, or 'barely known', relates to 'human+AI' failure modes, i.e. the likely failure types when people are using and interacting with various types of AI tools. We already have 'complacency', but again, this is a catch-all term, and knowing when it is likely or not, as a function of the human-AI system design, is unpredictable. This is problematic: how can a safety critical system be certified if complacency is a likely user characteristic and the AI can fail? There is a need for greater understanding of the evolution of human-AI inter-relationships. This may entail longer-term study of human-AI working partnerships, perhaps in extended simulations (lasting months rather than days or weeks), effectively constituting a safe 'sandbox' in which to see emergent behaviour of both human and AI.

 doi:10.20944/preprints202501.0974.v1

2.2.15. Just Culture and AI

Since the implementation of ASRS, aviation had generally been seen as having an effective reporting culture, enabling it to be an informed culture, learning from events to continually improve. The reporting culture is predicated on a Just Culture [84], in which the aviation system prefers to learn from mistakes than blame people for them, as long as there is no reckless behaviour or intention to cause damage or harm. In European aviation, this principle is enshrined in law[2]. But there is a potential double-bind in the future for human agents in aviation [85]. If they are advised by an AI to do something and it results in an accident, they may be asked to justify why they did not recognise the advice was faulty. Similarly, if they choose not to follow AI advice and there is an accident, they will be asked why they did not follow the AI's advice. Responsibility and justice are human constructs, and an AI, even if its future role is that of an executive agent in an operational aviation system, cannot be prosecuted in a court of law, and prosecuting its developers will likely be fraught with legal complexities. For example, what judge or jury will have sufficient AI literacy to understand what really happened in such an accident, and how will they overcome hindsight and other biases (e.g. 'the pilot should have known better') in forming their deliberations?

Just Culture could be a major deal-breaker for unions and professional associations who feel their members (pilots, controllers, others) are at risk of being prosecuted in an area with little legal precedent, and given juridical framework variations according to country, and potentially serious criminal charges. Hypothetical test cases should be carried out in legal sandboxes to anticipate legal argumentation and potential outcomes in this area, and more generally to expand the Just Culture 'playbook' [86]

As an aside, perhaps one useful outcome of the release of ChatGPT and other LLMs into the public domain is that many people see for themselves not only the benefits of these tools, but also how they can mislead or outright 'lie' to human end-users. In societal terms, which are important when it comes to justice, some of the initial 'shine' of AI has worn a little thin in the public eye.

2.2.16. HF Requirements Systems – CS25/1302, SESAR HP, SAFEMODE, EASA & FAA

Over the decades there have been various Human Factors requirements and assurance approaches. Of particular note in Europe is EASA Certification Standard (CS) 25.1302[3], which is concerned with controls and displays in cockpits, and contains detailed guidance on all aspects of information usage including display design, situation awareness, workload, alarms, etc. Essentially, all controls and displays need to be fit for human purpose whether in normal, degraded mode or emergency situations. Whilst there is no European-wide regulatory equivalent for Air Traffic Management (ATM) systems in Europe, there is comparable guidance available via the SESAR (Single European Sky ATM Research) programme and its Human Performance Assessment Process (HPAP) [4]. This comprehensive approach breaks down four high level areas – roles and responsibilities, the human-machine interface, teams and communications, and transition to operations – into detailed and measurable requirements in an argument-based structure. The SESAR HPAP allows a Human Factors 'case' to be built for a new system design or change, showing where the design complies with Human Factors principles, and where it does not (or does not need to). More recently, an EU research project called SAFEMODE[5] has developed a Human Factors assurance

[2] 1. REGULATION (EU) No 376/2014 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 3 April 2014.

[3] https://www.easa.europa.eu/en/search?keys=CS25%201302&f[0]=origin:EASA+Pro

[4] https://www.sesarju.eu/sites/default/files/documents/transversal/SESAR%202020%20-%20Human%20Performance%20Assessment%20Guidance.pdf

[5] https://www.safemodeproject.eu/

 doi:10.20944/preprints202501.0974.v1

platform called HF Compass, which refers to the HPAP but also highlights more than twenty tools and techniques that can be used to provide evidence that Human Factors has been assured for a new system. Very recently, EASA[6] has provided preliminary guidance on the Human Factors assurances required for AI-based systems in aviation, in the form of regulatory requirements. The Federal Aviation Agency (FAA) has also just released its own Roadmap[7] for safety assurance of AI-based systems in aviation, though it is currently focused more on safety than Human Factors.

The existing guidance from CS25 1302, the SESAR HPAP, and the EASA guidance on Human Factors aspects of human-AI Teaming systems, all offer excellent foundations for an approach focused on Human-AI systems. Section 3 accordingly presents a Human Factors Assurance system for Human-AI Teaming systems.

2.3. Contemporary Human Factors & AI Perspectives

Having reviewed the historical landmarks in Human Factors and their implications for Human-AI systems, this sub-section briefly reviews some of the more recent emerging Human Factors literature on Human-AI Teaming, focusing on a model of HAT, the somewhat tricky issues of personification (anthropomorphism) of AI, and emotion-mimicking AI.

2.3.1. HACO – a Human-AI Teaming Taxonomy

In a recent paper on Human AI Collaboration (HACO), a Human-AI Teaming (HAT) taxonomy has been usefully mapped out [87] and is shown in Figure 5 below:

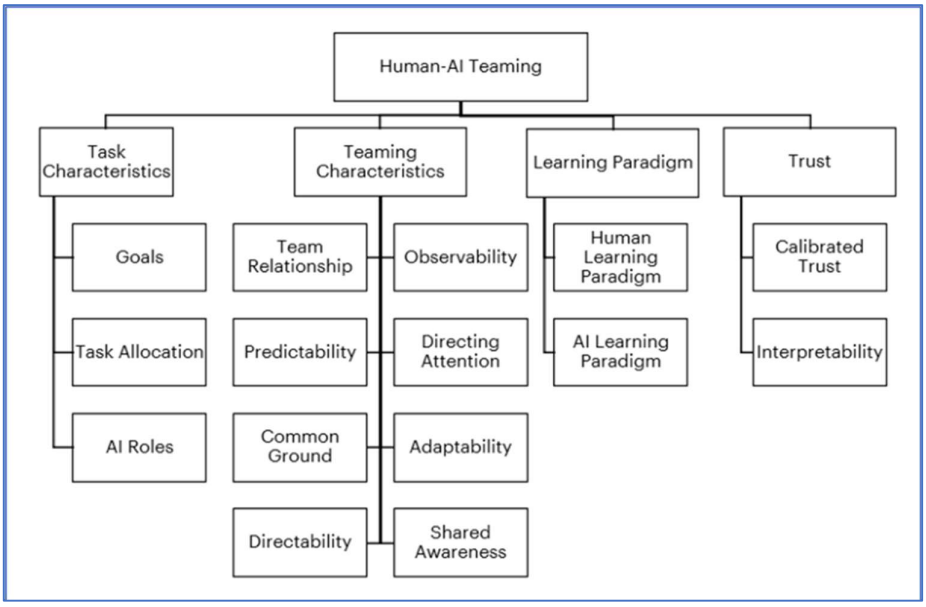

Figure 5. – the HACO Framework [87].

The HACO concept has six core tenets: context awareness, goal awareness, effective communication, pro-activeness, predictability and observability. Taken as a whole, one way to summarise these tenets is that they all ensure that the team members, including the AI, ***are on the same page*** with what is happening and what to do about it. In practice this should mean that the workflow remains smooth, without surprises, significant breaks or disruptions, or conflicts. As with human teams, it does not require perfect understanding of one another, but an appreciation of

[6] https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-concept-paper-issue-2

[7] https://www.faa.gov/aircraft/air_cert/step/roadmap_for_AI_safety_assurance

 doi:10.20944/preprints202501.0974.v1

individual behavioural norms (including style of interaction), pace of work, skill sets and capabilities, and limitations.

Context awareness is possibly a more useful (and less anthropomorphic) label than situation awareness. It seeks to establish what the AI was responding to within its inputs and datasets. However, this is not always simply the superficial data in the environment (e.g. weather patterns that might affect flight route), but the way the AI will use statistical and other algorithms to interpret such data according to the task and goals at hand. The design challenge is to maintain the usability adage of '*what you see is what you get'*, whether this is achieved via natural language dialogue or (more likely) visual media that can decrypt the AI's computational processes in meaningful and quickly 'graspable' ways for the end user.

Goal awareness (a precursor to ***goal alignment***) is a higher-level attribute, related to context awareness, and ensures that human and AI are on the same page at all times. This becomes important in work arrangement scenarios where the AI is able to modify the goal hierarchy, and where goals may dynamically shift during a scenario (e.g. for an emergency, as it worsens or becomes more stable). It can also be important where there are a mixture of safety and other goals, some of which may conflict with safety. As for contextual awareness, a design challenge is how to ensure the human is aware of the goals the Ai is working towards as they change and evolve.

Effective Communication can occur via various modalities, from natural language and even gestures, to digital displays and procedural textual responses to human prompts. The keyword here is "effective.'' Communication for AI tools or systems below EASA's category 2B may not need to be that advanced, though for 1B and 2A there may need to be an element of explainability if the AI's task or output is sufficiently complex. For human AI collaboration at the 2B level, there will likely not only need to be communication, but a degree of rapport, so that the AI is communicating in terms and contexts familiar to the human. This implies a shared understanding of the local operational environment conditions and practices. For example, this can refer not only to a specific aircraft type such as an A320 or B737, but how those aircraft types are fine-tuned by the airlines using them, along with their Standard Operating Procedures and day-to-day practices.

Proactiveness links again to EASA's 2B and above categorisations, whereby the AI can initiate its own tasks and even shift or re-prioritise goals, giving the AI a degree of autonomy and even agency (since it can act under its own initiative). However, in contrast to the distinctive categorisations of EASA these authors [87] suggest the concept of ***'sliding autonomy'***, wherein the human (or the system) determines the level of autonomy. This is of interest in aviation in cases where flight crew or air traffic controllers, for example, may be overwhelmed by a temporary surge in tasks or traffic respectively, and may wish to 'hand-off' certain tasks (especially low-level ones) to an AI. This is effectively an update of the Adaptive Automation concept.

Predictability can refer both to the degree to which the pace of work of an AI is understood, trackable and manageable, to the degree to which an AI might 'surprise' the humans in the team via its outputs. The former probably requires some kind of overview display to know what the AI is doing and the progress on its tasks or goals. The latter will depend on the amount of human-AI training afforded prior to teamworking in real operational settings.

Observability refers to the transparency of the progress of the AI agent when resolving a problem, and can be linked to explainability, though in practice the AI's workings might be routinely monitored via some kind of dashboard or other visualisation rather than a stream of textual explanations, albeit with the user able to pose questions as needed.

Three additional HAT attributes are mentioned in [87] and are worth reiterating here:

- ***Directing attention*** to critical features, suggestions and warnings during an emergency or complex work situation. This could be of particular benefit in flight upset conditions in aircraft suffering major disturbances.
- ***Calibrated Trust*** wherein the humans learn when to trust and when to ignore the suggestions or override the decisions of the AI.
- ***Adaptability*** to the tempo and state of the team functioning.

 doi:10.20944/preprints202501.0974.v1

With respect to Adaptability, one of the hallmarks of an effective team, it may be useful to resurrect High Reliability Organisation (HRO) theory [88]. All five of the pillars of HROs – *preoccupation with failure*, *commitment to resilience*, *reluctance to simplify explanations*, *deference to expertise* and *sensitivity to operations* – could well apply to Human-AI Teams. HRO theory as a whole leans towards collective mindfulness of the team, and here is where it is necessary to consider how humans think 'in the moment' compared to how an AI might build up its own situation representation or context assessment.

2.3.2. AI Anthropomorphism and Emotional AI

Human-AI Teaming can be considered to be an anthropomorphic term [33], suggesting that the AI is a team player, denoting human qualities to a machine. With GenAI systems the 'illusion of persona' can be stronger, since one has the impression of conversing with a person rather than a program, to the extent that we often add polite 'niceties' such as please and thank you in interactions with ChatGPT and equivalent LLMs, forgetting that they are simply very powerful algorithms calculating the next most likely word in a sequence.

Anthropomorphism relates to the identity we assign an AI system, and hence the degree of agency we accord it. The more we 'personify' AIs, the more the danger of delegating responsibility to it, or of surrendering authority to its logic and databases, or of basically second-guessing ourselves rather than the AI. In this vein it is worth recalling that AI systems are created by people (data scientists), and there are many human choices that go into development of AI systems, e.g. choice of training, validation and test datasets, choice of hyperparameters, selection of most appropriate algorithm, etc. [4].

The recent FAA Roadmap on AI is explicitly against the personification of AI, and given that AI has no sentience and will likely not do so for some time, it is right to do so, especially in safety critical operations. A key practical consideration, however, is whether treating future AI systems as a team member could enhance overall team performance. This is as yet unknown, but it leads to a second question of whether we can tell the difference between a human and an AI. In a recent study of 'emotional attachment' to AI as team members [89], most participants could tell the difference between an AI and a human based on their interactions with both. Another study [90] examined human trust in AIs as a function of the perception of the AI's identity. The study found that AI 'teammate' *performance* matters, whereas AI *identity* does not. The study authors cautioned against using deceit to pretend an AI is a human. Deception about AI teammate's identity (pretending it is a human) did not improve overall performance, and led to less acceptance of AI solutions. Knowing it is an AI improved overall performance.

In a non-safety-critical domain study [91], two key acceptance parameters for emotional AI were found to be the potential to erode existing social cohesion in the team, and authority (impacts on the status quo). As with other studies in this area, the authors found people judged machines by *outcomes*. A further study [92] found that monitoring people's behavior and emotional activity (speech, gestures, facial expressions, and physiological reactions), even if supposedly for health and wellbeing, can be seen as intrusive. Such monitoring activities can be for stress, fatigue and boredom monitoring, and error avoidance, and of course productivity. People may be uncomfortable with this level of personal intrusion of their behavior, bodies and personal data.

Overall, therefore, preliminary indications are that aviation needs neither anthropomorphic (personified) nor emotional AI. What matters is the effectiveness of the AI in the execution of its tasks.

3. Human Factors Requirements for Human-AI Systems

Having reviewed AI and Human Factors relevant to AI-based systems in aviation, this section outlines the development of a new and comprehensive set of Human Factors requirements for future HAT systems. First, any Human Factors Requirements System for Human-AI systems must satisfy several requirements of its own:

 doi:10.20944/preprints202501.0974.v1

It must capture the key Human Factors areas of concern with Human-AI systems.

11. It must specify these requirements in ways that are answerable and justifiable via evidence.
12. It must accommodate the various forms of AI that humans may need to interact with in safety critical systems (note – this currently excludes LLMs) both now and in the medium future, including ML and more advanced Intelligent Assistants (EASA's Categories 1A through to 3A).
13. It must be capable of working at different stages of design maturity of the Human-AI system, from early concept through to deployment into operations.

With respect to this last requirement, the NASA system of Technology Readiness Levels[8] (TRLs) is a useful maturity model for aviation system design maturity, as illustrated in Figure 6

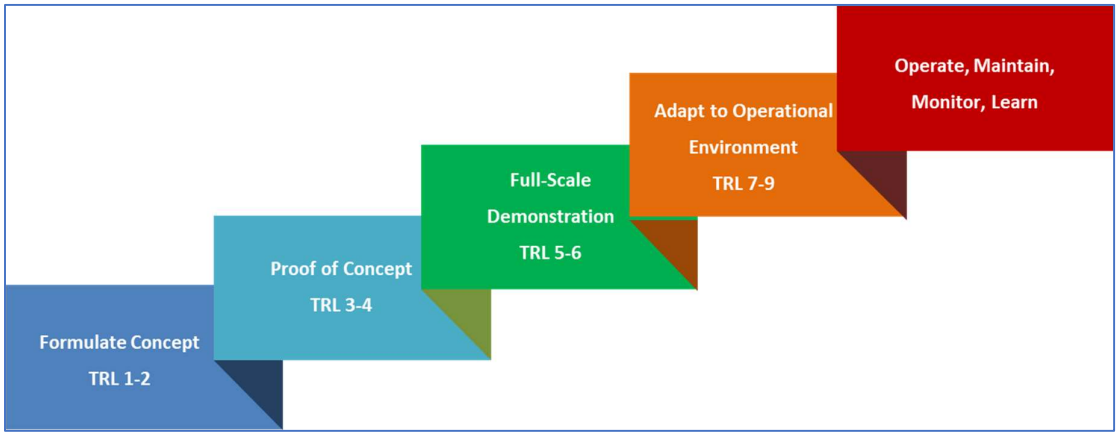

Figure 6. – Technology Readiness Levels.

In the early stages of a project, a number of the requirements questions will have to be returned to at a later stage in the development life cycle. But it is better to highlight them early rather than wait until the design is too advanced to be changed. Once the system is near deployment, all questions can usually be answered, and the organisation will have a robust basis upon which to deploy and operate the Human-AI system.

Following on from the principles of Human Centred Design, Human Factors needs to be integrated from the early design stages. Otherwise, the design of the AI will be a poor 'fit' to the human end user, and the only significant degree of freedom left to optimise HAT performance will be training. Training end users to operate a poorly designed system is a bad design strategy. Therefore, Human Factors requirements should ideally be applied from the early concept stages onwards, until the system is operational. The critical design stages however, in terms of integration of Human Factors to deliver optimal system performance, are TRLs 3-6, as these stages determine how the human-machine (AI) interactions will take place.

Typically, TRLs 1-2 are concerned with early concept exploration. For Human-AI System design, at this TRL usually the most important considerations focus on Roles and Responsibilities, namely who (or what) will be in charge when the AI is operating. Decisions about sense-making such as how shared situation awareness will be established, and primary human-AI interaction modes, e.g. 'conventional' (keyboard, mouse, touchscreen, etc.) or more advanced (speech, gesture recognition) can be made, as well as whether the AI will be used by a single person or a team. It is at this stage that key design choices are made concerning the adoption of a Human Centred Design approach, including the use of end users in informing the early 'foundational' concept.

TRLs 3-4 add a lot more detail, fleshing out the concept's architecture, and gaining a picture of what the AI might look and feel like to interact with, via early prototypes and walk-throughs of

[8] https://www.nasa.gov/directorates/somd/space-communications-navigation-program/technology-readiness-levels/

 doi:10.20944/preprints202501.0974.v1

human-AI interaction scenarios. The area of sense-making is crucial here, and communication and teamworking aspects will become clearer.

At TRLs 5-6 models and prototypes will be developed iteratively until a full-scope demonstration is completed and tested in a realistic simulation with operational end users. This period of design and development will see many issues 'nailed down' and solidified into the design and operational concept (CONOPS), having been validated via robust testing with human end users. Risk studies will focus on errors and failures seen, as well as those that could conceivably occur once the system is in operation.

TRLs 7-9 prepare the concept for deployment in real world settings. This is the time to consider in earnest the ramifications of entering a human-AI system into an operational organisation, with requirements for competencies and training of end users, as well as the often 'thornier' socio-technical systems issues such as staffing, user acceptance, ethics and wellbeing. These latter issues often determine whether the system will be accepted and used to its full extent.

After TRL 9 the system is in operational use. However, since most AI systems are learning systems, monitoring – particularly in the first six to twelve months – will be a critical determinant of sustainable system performance.

The relevance of the Human Factors areas to the different TRL are highlighted in Figure 7, to help the user adapt their evaluation approach. As illustrated in the figure, the application of Human Factors areas to TRLs are not always clear cut, so some degree of judgement must be used. For many projects, in early TRLs entire areas can be deemed 'TBD' (to be decided later) or even N/A (not applicable – for now) and returned to in later design stages. For example, many research projects will not consider Human Factors areas 7 & 8, although it is worth considering impacts on *staffing* early, since this can be a 'deal-breaker' for any project.

3.1. The Human Factors Requirements Set Architecture

Based on the literature review, including the SESAR HPAP and the recent EASA Guidance on Human Factors for AI-based systems [32], as well as the EU Act on AI, eight overall areas were identified, as shown in Figure 7 and outlined below.

Human Centred Design – this is an over-arching Human Factors area, aimed at ensuring the HAT is developed with the human end-user in mind, seeking their involvement in every design stage.

14. **Roles and responsibilities** – this area is crucial if the intent is to have a powerful and productive human-AI partnership, and helps ensure the human retains both agency and 'final authority' of the HAT system's output. It is also a reminder that only humans can have responsibility – an AI, no matter how sophisticated, is computer code. It also aims to ensure the end user still has a viable and rewarding role.
15. **Sense-Making** – this is where shared situation awareness, operational explainability, and human-AI interaction sit, and as such has the largest number of requirements. Arguably this area could be entitled (shared) situation awareness, but sense-making includes not only what is happening and going to happen, but why it is happening, and why the AI makes certain assessments and predictions.
16. **Communication** – this area will no doubt evolve as HATs incorporate natural language processing (NLP), whether using pilot/ATCO 'procedural' phraseology or true natural language.
17. **Teamworking** – this is possibly the area in most urgent need of research for HAT, in terms of how such teamworking should function in the future. For now, the requirements are largely based on existing knowledge and practices.
18. **Errors and Failure Management** – the requirements here focus on identification of AI 'aberrant behaviour' and the subsequent ability to detect, correct, or 'step in' to recover the system safely.
19. **Competencies and Training** – these requirements are typically applied once the design is fully formalised, tested and stable (TRL 7 onwards). The requirements for preparing end users to work with and mange AI-based systems will not be 'business as usual, and new training approaches and practices will almost certainly be required.

 doi:10.20944/preprints202501.0974.v1

20. **Organisational Readiness** – the final phase of integration into an operational system is critical if the system is to be accepted and used by its intended user population. In design integration, t is easy to fall at this last fence. Impacts on staffing levels (including levels of pay), concerns of staff and unions, as well as ethical and wellbeing issues are key considerations at this stage to ensure a smooth HAT-system integration.

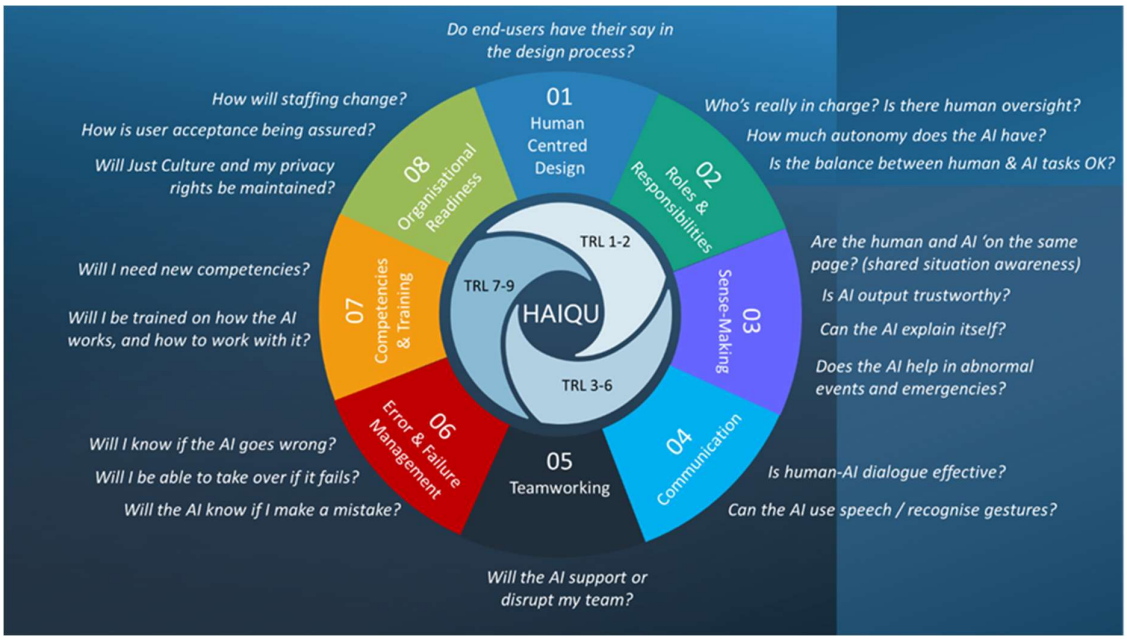

Figure 7. – Human Factors Requirements System Architecture.

This eight-area architecture goes beyond the SESAR HPAP four area structure (Roles & Responsibilities, Interface Design, Teamworking and Transition) and has added new sub-areas including trustworthiness, AI autonomy, operational explainability, speech recognition and human-AI dialogue, Just Culture, etc. Many of the requirements from both EASA's CS 25 1302 are evident at the specific requirements level, but have been re-oriented or augmented to focus on AI and HAT. The original requirements set was written before the most recent EASA HAT guidance release, and since that time alignment efforts have been made. The resultant HF Requirements set is significantly larger than the EASA one, as it deals with a number of areas and sub-areas outside EASA's current focus (e.g., related to roles and responsibilities organisational readiness, competencies, human agency, etc.).

The detailed structure is as follows, showing all sub-categories and the number of requirements in each of the total of seventeen categories and sub-categories (165 requirements in total, of which 51 are common requirements with EASA's):

1. Human Centred Design [5]
2. Roles and Responsibilities
a. Human and AI Autonomy [7]
b. Balance of Human and AI Tasks [13]
c. Human Oversight [10]
3. Sense-Making
a) Shared Situation Awareness [12]
b) Trustworthy Information [12]
c) Explainability [13]
d) Abnormal Events, Degraded Modes and Emergencies [7]
4. Communication
a. Human-AI Dialogue [10]
b. Speech and Gestures [11]

 doi:10.20944/preprints202501.0974.v1

5. Teamworking [11]
6. Errors and Failure Management [13]
7. Competencies and Training
a. New Competencies [9]
b. New Training Needs [9]
8. Organisational Readiness
a. Staffing [8]
b. User Acceptance [9]
c. Ethics and Wellbeing [8]

Extracts from the requirements question set are shown in Table 1. Those in red are common with EASA requirements (though they may have been reworded), and those in red towards the end (Organisational Readiness) are also embodied in the EU Act on AI. Whereas design requirements are typically stated as imperative statements (e.g. 'the colour coding should…'), the Human Factors requirements in Table 1 are phrased as questions. This is for three reasons. First, this requirements set's primary aim is to support designers and product teams, rather than serving as a regulatory certification tool. Second, since HAT design is a new and rapidly evolving area, there are still uncertainties about how exactly to best do something. It therefore seems more appropriate to raise the issue with the design or product development team as a question, letting them decide how best to satisfy it. Third, questions are more engaging than flat statements and may engender more 'purposeful creativity', which is welcome in a pioneering field of design application such as Human-AI Teaming.

Table 1. Example Human Factors Requirement Questions for HAT Systems.

HUMAN-CENTRED DESIGN
Are licensed end-users participating in design exercises such as focus groups, scenario-based testing, prototyping and simulation (e.g. ranging from desk-top simulation to full scope simulation)?
Are end-user opinions helping to inform and validate the design concept, as part of an integrated project team including product owner, data scientists, safety, security, Human Factors and operational expertise?
Are end-users involved in any hazard identification exercises (e.g. HAZOP, STPA, FRAM etc.)?
Do licensed end-users participate in validation and testing activities?
ROLES & RESPONSIBILITIES
Have any of the various human roles / responsibilities changed?
Are there any new roles, or suppressed roles?
What is the overall level of autonomy – is the human still in charge?
Has the basis for task allocation been articulated, and does it follow Human Factors guidance?
Are there criteria for task-switching between human and AI?
Is task-switching controlled by the human, by the AI, or a mixture of both?
In the case of AI control, can the human reject a task?
Can the AI detect poor decision-making by the user and offer alternatives?
Can the human monitor and adjust the AI's goal formulation?
If human-AI negotiation is possible, who makes the final decision?
SENSE-MAKING
Is all the required information presented to the user in an uncluttered way?
Is the interaction medium appropriate for the task, e.g. keyboard, touchscreen, voice, and even gesture recognition?

 doi:10.20944/preprints202501.0974.v1

Do visual/oral/auditory displays and controls follow Human Factors guidance (e.g. colour coding, luminance, auditory range etc.)?
Does the location of a new AI interface with the operational workplace support rather than hinder critical operations?
Does the AI-human interface reinforce the end-user's situation awareness.
In cases of moderate (or higher) uncertainty, are alternative suggestions/explanations offered?
Can the human query the information/decision via an explainability function?
Can the AI provide alternatives, along with likelihood estimates?
Is the end-user aware of the key parameters the AI is optimising? Can the end-user alter them?
Can the human view both data that were used and data that were ignored by the AI, e.g. anomalies or outliers?
COMMUNICATION
Is human-AI communication via a proceduralised phraseology (e.g. as pilots and controllers use)?
If natural language is used, how is context-sensitivity ensured, so that misunderstandings are avoided?
Are humans made aware of what the AI may be monitoring, e.g. speech aspects (to detect stress or fatigue), gestures, etc.?
Is the human-AI communication load acceptable to the human?
If spoken natural language is used, can the AI process end-user requests, responses and reactions, and acknowledge the user's intentions.
In case of confirmed misunderstanding by the end user, can the IA resolve the issue?
TEAMWORK
Has it been determined which roles will interact with the AI?
How will the impact of suppressed or new roles (if any) on teamwork be managed?
Is the AI role, its task allocation and authority level made clear to each team member?
Is AI-derived situation awareness communicated to the entire team to ensure coherent team situation awareness?
Does the human rather than the AI control or oversee task allocation and the overall workflow?
Are there sufficient skilled human crew to operate or recover the system in case of AI failure or erroneous behaviour?
Can the AI detect a breakdown of team functioning and signal this to the team or compensate accordingly?
ERROR & FAILURE MANAGEMENT
Is the human trained on AI error modes and how to verify AI results?
Are users trained to recognise and take corrective action on strange or erroneous AI outputs?
Has the human seen examples of AI incorrect information / advice in simulation training?
Does the AI inform the end-user of a detected / predicted shift from normal to emergency or degraded mode?
Are there sufficient independent (non-AI) displays of critical functions to allow the human to verify that context is changing?
Are end users trained to recognise unsafe AI operating conditions and to take corrective action?
Is it clear that the AI won't interfere with existing alerts and warning systems (e.g. TCAS, STCA etc.)?
COMPETENCIES & TRAINING
Will working with the AI technology be a competency requirement?
Will there be new competency requirements to maintain the AI
Is personnel licensing affected in any way? Are there any new licensing requirements?

 doi:10.20944/preprints202501.0974.v1

Will the AI's interactions / monitoring of the team inform competency assessment, and will end-users be aware?
If the AI is a learning system, how will human competencies be updated?
Have the risks of skills loss been identified and mitigated where required (e.g. with fallback training)?
Do existing training systems need to change (computer-based training, classroom training, simulator training, on-the-job training)?
Will humans help train the AI (supervised human-in-the-loop training)?
Will simulator exercises show what AI failure / misbehaviour looks like?
ORGANISATIONAL READINESS
Will the overall number of staff increase, decrease or remain the same (Management, Operational, Engineering, Support)?
Will new or existing staff be paid lower wages?
Do end-users think implementing the AI is a good idea?
Do the human team members feel the AI output exceeds that of a human, and improves overall task/system performance?
Are end-users comfortable with any legal liability implications of using the AI in safety-critical contexts?
Will existing protections afforded by Just Culture in the organisation be affected by the implementation of the AI?
Does the system comply with national and EU data protection provisions (e.g., GDPR)?
Are users always aware whenever they are interacting with the AI?
Could the system have any negative effects on the physical or mental wellbeing of the end user, including moral integrity?

4. Application of Human Factors Requirements to a HAT Prototype

The aim of the questionnaire is to help developers realise and deliver a highly usable and safe Human-AI system, one that end users can use effectively and will want to use (avoiding misuse, disuse and abuse as mentioned earlier). The approach has been tested on several Human-AI Teaming use cases from the HAIKU research project (https://haikuproject.eu), outlined in Figure 8. Three of these are at TRL levels 3-6, and one (Urban Air Mobility) is at TRL 1-2.

Figure 8. Examples of HAIKU HAT Use Cases.

The first use case, UC1, addresses support to a single pilot in the event of startle response, wherein a sudden unexpected serious event (e.g. a lightning strike) can cause 'startle', leading to diminished cognitive performance for a short period of time (e.g. 20 seconds) [93, 94, 95]. The FOCUS (Flight Operational Companion for Unexpected Situations) AI supports the pilot first by detecting startle via various physiological sensors analysed by a trained AI, using an AI technique known as

 doi:10.20944/preprints202501.0974.v1

Extreme Gradient Boosting. Second, it analyses where the pilot is looking compared to where the pilot should be looking given the situation, and if different, the AI highlights the relevant parameter on the cockpit displays (see Figure 9). This is effectively directed (or supported) situation awareness.

Figure 9. Images from HAIKU Use Case 1 simulation [26; 28] – *Human-AI Teaming for Startle Response – showing the cockpit simulator (top right), a pilot in the cockpit being supported after startle (top left), an extract from the eye tracking analysis of key parameters the pilot may need to focus on (bottom right) and red-box-highlighted directed situation awareness on the primary flight display (PFD))bottom left).*

This use case is currently TRL5, and experiments have been run in the simulator with commercial pilots. The HF requirements set was applied to the use case in December 2023, via a one-day session with the design and Human Factors team at ENAC (École Nationale Aviation Civile, Toulouse: https://www.enac.fr/), shortly after a series of experimental trials with pilots. An extract from the evaluation is shown in Table 2 for a sample of the questions from the first four Human Factors Areas.

Table 2. – Example responses to HF Requirements Questions from HAIKU UC1 following the first real-time cockpit simulation with airline pilots.

Human Factors Requirement Question	**Y/N N/A TBD**	**Justification**
Human Centred Design		
Are licensed end-users participating in design exercises such as focus groups, scenario-based testing, prototyping and simulation (e.g. ranging from desk-top simulation to full scope simulation)?	Y	ENAC pilots and commercial airline pilots are involved in Val1 and Val2 simulation exercises. ENAC pilots are involved in the design process, and the product owner is also a pilot.
Are end-user opinions helping to inform and validate the design concept, as part of an integrated project team including product owner, data	Y	ENAC pilots are involved, and the product owner is a pilot. Additionally there are data science and Human Factors experts in the

 doi:10.20944/preprints202501.0974.v1

scientists, safety, security, Human Factors and operational expertise?		design team. Security is outside the scope of the study as it is TRL4-5.
Are end-users involved in any hazard identification exercises (e.g. HAZOP, STPA, FRAM etc.)?	Y	An ENAC pilot was involved in the first HAZOP. Additionally, points raised by pilots during Val1 simulation were fed into the first HAZOP.
Roles & Responsibilities		
Are there any new roles, or suppressed roles?	Y	This is a Single Pilot Operation (SPO) concept study, so one flight crew member is no longer there.
What is the level of autonomy – is the human still in charge?	Y	The pilot remains in charge
Does this level of autonomy change dynamically? Who/what determines when it changes?	TBD	This aspect is being developed in the Concept of Operations (CONOPS)
Are the new / residual human roles consistent, and seen as meaningful by the intended users?	Y	Yes, for pilots it is like a clever flight director or attention director, but the pilot is still in control.
Sense-Making		
Is the interaction medium appropriate for the task, e.g. keyboard, touchscreen, voice, and even gesture recognition?	Y / TBD	Startle and SA support colour etc. was appreciated in Val 1; Electronic Flight Bag (EFB) not used due to the emergency nature of the event. Red was seen as too strong. Voice was suggested to back up the visual direction of SA (this has since been implemented). Changes being tested in Val2 experiment.
Does the AI build its own situation representation?	Y	Yes, coming from the aircraft data-bus, and from the pilot's attentional behaviour (eye-tracking). Context is also from the SOPs (Standard Operating Procedures) for the events.
Is the AI's situation representation made accessible to the end user, via visualisation and/or dialogue?	Y	It is explained to pilots how the AI works and how it uses eye tracking, the parameters it uses and how it has been informed by pilots for different flight phases (supervised training). The EFB summarises the AI's situation assessment.
Does the AI-human interface reinforce the end-user's situation awareness, so that human and AI can remain 'on the same page'?	Y	Pilots feel it helped their SA, and speed of gaining a situational picture.

 doi:10.20944/preprints202501.0974.v1

Can the human modify the AI's parameters to explore alternative courses of action?	N/ TBD	The pilot can follow an alternative course of action, though there is no 'interaction' on this with the IA. The acceptability of the directed attention will be discussed with pilots in Val2, and whether they would want the ability to alter parameters during an event.
Is at least some operational explainability possible, rather than the AI being a 'black box'?	Y	Explainability is via the EFB. However, due to the very short response times in a loss of control in flight scenario, pilots had little time for explainability.
Can the humans still detect and correct their own errors, or those of their colleagues?	Y	SPO (Single Pilot Operations) concept, so only one crew member. AI is there to guide the pilot out of startle response / temporary loss of situation awareness, and to prevent loss of control in flight.
Does the IA possess the ability to detect human errors or misjudgement and notify them or directly correct them?	Y	The IA is intended to detect temporary performance decrement due to startle, and to guide the pilot, but does not go as far as correcting his/her action.
Does the level of human workload enable the human to remain proactive rather than reactive, except for short periods?	Y	For these types of sudden scenarios the pilot is in reactive mode. The answer 'yes' is given as it is a short period, and the aim is to reduce cognitive stress, giving them more 'headspace' to deal with the event.
Can the human detect errors by the AI and intercede accordingly?	Y	Pilots in the first set of simulations detected when the AI should have deactivated but did not (they could deactivate it manually). This will be improved for the second round of simulations.
Does the human trust the AI, but not over-trust it? Is the human taught how to recognise AI malfunction or bad judgement?	Y	Pilots detected some smart agent anomalies during the experiment so did not over-trust it. Can be further explored in Val 2, e.g. using a trust scale.
Can the human suggest changes to the parameters the AI is using, to explore alternative courses of action?	N	The pilot can follow an alternative course of action, though there is no 'interaction' on this with the AI.
Communications		
Is human-AI communication via a proceduralised phraseology (e.g. as pilots and controllers use)?	Y	Aural alerts for Val2 are using standard (procedural) phraseology (e.g. "Vertical Speed")

 doi:10.20944/preprints202501.0974.v1

Are humans made aware of what human performance aspects the AI may be monitoring or recording, e.g. speech (to detect stress or fatigue), gestures, psychophysiological parameters (EEG skin conductance, eye movements, etc.?	Y	Eye tracking, and other measures (ECG, Galvanic Skin Response, etc.) are employed for Val 1 & 2 as inputs to detect startle. Pilots are fully briefed on all measures used in the simulations and sign a consent form. Data are de-identified from pilot details.
Does the AI communication mode, whether oral or audio-visual, avoid the use of emotional mimicry (i.e. mimicking human emotions)?	Y	There is no emotional mimicry employed in the FOCUS AI CONOPS
Is the AI clear and concise in times of high stress (e.g. emergencies)?	Y	Yes. No clarity issues identified in Val 1.

At several points during the HF Requirements review with the design team, potential improvements were derived for the AI system, in terms of how it is visualised and used by the pilot. In total, 27 issues were raised for further consideration by the design team, in terms of potential changes to the design or further tests to be carried out in the second set of simulations (Val2). A number of requirements were deemed 'not applicable' to the concept being evaluated, and many of the requirements from the last two areas (Competencies & Training; Organisational Readiness) were not answerable at this design stage. Figure 10 gives an overview of the results at the end of Val1 (first set of simulation runs with pilots) in terms of HF compliance ('Yes' responses in green, red = not compliant, grey = N/A and orange = TBD).

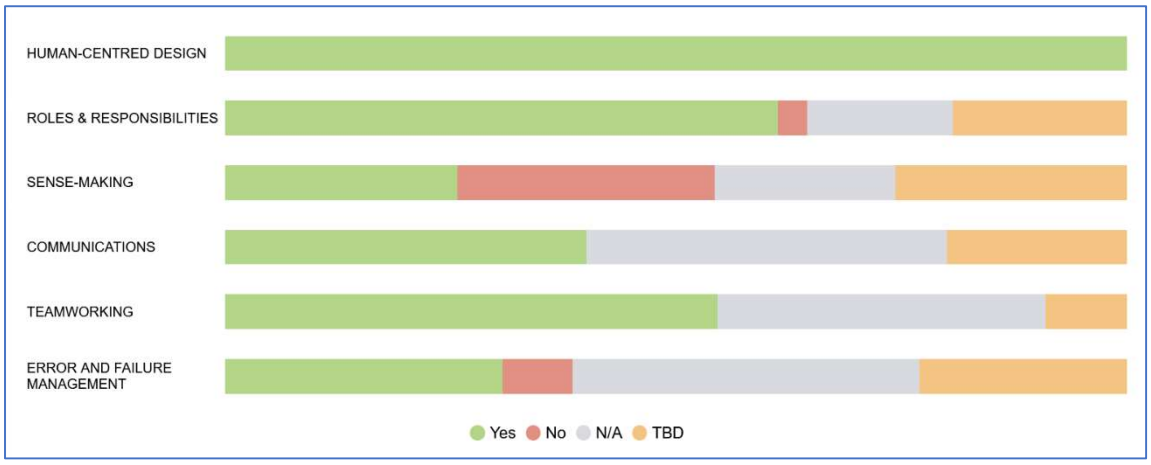

Figure 10. – Overview of results of requirements evaluation for UC1 post Val1.

The Human Factors requirements evaluation process for UC1 has resulted in a number of refinements to the HAT design, including: AI display aspects related to the 'on/off' status of the AI support; use of aural SA directional guidance; application of workload measures; consideration of how to better maintain a strategic overview during the emergency; consideration of value of prioritisation by AI during the emergency; pilot trust issues with the AI; consideration of the utility of personalisation of the AI to individual pilots; consideration of potential interference of the AI support with other alerts during an emergency; and use of HAZOP for identification of potential failure modes and recovery/mitigation measures. More generally it raised a number of issues that could be tested in Val2.

 doi:10.20944/preprints202501.0974.v1

The Human Factors requirements set has also been applied to three other use cases: a Tower use case (UC4: TRL5-6 EASA Category 2A); a second cockpit use case (UC2: TRL 4-5, EASA Category 2A/2B), and an airport safety intelligence support system (UC5: TRL 8, EASA Category1A). In all cases so far the requirements evaluation process was found by the participating design teams to be useful and even provocative in elaborating and refining the concept: e.g. in UC2 on Roles & Responsibilities and Sense-Making; in UC4 on the detail of Human-AI Teaming interaction as well as Error and Failure Management, and in the airport use case on Sense-Making (including the usability and explainability of the system) and Organisational Readiness (e.g. staff allocation to the AI, new competencies and training requirements, etc.). Overall, the requirements evaluation process tends to enrich the collective understanding of the HAT design and intended operational modes of use. There is an intention to apply the requirements set to one final Urban Air Mobility use case (TRL 1-2, EASA Category 3A) when its CONOPS reaches sufficient maturity. All results of these applications will be published at the end of the HAIKU project at the end of 2025.

5. Conclusions

The review section of this paper, charting the evolution of both AI and Human Factors, has argued that there is a need for a new or upgraded approach to the determination and evaluation of the Human Factors adequacy for future Human-AI Teaming systems in commercial aviation. A comprehensive and detailed Human Factors requirements system comprising eight Human Factors areas and 165 detailed requirements questions, has been developed and applied to four contemporary HAT research use cases. The application of the requirements approach has been found to be both relevant and useful by the use case owners and design teams, and has resulted in enhancements and refinements to the HAT system designs. In this sense, the approach appears to be fit for purpose, and scalable to projects of differing levels of AI autonomy and design maturity.

However, this is but a preliminary step. There is much research to do and experience to be gained with HAT systems. Certain Human Factors areas stand out as priorities for research, e.g. Human-AI Teamworking arrangements and AI error and failure management. As more HAT prototypes are developed and tested, expertise of how best to design HATs will accrue, and the requirements can evolve in tandem, even adding new HF areas or sub-areas where advisable.

As a final conclusion, all the HAIKU use cases are augmenting human performance and capabilities, rather than seeking to supplant human capabilities in aviation systems. This is not only to maintain human agency and final authority, but also because it appears to be the best way to optimize overall aviation transportation system safety and performance. Human-AI Teaming therefore appears a useful and potentially valuable research and development avenue to pursue.

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

Acknowledgments: The author would like to acknowledge the support of the ENAC UC1 Team, in particular Jean-Paul Imbert and Alexandre Duchevet, as well as Roberto Venditti and Nikolas Giampaolo of Deep Blue and Miguel Villegas Sanchez of Skyway for their support on UC4, Jaime Diaz-Pineda (Thales Avionics), Ricardo Reis and Anais Villani (Embraer), Théodore Letouze and Charles Dormoy (Bordeaux University: ENSC and CATIE respectively) for their support on UC2, and Ryan Elliott and Liam Bolger of London Luton Airport for their support on UC5.

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

References

1. Turing, A.M. (1950) Computing machinery and intelligence. Mind, 49, 433-460. https://doi.org/10.1093/mind/LIX.236.433
2. 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.

3. Dasgupta, Subrata (2014). It Began with Babbage: The Genesis of Computer Science. Oxford University Press. p. 22. ISBN 978-0-19-930943-6.
4. Burkov, A. (2019) The Hundred Page Machine Learning Book. ISBN: 199957950X
5. Mueller, J.P., Massaron, L. and Diamond, S. (2024) Artificial Intelligence for Dummies (3rd Edition). Hoboken New Jersey: John Wiley & Sons.
6. Lighthill, J. (1973) Artificial Intelligence: a General Survey. Artificial Intelligence – a paper symposium. UK Science Research Council. http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm
7. https://www.perplexity.ai/page/a-historical-overview-of-ai-wi-A8daV1D9Qr2STQ6tgLEOtg
8. OpenAI. 2023. GPT-4 Technical Report. ArXiv preprint abs/2303.08774 (2023). https://arxiv.org/abs/2303.08774
9. Huang, L. Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, J., Qin, B., and Liu, T. (2024) A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Question. ACM Transactions on Information Systems. https://doi.org/10.1145/3703155
10. Weinberg, J., Goldhardt, J., Patterson, S. and Kepros, J. (2024) Assessment of accuracy of an early artificial intelligence large language model at summarizing medical literature: ChatGPT 3.5 vs. ChatGPT 4.0. Journal of Medical Artificial Intelligence, 7, 33.
11. Grabb, D. (2023) The impact of prompt engineering in large language model performance: a psychiatric example. Journal of Medical Artificial Intelligence, Vol 6, paper 20.
12. Hable, I., Sujan, M. and Lawton, T. (2024) Moving beyond the AI sales pitch – Empowering clinicians to ask the right questions about clinical AI. Future Healthcare Journal, 11, 3. https://doi.org/10.1016/j.fhj.2024.100179
13. DeCanio, S. (2016) Robots and Humans – complements or substitutes? Journal of Macroeconomics, 49, 280-291.
14. Defoe, A. (2017) AI Governance » A Research Agenda. Future of Humanity Institute. https://www.fhi.ox.ac.uk/ai-governance/#1511260561363-c0e7ee5f-a482
15. Dalmau-Codina, R. and Gawinowski, G. (2023) Learning With Confidence the Likelihood of Flight Diversion Due to Adverse Weather at Destination.IEEE Transactions on Intelligent Transportation. http://dx.doi.org/10.1109/TITS.2023.3235741
16. Dubey, P., Dubey, P., and Hitesh, G. (2025) Enhancing sentiment analysis through deep layer integration with long short-term memory networks. Int. J of Electrical and Computer Engineering, 15, 1, 949-957.
17. Luan, H., Yang, K., Hu, T., Hu, J., Liu, S., Li, R., He, J., Yan, R., Guo, X., Qian, X. and Niu, B. (2025) Review of deep learning-based pathological image classification: From task-specific models to foundation models. Future Generation Computer Systems, Vol 164. https://doi.org/10.1016/j.future.2024.107578
18. Elazab, A., Wang, C., Abdelaziz, M., Zhang, J., Gu, J., Gorriz, J.M., Zhang, Y. and Chang, C. (2024) Alzheimer's disease diagnosis from single and multimodal data using machine and deep learning models: Achievements and future directions. Expert Systems with Applications, Vol 255 Part C. https://doi.org/10.1016/j.eswa.2024.124780
19. Lamb, H., Levy, J. and Quigley, C. (2023) Simply Artificial Intelligence. DK Simply Books, Penguin: London.
20. Macey-Dare, R. (2023) How Soon is Now? Predicting the Expected Arrival Date of AGI- Artificial General Intelligence. 1204 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4496418
21. Hawking, S. (2018) Brief answers to the big questions. Hachette: London, pp.181-196.
22. HAIKU EU Project (2023) https://haikuproject.eu/ see also https://cordis.europa.eu/project/id/101075332
23. SAFETEAM EU Project (2023) https://safeteamproject.eu/1186
24. Eurocontrol (2023) Technical interchange meeting (TIM) on Human-Systems Integration. Eurocontrol Innovation Hub, Bretigny. https://www.eurocontrol.int/event/technical-interchange-meeting-tim-human-systems-integration
25. European Parliament (2023) EU AI Act: first regulation on artificial intelligence. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
26. Duchevet, A., Dong-Bach V., Peyruqueou, V., De-La-Hogue, T., Garcia, J., Causse, M. and Imbert, J-P. (2024) FOCUS: An Intelligent Startle Management Assistant for Maximizing Pilot Resilience. ICCAS 2024, Toulouse, France.
27. Bjurling, O., Müller, H., Burgén, J., Bouvet, C.J. and Berberian, B. (2024) Enabling Human-Autonomy Teaming in Aviation: A Framework to Address Human Factors in Digital Assistants Design. J. Phys.: Conf. Ser. 2716 012076
28. Duchevet, A., Imbert, J-P., De La Hoguea, T., Ferreirab, A., Moensb, L., Colomerc, A., Canteroc, J., Bejaranod, C. and Rodríguez Vázquez, A.L. (2024) HARVIS: a digital assistant based on cognitive computing for non-stabilized approaches in Single Pilot Operations. In 34th Conference of the European

Association for Aviation Psychology, Athens, Transportation Research Procedia 66 (2022) 253–261. https://doi.org/10.1016/j.trpro.2022.12.025
29. Minaskan, N., Alban-Dormoy, C., Pagani, A., Andre, J-M. and Stricker, D. (2022) Human Intelligent Machine Teaming in Single Pilot Operation: A Case Study," in Augmented Cognition, Bd. 13310, Cham, Springer International Publishing, pp. 348–360.
30. Kirwan, B., Venditti, R., Giampaolo, N. and Villegas Sanchez, M. (2024) A Human Centric Design Approach for Future Human-AI Teams in Aviation. In Ahram, T., Casarotto, L. and Costa, P. (Eds) Human Interactions and Emerging Technologies IHIET 2024, Venice, August 26-28. DOI: http://doi.org/10.54941/ahfe1005464
31. 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
32. 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
33. 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
34. Kirwan, B. (2024) 2B or not 2B? The AI Challenge to Civil Aviation Human Factors. In Contemporary Ergonomics and Human Factors 2024, Golightly, D., Balfe, N. and Charles, R. (Eds.). Chartered Institute of Ergonomics & Human Factors. ISBN: 978-1-9996527-6-0. Pages 36-44.
35. Kilner, A., Pelchen-Medwed, R., Soudain, G., Labatut, M. and Denis, C. (2024) Exploring Cooperation and Collaboration in Human AI Teaming (HAT) – EASA AI Concept paper V2.0. In 14th European Association for Aviation Psychology Conference EAAP 35, 8-11 October, Thessaloniki, Greece.
36. https://en.wikipedia.org/wiki/Ergonomics
37. Nick Joyce. (2013). "Alphonse Chapanis: Pioneer in the Application of Psychology to Engineering Design". Association for Psychological Science. https://www.psychologicalscience.org/observer/alphonse-chapanis-pioneer-in-the-application-of-psychology-to-engineering-design
38. Meister, D. (1999) The History of Human Factors and Ergonomics. CRC Press Imprint: Abingdon, UK.
39. Fitts P M (1951) Human engineering for an effective air-navigation and traffic-control system. Columbus, OH: Ohio State University. https://apps.dtic.mil/sti/citations/tr/ADB815893
40. Chapanis, A. (1965). On the allocation of functions between men and machines. Occupational Psychology, 39(1), 1–11.
41. De Winter, J.C.F. and Hancock, P. A. (2015) Reflections on the 1951 Fitts list: Do humans believe now that machines surpass them? Paper presented at 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and theAffiliated Conferences, AHFE 2015. Procedia Manufacturing 3 (2015) 5334 – 5341.
42. Helmreich, R.L., Merritt, A.C. and Wilhelm, J.A. (1999) The Evolution of Crew Resource Management Training in Commercial Aviation. International Journal of Aviation Psychology, 9 (1), 19-32.
43. https://en.wikipedia.org/wiki/Crew_resource_management
44. https://skybrary.aero/articles/team-resource-management-trm
45. https://en.wikipedia.org/wiki/Maritime_resource_management
46. Norman, D.A. (2013). The design of everyday things (Revised and expanded editions ed.). Cambridge, MA London: The MIT Press. ISBN 978-0-262-52567-1.
47. Norman, D.A. (1986). User Centered System Design: New Perspectives on Human-computer Interaction. CRC. ISBN 978-0-89859-872-8.
48. Schneiderman, B. and Plaisant, C. (2010) Designing the User Interface: Strategies for Effective Human-Computer Interaction. Addison-WesleyISBN: 9780321537355.
49. Schneiderman, B. (2022) Human-Centred AI. Oxford University Press: Oxford.
50. Gunning, D. and Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine, [online] 40(2), pp.44–58. doi: https://doi.org/10.1609/aimag.v40i2.2850
51. Hollnagel, E. and Woods, D.D. (1983). "Cognitive Systems Engineering: New wine in new bottles". International Journal of Man-Machine Studies. 18 (6): 583–600. doi: https://doi.org/10.1016/S0020-7373(83)80034-0
52. Hollnagel, E. & Woods, D.D. (2005). Joint Cognitive Systems: Foundations of Cognitive Systems Engineering. CRC Press.
53. Hutchins, Edwin (July 1995). "How a Cockpit Remembers Its Speeds". Cognitive Science. 19 (3): 265–288.
54. Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775-779. https://doi.org/10.1016/0005-1098(83)90046-8
55. Endsley, M.R. (2023) Ironies of artificial intelligence. Ergonomics, 66, 11.
56. Sheridan T B and Verplank W L (1978) Human and computer control of undersea teleoperators. Department of Mechanical Engineering, MIT, Cambridge, MA, USA.

57. Sheridan, T.B. (1991) Automation, authority and angst – revisited. In: Human Factors Society, eds. Proceedings of the Human Factors Society 35th Annual Meeting, Sep 2-6, San Francisco, California. Human Factors & Ergonomics Society Press. 1991, 18–26.
58. Parasuraman, R., Sheridan, T.B., & Wickens, C.D. (2000). A Model for Types and Levels of Human Interaction with Automation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30(3), 286-297. https://doi.org/10.1109/3468.844354.
59. Wickens, C., Hollands, J.G., Banbury, S. and Parasuraman, R. (2012) Engineering Psychology and Human Performance. London: Taylor and Francis.
60. Endsley, M.R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors, 37(1), 32-64. https://doi.org/10.1518/001872095779049543.
61. Moray, N. (Ed.) (2013) Mental workload – its theory and measurement. NATO Conference Serioes III on Human Factors, Greece, 1979. Springer, US. ISBN: 9781475708851.
62. Klein, G., Moon, B., and Hoffman, R. (2006) Making Sense of Sensemaking 1: Alternative Perspectives. IEEE Intelligent Systems, 21, 4, 70-73. DOI: https://doi.org/10.1109/MIS.2006.75
63. Rasmussen, J. (1983) Skills, rules, knowledge; signals, signs, and symbols; and other distinctions in human performance models," and IEEE Transactions on Systems, Man, Cybernetics, vol. SMC-13, no. 3,1983.
64. Vicente, K.J. and Rasmussen, J. (1992), Ecological interface design: theoretical foundations. IEEE Transactions on Systems, Man, and Cybernetics, 22, 589-606.
65. Reason J.T. (1990), Human Error. Cambridge: Cambridge University Press.
66. Billings, C.E. (1996). Human-centred aviation automation: principles and guidelines (Report No. NASA-TM-110381). National Aeronautics and Space Administration. https://ntrs.nasa.gov/citations/19960016374
67. Shappel, S.A. and Wiegmann, D.A. (2000) The Human Factors Analysis and Classification System—HFACS; DOT/FAA/AM-00/7; U.S. Department of Transportation: Washington, DC, USA, 2000.
68. Dillinger, T.; Kiriokos, N. NASA Office of Safety and Mission Assurance Human Factors Handbook: Procedural Guidance and Tools; NASA/SP-2019-220204; National Aeronautics and Space Administration (NASA): Washington, DC, USA, 2019.
69. Stroeve, S., Kirwan, B. Turan, O., Kurt, R.E., van Doorn, B., Save, L., Jonk, P., Navas de Maya, B., Kilner, A., Verhoeven, R., Farag, Y.B.A., Demiral, A., Bettignies-Thiebaux, B., de Wolff, L., de Vries, V., Ahn, S., and Pozzi, S. (2023) SHIELD Human Factors Taxonomy and Database for Learning from Aviation and Maritime Safety Occurrences. Safety 2023, 9, 14. https://doi.org/10.3390/safety9010014.
70. IAEA. Safety Culture; Safety Series No. 75-INSAG-4; International Atomic Energy Agency: Vienna, Austria, 1991.
71. Kirwan, B. (2024) The Impact of Artificial Intelligence on Future Aviation Safety Culture. Future Transportation, 4, 349-379. http://dx.doi.org/10.3390/futuretransp4020018
72. Salas, E., Sims, D.E., & Burke, C.S. (2005). Is There a "Big Five" in Teamwork? Small Group Research, 36(5), 555-599. https://doi.org/10.1177/1046496405277134
73. Parasuraman, R. (1997) Humans and Automation: Use, Misuse, Disuse, and Abuse. Human Factors, 39, 2, 230-253.
74. Parasuraman, R. and Manzey, D.H. (2012) Complacency and bias in human use of automation: an attentional integration. Human Factors, 52, 3. https://doi.org/10.1177/0018720810376055
75. Tversky, A. and Kahneman, D. (1974) Judgment under Uncertainty: Heuristics and Biases. Science, Vol. 185, No. 4157, Sep. 27, pp. 1124-1131.
76. Klayman, J. (1995) Varieties of Confirmation Bias. Psychology of Learning and Motivation. 32, 385-418. https://doi.org/10.1016/S0079-7421(08)60315-1
77. Hawkins, F. H. (1993). Human Factors in Flight (H. W. Orlady, Ed.) (2nd ed.). Routledge. https://doi.org/10.4324/9781351218580
78. Leveson, N.G. (2018) Safety Analysis in Early Concept Development and Requirements Generation. Paper presented at the 28th annual INCOSE international symposium. Wiley. https://hdl.handle.net/1721.1/126541
79. Kletz, T. (1999) HAZOP and HAZAN: Identifying and Assessing Process Industry Hazards (4th edition). Taylor and Francis, Oxfordshire, UK.
80. Kirwan, B., Venditti, R., Giampaolo, N. and Villegas Sanchez, M. (2024) A Human Centric Design Approach for Future Human-AI Teams in Aviation. In Ahram, T., Casarotto, L. and Costa, P. (Eds) Human Interactions and Emerging Technologies. IHIET 2024, Venice, August 26-28. DOI: http://doi.org/10.54941/ahfe1005464
81. Hollnagel, E. (2012) FRAM: The Functional Resonance Analysis Method. Modelling Complex Socio-technical Systems. CRC Press, Boca Raton, Florida. https://doi.org/10.1201/9781315255071
82. Hollnagel, E. (2013). A Tale of Two Safeties. International Journal of Nuclear Safety and Simulation, 4(1), 1-9. Retrieved from https://www.erikhollnagel.com/A_tale_of_two_safeties.pdf
83. 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.
84. https://skybrary.aero/articles/just-culture

85. Franchina, F. (2023) Artificial Intelligence and the Just Culture Principle. Hindsight 35, November, pp.39-42. EUROCONTROL, rue de la Fusée 96, B-1130, Brussels. https://skybrary.aero/articles/hindsight-35
86. MARC Baumgartner & Stathis Malakis (2023) Just Culture and Artificial Intelligence: do we need to expand the Just Culture playbook? Hindsight 35, November, pp43-45. EUROCONTROL, Rue de la Fusee 96, B-1130 Brussels. https://skybrary.aero/articles/hindsight-35
87. Dubey, A., Kumar, A., Jain, S., Arora, V., & Puttaveerana, A. (2020). HACO: A Framework for Developing Human-AI Teaming. In Proceedings of the 13th Innovations in Software Engineering Conference on Formerly Known as India Software Engineering Conference (pp. 1-9). Association for Computing Machinery. https://doi.org/10.1145/3385032.3385044
88. Rochlin, G.I. (1996). "Reliable Organizations: Present Research and Future Directions". Journal of Contingencies and Crisis Management. 4 (2): 55–59. ISSN 1468-5973.
89. 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.
90. 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.
91. 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.
92. 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.
93. Koch, M. (1999). The neurobiology of startle. Progress in Neurobiology, 59(2), 107-128. https://doi.org/10.1016/S0301-082(98)00098-7
94. Thackray, R. I., Touchstone, R. M., & Civil Aeromedical Institute. (1983). Rate of initial recovery and subsequent radar monitoring performance following a simulated emergency involving startle. (FAA-AM-83-13). ttps://rosap.ntl.bts.gov/view/dot/21242
95. BEA. (2012). Rapport final, accident survenu le 1er Juin 2009, vol AF447 Rio de Janeiro—Paris. https://bea.aero/docspa/2009/f-cp090601/pdf/f-cp090601.pdf