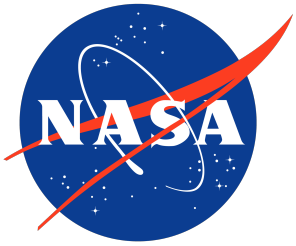
NASA/TM–20250002888



Challenges, Research, and Opportunities for Human–Machine Teaming in Aviation

Mallory S. Graydon
Langley Research Center, Hampton, VA

Jon B. Holbrook
Langley Research Center, Hampton, VA

Natasha A. Neogi
Langley Research Center, Hampton, VA

Jeffrey M. Maddalon
Langley Research Center, Hampton, VA

G. Frank McCormick
Certification Services Inc. (Ret.), Seattle, WA

April 2025

NASA STI Program Report Series

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI Program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI Program provides access to the NTRS Registered and its public interface, the NASA Technical Report Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

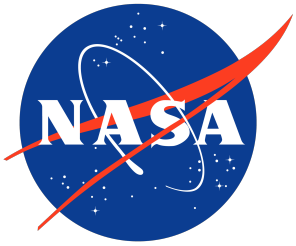
- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers, but having less stringent limitations on manuscript length and extent of graphic presentations.
- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.
- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.
- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.
- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.
- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI Program, see the following:

- Access the NASA STI program home page at `http://www.sti.nasa.gov`

NASA/TM–20250002888



Challenges, Research, and Opportunities for Human–Machine Teaming in Aviation

Mallory S. Graydon
Langley Research Center, Hampton, VA

Jon B. Holbrook
Langley Research Center, Hampton, VA

Natasha A. Neogi
Langley Research Center, Hampton, VA

Jeffrey M. Maddalon
Langley Research Center, Hampton, VA

G. Frank McCormick
Certification Services Inc. (Ret.), Seattle, WA

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

April 2025

Acknowledgments

We thank the branch and directorate reviewers for their feedback on this work. We thank the System Wide Safety Project for funding the successful NASA Research Announcement grants 80NSSC20M0004, 80NSSC20M0005, and 80NSSC20M0080.

Available from:

NASA STI Program / Mail Stop 050
NASA Langley Research Center
Hampton, VA 23681-2199

Abstract

The advent of increasingly autonomous systems (IAS) are likely to herald the proliferation of non-traditional role allocations in human–machine teams (HMTs) in aviation contexts. Ensuring the safety of these novel human machine teaming paradigms and their attendant IAS will require an appropriate body of knowledge created through relevant, reproducible research. In this report, we examine the meaning of teaming; current regulation, standards, and guidance; the knowledge required to build resilient HMTs; and recent research performed under NASA auspices to derive recommendations for an overarching research agenda that will be needed to safely enable the deployment of HMTs with IAS. We identify the need for research into the holistic performance of HMTs; the effect of novel allocations of roles between humans and machines; the ability of humans to provide resilience to unforeseen dangers when acting as a part of these teams; means of assuring the correctness and innocuity of AI/ML components enabling these systems; and the characteristics required for clear, timely, and accurate communication between the humans and machines.

Contents

Executive summary

Novel roles in human–machine teams (HMTs) and the entry of increasingly autonomous systems (IAS) in piloting, air traffic control, or other aspects of aircraft or airspace design, operation and management, maintenance, and retirement raise fundamental questions about how these systems should be designed and assessed for safety. HMTs in aviation should be designed to give rise to safety-producing behaviors and it must be possible to accurately predict both the performance of human team members and the effect of these teams on humans. The strategy of allocating tasks to machines that they are good at can leave humans without the ability to perform their tasks or intervene when required, thereby setting humans up to fail and degrading system safety. We must also better understand the historically overlooked human contributions to everyday aviation safety (and not just to accidents) in order to capture these behaviors in novel HMTs.

In this report, we examine the meaning of teaming; current regulation, standards, and guidance for the adoption of IAS; the knowledge required to build resilient HMTs; and recent research performed under NASA auspices to derive recommendations for a research agenda surrounding the safe deployment of novel HMTs with IAS. We show that fundamental research is required to inform the creation of safe and effective HMTs:

- *The performance of HMTs as a whole.* It will be necessary to develop well-validated, reproducible means of evaluating both the predictability of systems implementing the HMT and the ability of the aircrew to intervene and/or perform their safety role. Using these means, it will be necessary to create a body of knowledge on the crew's ability to perform their safety roles under both nominal and failure conditions.
- *The effects of novel task allocation on resilience.* The complexities of human cognition will lead to unintuitive performance differences across different task allocations. Because aviation will continue to rely on humans for resilience against unanticipated hazards, it is necessary to identify systematic and repeatable means for exploring the function allocation trade-space for HMTs. Using these means, a body of knowledge can be created on crew resilience under different task allocation patterns.
- *The safety implications of AI/ML implementations enabling novel HMT paradigms.* Implementing novel HMT paradigms using machine learning (ML) will require research into well-validated means of enumerating and assessing the unintended behaviors of ML components and ensuring that they operate as specified. It also requires research into the characteristics of explainable ML decisions, which must be clear and provide timely, correct information to humans at a valuable level of abstraction.

Adequately addressing these topics will require an interdisciplinary approach. Agencies and organizations in government, industry, and academia that are involved in research, development, deployment, and operation of HMT systems should actively engage "whole of community" teams with expertise in machine systems, human behavior, system safety, and societal impacts. NASA has a leading role in such research—both in-house and through NRAs—and should maintain a leadership role to help guide the community. As this research matures, NASA should continue to participate in the standards-making process to facilitate research transfer and ensure that standards remain relevant, sufficient, and well-grounded in both research findings and practical experience.

Overarching recommendations

The recommendations of this document—listed in full in Table 2 and referenced in curly braces below—form a research agenda that is aimed at three (potentially overlapping) segments of the aviation community: (1) NASA, (2) the aviation community interested in novel HMT paradigms, and (3) the aviation community interested in deploying AI/ML systems (potentially in HMTs).

NASA should:

1. Leverage, through targeted funding such as NASA Research Announcements (NRAs), expertise in organizations to create a body of knowledge that enables the understanding, design, deployment, and operation of HMTs that create safety-producing behaviors. {9}
2. Continue to participate in the standards-making process as novel HMT concepts mature to ensure that aviation development assurance standards remain relevant, sufficient, and well-grounded in a common body of knowledge based both on research and practical experience. {2}

Agencies and organizations in government, industry, and academia that are involved in research, development, deployment, manufacture, certification, operation, and regulation of aircraft embodying novel HMT concepts should:

1. Develop methodical and reproducible means of evaluating (a) the predictability of the systems and equipment under novel teaming paradigms and (b) the ability of the aircraft crew to intervene and/or perform their safety role. {1}
2. Perform reproducible research studies on the crew's ability to perform their roles under failure conditions, in order to create a body of evidence that can be used to create standards and/or best practices for such assessments. {3}
3. Identify systematic and repeatable means for exploring the function allocation trade space for HMTs and justifying resulting tradeoffs. {5}
4. Actively engage interdisciplinary teams with expertise in machine systems, human behavior, system safety, and societal impacts. {4}

Organizations responsible for the research, development, deployment, assurance, and operation of AI-based HMT systems should:

1. Identify the necessary characteristics of AI-explanations that are (a) clear, (b) correct, (c) timely, (d) enhance situation awareness, and (d) at a valuable level of abstraction to the required end users. {7}
2. Perform research to (a) identify novel risks posed by the introduction of AI; (b) characterize AI-specific behaviors leading to unexpected modes and malfunctions; and (c) evaluate the efficacy of assurance techniques in AI-based HMT systems. {8}
3. Characterize which HMT task allocation paradigms provide the necessary data and experience to enable any other paradigm via reproducible research. {6}

1 Introduction

Increasingly autonomous systems (IAS) incorporate more automation and/or autonomous functions than are in use today, but are not yet fully autonomous [1,2]. The locus of control in these systems is shifting from humans to machines [3,4]. The National Research Council (NRC) report "Autonomy Research for Civil Aviation" calls out the assurance challenge for IA systems—that is, their verification, validation and certification—as a critical barrier to the adoption and use of these systems in the National Airspace System (NAS) [5]. Any safety analyses or verification and validation techniques used to assure IA system will not be able to rely (implicitly or explicitly) on the current human-centric approach to ensuring safety, and thus must address the challenges of novel human–machine interaction paradigms. For example, a single pilot interacting with advanced automation to control a passenger transport aircraft will not have the benefits of a second pilot and contemporary crew resource management (CRM) techniques. Because modern operational practices and aircraft safety assessments implicitly assume both, assessments of crew performance, crew impact, and related aspects of the aircraft design will need to account for a different interaction paradigm. Likewise, a single operator overseeing several high-velocity, high mass cargo delivery aircraft would interact with those aircraft differently than either a contemporary pilot or airline dispatcher, leading again to different safety concerns and obligations.

This document investigates different possible human–machine teaming paradigms and offers recommendations for research and evidence-based safety assurance for aircraft and operations implementing them. We draw our findings and base our recommendations on a multitude of sources. Firstly, research performed under the NASA research announcement (NRA) entitled "Assuring Increasingly Autonomous Systems with Non-Traditional Human Machine Roles" forms the keystone for this work. This NRA has been executing for the past four years and focuses on the evaluation of both traditional and novel human–machine interaction paradigms in a safety context. Secondly, we incorporate insight from current regulation, industry consensus standard-making, and policy and advisory material from civil aviation authorities. We draw on the current state of the art and practice for the use of novel human–machine paradigms in industry, government, and academia. Finally, we briefly consider the human–machine-interaction-related aspects of emerging technologies (e.g, machine learning and artificial intelligence).

1.1 Problem statement

The demand for automation to perform increasingly complex and adaptive tasks in concert with humans is a driving influence in aviation. However, it is difficult to use current human–machine control paradigms to determine the full contribution of the human to the overall safety of such systems under nominal, off nominal and contingency conditions [6,7]. Contingency management for IA systems is regarded as a critical topic. This is because automation is commonly regarded as performing well under nominal conditions, but performing poorly in undefined or unspecified situations, leading to a degradation of overall system performance, safety, and trust [8,9]. Additionally, the methods required to certify IA systems having components whose behavior is not easily predicted are poorly understood.

Frameworks that enable the specification, modeling, analysis, and evaluation of the safety of IAS, including both human and increasingly autonomous agents, will likely be

required in order to advance the state of practice in verification, validation and certification processes [1–3, 6, 7]. Specifically, the ability to identify the artifacts required to demonstrate the IAS's capacity to mitigate system hazards and unanticipated events, as well as potentially generate those artifacts by analytic or other means, is a fundamental necessity in establishing a certification basis for these systems. Note that these artifacts should be able to be used to assess IAS's compliance to safety requirements throughout the system lifecycle (e.g., design, implementation, deployment and retirement).

We wish to focus on three main problems encountered in the adoption of IAS:

1. Lack of requirements that establish the safety of IAS that use novel human–machine role allocations
2. Lack of verification and validation (V&V) methods for IAS possessing non-traditional characteristics (e.g., non-deterministic/adaptive/learning functions)
3. Lack of understanding regarding what evidentiary artifacts are necessary to demonstrate an IAS compliance to safety requirements under novel human–machine role allocations (e.g., lack of established certification bases and means of compliance)

A subset of these challenges were explored through multiple research initiatives, with a specific focus on new or emerging aviation applications, such as urban air mobility (UAM). Simple use cases, comprising the initial state of operation for UAM (including initial autonomy paradigms) were considered as operational examples for the deployment of IASs, with additional complexity being considered via increases in scalability of the operator-vehicle ratio.

1.2 Purpose

The purpose of this document is to provide recommendations that would inform and support a safe and measured implementation of IAS under diverse human–machine teaming paradigms in civil aviation. Failure to implement IAS in a careful and deliberate manner could conceivably reduce safety and efficiency, increase life-cycle costs, and damage the path to viability for a series of emerging aviation operations (e.g., UAM, uncrewed aircraft system (UAS) cargo delivery, etc.). The prospective benefits, the associated costs, and the unintended consequences that are likely to arise from the integration of IAS in civil aviation will not fall equally on all stakeholders, rendering their (public) acceptance potentially problematic.

Specifically, the recommendations in this document aim to enable the formulation of well-considered, human–machine operational paradigms that ensure the safe and harmonious operation of IAS in the NAS. Many of the recommendations in this document center around the research, technology, and operational procedures that may be needed to safely (1) enable these IASs under traditional and non-traditional human–machine role allocations; (2) demonstrate IAS capabilities for crewed, remote, and uncrewed aircraft; (3) predict the system level effects of incorporating IAS into the NAS; and (4) assure IAS through the design, development, operation, maintenance, and retirement life cycle phases. Additionally, a wide variety of organizations possess key expertise and are making advances in technology, regulation, policy, and standards directly related to the potential adoption of IA systems in civil aviation. The analysis of state-of-the-art practices and documentation

produced from industry, civil aviation authorities, and standards development organizations form a significant portion of the observations, findings, and recommendations included in the document.

Objective. The objective of this report is to outline recommendations that help:

1. Characterize the human contribution to safety in current aviation and air traffic systems, including (a) the explicit actions that humans take in order to mitigate hazards, anomalous events, and unanticipated scenarios as well as (b) the implicit actions that maintain safety and avoid potentially hazardous situations that may arise in the future
2. Identify safety requirements associated with human–machine teams deployed in novel role allocations, specifically with respect to new responsibilities allocated to IASs for safety critical tasks
3. Identify and/or outline the development of necessary capabilities and methods for the modeling, analysis, verification and validation of IAS participating in novel human–machine roles
4. Identify and/or evaluate assessment means and methods for determining whether IASs are compliant to their safety requirements at all phases of their lifecycle (e.g., design, implementation, deployment, maintenance and retirement)

In order to consider the integration of IAS into safety critical aviation contexts, expertise was required in (but not limited to) (1) human–machine interaction in a safety critical context, (2) assurance techniques (verification and validation) for complex, safety critical systems, and (3) aviation or air traffic control systems.

1.3 Scope

IAS are characterized by their ability to perform complex mission-related tasks with substantially less human intervention for extended periods of time, sometimes at remote distances. IAS span a range of domains including but not limited to agriculture, aeronautics, building design, civil infrastructure, energy, environmental quality, healthcare and medicine, manufacturing, and transportation. In this document, we will focus on IAS in civil aviation applications.

There are a multitude of human–machine operational paradigms that have been posited in the past few years (or decades) in civil aviation. Industry has proposed a variety of roles for a pilot or operator which span the gamut from the traditional role of an on-board pilot in command of both the aircraft and other crew members (including a "pilot not flying"), to remote pilots (who take the role of the "pilot in command" but do not sit in the cockpit), to operators (who "supervise" or "manage" the flight of the aircraft, either onboard or remotely). Industry has also proposed fully autonomous flights, which do not possess a "pilot" per se. Similarly the ratio of "pilots" to aircraft can range from one-to-one (where each pilot is in command of a single aircraft), to one-to-many (where a pilot is in command of multiple aircraft, to many-to-one (where there are multiple pilots and/or other crewmembers with diverse roles, such as "safety pilot", associated with each aircraft), to many-to-many (where a pool of crewmembers potentially including pilots for different stages of flight are assigned to a pool of aircraft). All of these paradigms are considered in the scope of this

document, including the different numbers of crewmembers associated with aircraft and their range of ability to intervene during different phases and modes of aircraft operation.

In civil aviation, IAS are being envisioned as being integrated into aircraft, air traffic management (ATM), and other ground-based elements of the national airspace system (NAS). However, there are serious, unanswered questions about how to safely integrate these IASs into a well-established, safe, and efficient NAS governed by operating rules that can only be changed after extensive deliberation and consensus [5]. While the document will address research, technology, and operational issues related to the integration of IAS into civil aviation, we will not include regulatory recommendations, nor will we include recommendations related to ethics, economics or governance, although these factors have important influences in the environment in which these systems will be deployed.

1.4 Audience

This document is intended to provide information, analysis, and potential guidance to:

- Research and development organizations such as NASA, in order to help identify research priorities for IASs
- Regulatory, policy, and standards organizations, in order to help synthesize potential avenues for assessment of IASs
- Practitioners (in industry and otherwise), in order to highlight potential barriers and challenges to the adoption of IASs

2 What do we mean by 'teaming'?

Much has been written about automation, autonomy, artificial intelligence, and human–machine teams. Making sense of this is complicated by subtle variations in how these terms are used. In this section, we briefly explain what we mean here by their use and identify some of the characteristics of teams.

2.1 A word on terminology

In this report, the term *human–machine team* (HMT) is used to describe one or more autonomous machine agents and one or more human agents working in tandem interdependently to achieve a collective goal [10]. A survey of the literature reveals a variety of terms used to describe ways in which humans and machines interact, ranging from *human–automation interaction* to *human–autonomy teaming* to *human–artificial-intelligence teaming*. Loose use of terminology can lead to confusion, creating issues when comparing across studies and making it difficult to determine whether a given design aligns with the theoretical concept it is purported to represent. Likewise, clear statements that describe what does, and what does not, comprise a HMT are needed to support efficient progress and avoid the risk of "re-discovering" what has already been discovered in related literature [10].

Returning to the description of HMTs provided above, a *machine agent* is non-organic and powered by a computational algorithm [10]. To function in the role of a "teammate" with a human, a machine system must be able to perform its roles with some degree of operational independence, or autonomy [11]. *Artificial intelligence* describes a form of advanced automation with the capability to "sense and interpret situations, adapt to changes in conditions and the environment, prioritize and optimize based on changes in goals, and refine its abilities through learning" [11, p. 7]. Teammates need some level of autonomy (freedom to act and decide independently of others to support teaming behaviors, as well as the capabilities to scan their environment, analyze it, make decisions based on their assigned goal, and learn from what happens). Today, many systems built with AI software fall short of full realization of these capabilities, but can still be employed to enhance or facilitate operations that are conducted by humans, albeit not as "teammates" [11]. *Interdependence* can refer to activity (i.e., dependence upon another's tasks in one's own role) or outcome (i.e., the extent to which team outcomes have consequences for the individual). Interdependence requires that humans and machine agents leverage their unique capabilities and information for the good of the other team members and the team as a whole [10]. That is, work is performed better in, or even requires, the presence of the other agent(s). A goal refers to a focus of collective effort [10].

In conventional human–machine (or human–automation) interactions, the machine is treated as a tool. When those interactions are fluid, highly interdependent, and aligned toward a collective goal, and when the machine possesses (or least appears to possess) the capabilities to act independently, adapt, and learn, this can engage propensities for social cooperation in the human agent [12]. It is this engagement which elevates the machine agent to being perceived by the human agent as a teammate rather than as a tool.

In any discussion of the performance of human cognitive tasks by artificial agents, it is often necessary (or at least practical) to use the terminology of human cognition to describe what the artificial agent is doing. In the same way, when humans and artificial agents

interact, the terminology of human–human teaming is often used to describe these interactions between human and artificial agents. Use of the same terms to describe information processing by human and artificial agents, however, can mask fundamental differences between those processes. Thus, use of human-cognition terms to describe machine processing should be viewed as a convenient analogy to convey a concept rather than as an indication that the same processes are at work in the machine. The benefits of using the shared terminology include simplified dialog and description of concepts. However, it can be easy to lose sight of the fact that shared language is used as an analogy, which can mask important differences and promote personification of the artificial agent. Ironically, a human proclivity toward personification of technologies that seemingly evince "humanlike" interactions may ease acceptance of machines performing team-like roles (e.g., mental health chatbots [13]).

Personification refers to (mis)attributing human characteristics to nonhuman things. In this case, use of terms describing complex mental activity (e.g., cognition, intelligence, learning, reasoning, understanding, knowing, intending, etc.) comes with a lot of attributions tied to the resulting behaviors (e.g., about how and why those behaviors occurred). Misattribution of the reasons for an artificial agent's behavior can lead to faulty prediction of its future behavior and inaccurate assessment of its capabilities and limitations. Machines may exhibit forms of intelligence and behavior that are qualitatively different—even alien—from those seen in biological agents [14]. Furthermore, people have little to no direct or "true" introspective access to their own cognition, leading to inaccurate or misconceived notions and assumptions of how human mental processes work [15]. Thus, attempts to use these misconceptions about human cognition to design HMTs can have the unintended effect of reducing rather than enhancing the human's ability to perform cognitive work.

Sustaining safe operations in complex, unbounded, and dynamically changing environments requires consideration not only of failure prevention, but also preparing for and recovering from both expected and unexpected failures. While teammate roles in these environments can be well-defined, specific procedures and tasks must be evaluated and performed within the specific context at the time a given action is implemented. This is something that humans successfully do routinely, but these behaviors have not historically been well-documented and are poorly understood. Recognition of and support for these common behaviors, however, is necessary to ensuring effective teaming. This issue is explored in more detail in Section 4.

2.2 Creating successful human–machine teams

What are the factors within HMTs that will lead to successful team outcomes? Given the criticality of human propensities for social cooperation in establishing team situations, it is appropriate to consider the decades of empirical research on human teams as one part of a foundation for HMT research, but the linkage from human-human teaming to human-autonomy teaming "will not be a one-to-one transfer of knowledge" [10, p. 1]. In the context of all-human teams, the following abilities that promote team effectiveness have been identified (e.g., [16, 17]):

- Ability to direct and coordinate the activities of other team members, assess team performance, assign tasks, and develop team knowledge, skills, and abilities

- Ability to develop common understandings of the team environment and apply appropriate task strategies to accurately monitor teammate performance
- Ability to anticipate other team members' needs through accurate knowledge about their responsibilities, capabilities, and limitations
- Ability to adjust strategies and reallocate resources based on information regarding changing conditions gathered from the environment
- Ability to focus on achievement of team goals over individual members' goals
- Ability to engage in conflict resolution around ideas
- Ability to commit to decisions and plans of action and to hold one another accountable for delivering against those plans

Additionally, coordinating mechanisms that enable effective human teamwork include the following (e.g., [16, 18]):

- Common goals
- Interdependence in terms of assigned tasks and outcomes
- Distinctive roles within the team
- Shared knowledge structure of the relationships among the task the team is engaged in and how the team members will interact
- Mutual trust, which is the shared belief that team members will perform their roles and protect the interests of their teammates
- Closed-loop communication, which specifies the exchange of information between a sender and a receiver (irrespective of the medium)

O'Neill et al. suggested that researchers should carefully describe how they designed their teams and autonomous agents to meet the criteria of a HMT. This is based on well-defined criteria that enables cross-study comparisons and highlights the relevance of existing literature [10]. For example, low levels of interdependence would not characterize HMTs but instead represent a form of human-automation interaction, for which a vast literature spanning several decades already exists. To provide a structure for systematic comparisons and organization of the HMT literature, O'Neill et al. propose using the "Input-Mediator-Output" (IMO) model, which has been used to support analysis of the literature on human teams [10, 19]. The IMO framework comprises the following:

I: Inputs to team processes (e.g., level of autonomy, transparency, reliability, team composition, task characteristics, training, etc.)

M: The "how" of team interactions (i.e., team processes such as planning, communication, coordination, conflict management, etc.), along with mediators (i.e., emergent states such as trust, shared situation models, situational awareness, workload, etc.)

O: Team outputs (e.g., individual and team task performance, team viability, individual learning, etc.)

In a review of HMT research in 2022, O'Neill et al found that, in terms of mediators, little has been examined systematically, with most focus being on direct input-to-output mapping without considering mediating mechanisms [20]. Further exploration of mediators and processes of interaction will provide a stronger understanding of "how" HMTs work and, therefore, how to make them more effective [10].

2.3 Rising authority-responsibility mismatches in human machine teams

There has been a rise in the use of novel function allocations, whereby tasks that were formerly performed by humans are now being performed by machines. This has led to a prevalence of mismatches between authority and responsibility. Authority-responsibility mismatches are created when one agent is authorized (has authority) to perform an activity, but a different agent is responsible for its outcome. An authority-responsibility mismatch demands monitoring by the responsible agent that itself requires additional information transfer and taskload [21]. As the authority to execute actions is increasingly being given to machines, the responsibility for the outcomes is still assigned to the human operator, which results in a worrying trend where mismatches will proliferate through novel functional allocations creating many more potential hazards.

Additionally, this gives rise to the question: Is it reasonable for the human to monitor and intervene in these tasks or actions that have been assigned to the machine? It is unclear as to whether it is reasonable, fair, or ethical to expect that the human will be responsible for questionable outcomes predicated on machine actions that are opaque to them as they are pushed farther from the activity being executed.

Finally, we'd like to take a moment to consider whether humans are inherently "bad" at monitoring, a phrase that is frequently used to explain the human's inability to "catch" undesirable systems behavior. The Fitts report [22], which is often cited in such cases, states that "human operators are poor monitors when all they have to do is to watch the performance of automatic equipment." It is noteworthy that the claim in this statement is that humans are poor passive monitors, which is quite distinct from "humans are bad at monitoring." Indeed, there is ample evidence to show that human performance on a wide variety of cognitive tasks is quite different under mentally active versus passive conditions. For example, task-relevant objects and features receive more attention than task-irrelevant ones [23–25]; cognitively engaging activities improve memory performance [26–28]; and "cognitively active" learning is generally superior to "cognitively passive" learning [29]. Taken together, empirical evidence suggests a human monitoring system that is highly tuned to certain characteristics of sensory information, derived from goals, expertise, and to how human sensory and cognitive systems evolved to detect and attend to environmental change [30]. Thus, one should take care to distinguish poor performance resulting from (design independent) fundamental human cognitive limitations versus poor performance resulting from (design dependent) environments, tools, or tasks that are misaligned with how human cognitive systems work.

Section 6 will discuss what has been learned from NASA's NRA partners within the context of the descriptions and teaming framework posed in this section, along with problems with authority and responsibility in novel human-machine teaming paradigms.

3 Current regulation, certification, standards, and guidance

Aircraft and operations employing novel human–machine teaming will need to exist in the aviation regulatory environment. Both assuring safety and achieving certification require understanding this regulatory environment. While the need to change regulation and certification practices is sometimes overstated, some change might be required [31, 32]. Where there is a need, sufficient knowledge, and appropriate experience, standards and guidance can be updated to better address novel aircraft and operations that are nearing maturity.

As a leader in aeronautics research, NASA is helping to guide the evolution of development assurance and safety assessment practices, standards, and guidance. In this section, we describe the current regulatory environment, the aircraft certification process, existing aircraft development and safety assurance processes, and how these processes currently address cockpit automation. We relate findings concerning gaps that must be addressed to enable future aviation concepts and offer recommendations.

3.1 Federal airworthiness standards, certification, and approvals

New civil aircraft types must receive a *type certificate* before instances of that type can be flown. *Civil aviation authorities,* such as the US Federal Aviation Administration (FAA), *approve* aircraft systems when applicants demonstrate that they have complied with applicable regulations in accordance with a *certification basis* negotiated at the outset of the certification process [32]. Certification of an aircraft's type follows and is supported by approval of its systems.

US aviation regulations are given in Title 14 of the Code of Federal Regulations (CFR). Part 25 gives airworthiness regulations for transport-category airplanes[1], with regulations for smaller general aviation craft given in Part 23. Parts 27 and 29 give regulations for normal and transport-category rotorcraft (helicopters), respectively. There are separate parts of 14 CFR for engines (Part 33) and propellers (Part 35). Title 14 also covers general operating rules for aircraft, including Part 91's general operating and flight rules for pilots.

One general-purpose element of the regulations addresses the possibility that installed equipment might fail. Section 1309 of 14 CFR Part 25 requires applicants to: (a) limit the rate at which equipment, systems, and installations enter states that could lead to loss of the aircraft (including ensuring that no single failure leads to a catastrophic failure condition and addressing combinations of latent failures with other failures) and (b) alert the crew to unsafe operating conditions to facilitate corrective action. The development assurance practices described in Section 3.2 are recognized as a means of compliance with this part of the regulations [33–35].

Another general-purpose element addresses designing systems and equipment to facilitate their use by the flight crew. Section 1302 of 14 CFR Part 25 requires designers to make flight deck controls and information intended for the flight crew's use clear, unambiguous, and accessible and to enable awareness of the effects their actions might have. The same section requires the behavior of installed equipment to be "predictable and unambiguous" and "designed to enable the flightcrew to intervene in a manner appropriate to the task." Designers must also, to the extent practicable, design equipment so that appropriately

[1] Transport-category airplanes are fixed-wing aircraft with more than 19 seats or a maximum takeoff weight greater than 19,000 lbs.

skilled flight crew can manage errors resulting from foreseeable, non-malicious use of that equipment. Aircraft and system developers address these regulatory requirements with separate human factors processes, guided where appropriate by industry consensus standards such as those published by SAE International's G-10 *Aerospace Behavioral Engineering Technology* committee [36].

> **Finding 1:** *Designers must design flight deck controls and information intended for the flight crew's use to be clear, unambiguous, and accessible and to enable awareness of the flight crew's actions.*

> **Finding 2:** *Behavior of installed equipment must be (a) predictable and unambiguous and (b) designed to enable the flight crew to intervene in a manner appropriate to the task.*

> **Finding 3:** *Designers must design equipment so that appropriately skilled flight crew can manage errors resulting from foreseeable, non-malicious use of that equipment.*

> **Recommendation 1:** *Organizations responsible for the research, development, deployment, certification, and operation of aircraft embodying novel HMT concepts should perform research to develop methodical and reproducible means of evaluating (a) the predictability of the systems and equipment under the novel teaming paradigm and (b) the ability of the aircraft crew to intervene and/or perform their safety role.* [See findings: 1, 2, 3]

These regulations were written to address the aircraft we have been building—e.g., traditional airplanes and rotorcraft—often in the light of accidents and incidents that revealed a need. In Section 3.2, we discuss the aircraft development and safety assurance processes currently used to build safe aircraft in the context of these regulations, and how they might need to change to address novel human–machine teaming concepts. In Section 3.3, we discuss current human-factors aspects of these processes, and how those might need to change. In Section 4, we discuss the knowledge we'd need to make these changes. In Section 5, we discuss technologies that might be used to implement the machine to be teamed with and how civil aviation authorities envision regulating those technologies. And in Section 6, we discuss research that might inform and create a foundation for taking such future directions as novel HMT with IAS.

3.2 The aircraft development and safety assurance process

Aircraft systems, equipment, and installations might fail to function or malfunction (including both functioning when not intended and functioning in an incorrect or inconsistent manner). Because this can lead to loss of the aircraft and its occupants (and possibly damage to persons or structures in other aircraft or on the ground), airframers and their suppliers conduct safety assessments and apply development assurance to address the possibility of design errors leading to such failures and losses. That is, they (1) analyze the aircraft concepts and designs to identify how systems, equipment, and installations might fail in a way that contributes to loss; (2) design the aircraft to eliminate, minimize, or tolerate such

failure conditions; and (3) apply *development assurance* (e.g., verification and validation) sufficient to ensure that they have done so despite the possibility of design error.

SAE International's *Aircraft and Systems Development and Safety Assessment Committee*, S-18, publishes two documents that give recommended practices for aircraft development and safety assurance [37]:

- ARP4754B, *Guidelines for Development of Civil Aircraft and Systems*, describes the overall process, setting objectives for what must be accomplished
- ARP4761A, *Guidelines and Methods for Conducting the Safety Assessment Process on Civil Airborne Systems and Equipment*, describes ways to meet the safety assessment objectives set out in ARP4754B

These documents are recent updates of widely-used prior versions. ARP4754A (which references ARP4761) is currently recognized as a means of establishing development assurance practices appropriate to addressing development error in the context of development regulations such as 14 CFR 25.1309 (see Section 3.1). We expect the updated versions, ARP4754B and ARP4761A, to be recognized by civil aviation authorities in due time.

ARP4754B (like ARP4754A) describes a development process *v-model* comprising:

1. Aircraft function and requirement development
2. [Development of] aircraft architecture and system functions
3. [Development of] system requirements and architectures
4. Hardware and software implementations and verifications
5. Subsystem verifications
6. System verifications
7. Aircraft verification

During this, staff (a) perform safety assessment; (b) assign *development assurance levels* that indicate how much assurance is needed; (c) capture requirements, including safety requirements; (d) validate requirements; (e) manage configurations; and (f) perform process assurance [38].

Safety assessment comprises:

1. An aircraft functional hazard assessment (AFHA) during aircraft function and requirement development
2. A preliminary aircraft safety assessment (PASA) during development of the aircraft architecture and system functions
3. System functional hazard assessments (SFHAs) during development of the aircraft architecture and system functions
4. Preliminary aircraft safety assessments (PSSAs) during development of system requirements and architectures
5. System safety assessments (SSAs) during subsystem and system verifications
6. An aircraft safety assessment (ASA) during aircraft verification

The AFHA and SFHA identify failure conditions of the aircraft and its systems, identify the consequences for the aircraft, crew, and passengers, and rate the severity of the failure. This assessment includes documentation of the expected crew response to failure, the assessment of which involves coordination with human-factors experts as described in Section 3.3. Processes for conducting AFHA and SFHA are given in ARP4761A [39]. The preliminary safety assessments use analyses such as fault tree analysis (FTA), common mode analysis (CMA), and zonal safety analysis (ZSA) to determine how systems and items might contribute to failure modes [38, 39]. During PASA and PSSA, developers allocate safety requirements to systems and items and set development assurance levels for functions and items (FDALs and IDALs). After the PSSAs, items are developed following appropriate development assurance guidance. For instance, software might be developed in conformance with RTCA DO-178C [40]. During subsystem validation, system validation, and aircraft validation, safety assessments confirm satisfaction of safety requirements.

The ARP4754B and ARP4761A processes were written to be applied to traditional, directly piloted aircraft with automation limited to what might be called a tool rather than a teammate. This focus is a deliberate choice by those documents' authors: they constrained their recommended practices to the kinds of aircraft they had the knowledge and experience to write best practices for.

> **Finding 4:** *Current standards for the development and safety assessment of civil aircraft (e.g., ARP4754B, ARP4761A etc.) are written to be applied to aircraft which are operated according to traditional piloting paradigms (i.e., directly piloted aircraft from within the aircraft cockpit).*

Because of this focus, the aircraft development and safety assessment processes described in ARP4754B and ARP4761A might require adaption for use on the kinds of air vehicles and operations that might employ novel human–machine teaming. For example, consider the case of a single operator remotely controlling several uncrewed aerial vehicles. Because one operator controls several aircraft, aspects of the AFHA and SFHA that consider crew actions and impact on crew might need to be done differently. Systems on these aircraft might be allocated responsibility for actions like "seeing and avoiding" other air traffic under visual flight rules. Some equipment necessary for the control of the aircraft—analogues of what might be installed in current cockpits—might be installed in a ground station. These kinds of differences create gaps between the ARP4754B and ARP4761A recommended practices and what is needed to develop and field such aircraft safely. The S-18A subcommittee of SAE S-18 is developing an aerospace information report, AIR7121, that already identifies several such gaps [41,42]. We anticipate its publication in the coming months.

> **Finding 5:** *SAE International's S-18 committee is analyzing its recommended practices for aircraft development and safety assessment (ARP4754B and ARP-4761A) to identify gaps between those practices and the development assurance needs of novel air vehicles, some of which would employ novel forms of human–machine teaming.*

At the same time, technology for performing hazard assessments and safety assessments continues to be developed. ARP4761A revised ARP4761 to include new analyses that have

gained traction in industry, including model-based safety assessment. SAE S-18 also addresses new techniques that have promise but have not yet become established best practices in aircraft development. For example, S-18 is examining how a new hazard assessment, the System-Theoretic Process Analysis (STPA), might be used in an aviation context [43]. As discussed in Section 6.2, STPA is being adapted to analyze aviation systems that embody novel human–machine teams. SAE is developing an aerospace information report on the use of STPA in aviation, which might be published as early as next year.

> **Finding 6:** *SAE International's S-18 committee is analyzing hazard assessment techniques that might be used to assess aircraft embodying novel human–machine teaming concepts and preparing information reports on their use in an aviation context.*

NASA Langley civil servants have been active in SAE S-18, participating in the production of the ARP4754B and ARP4761A revisions and contributing to the development of the AIR7121 gap analysis, the STPA report, and several other documents. One of us, Dr. Graydon, is a voting member of the committee. This participation has been informally recognized as providing value to the committee's work, with many of Dr. Graydon's suggestions having been accepted as changes to the committee's documents and her insight having been sought on particular matters. We expect that continued NASA participation in SAE S-18 will help the committee to produce updated documents that address future aviation challenges such as those associated with novel human–machine teaming concepts.

Additionally, NASA Langley civil servants have been active participants in ASTM International standards committees. Committees such as ASTM F38 Unmanned Aircraft Systems, F39 Aircraft Systems (for electrical aircraft), F44 General Aviation Aircraft, and AC 337 Autonomy Design and Operations in Aviation all explore standards for non-traditional human machine teaming paradigms. One of us, Dr. Neogi, is NASA's designated voting member of the F44.50 subcommittee on Systems and Equipment, and she is also a founding member of AC 377. She has worked on the ASTM standard F3230-21a, "Standard Practice for Safety Assessment of Systems and Equipment in Small Aircraft", and she is co-leading the working group WK 60748 in developing the "Standard Guide for the application of Systems-Theoretic Process Analysis (STPA) to Aircraft", in order to facilitate the use of emerging hazard assessment methods for novel human–machine teaming paradigms. Similarly, she has provided a detailed review for ASTM F3269-21, "Standard Practice for Methods to Safely Bound Behavior of Aircraft Systems Containing Complex Functions Using Run-Time Assurance", which is a product of AC 377 and provides guidance on the use of runtime assurance in aviation systems. Multiple members of NASA's civil servant workforce regularly provide input to, review of, and commentary on ASTM committee's work products. This creates a whole-of-agency approach to the standards-making process that leverages NASA's expertise in multiple domains and advances NASA's mission in providing guidance to industry both in the United States of America and abroad.

> **Finding 7:** *NASA participation in the making of relevant development assurance standards, such as those published by SAE International's S-18 committee and ASTM International's F44 committee, has resulted in improvements to those documents.*

Recommendation 2: *NASA should continue to participate in the standards-making process as aircraft designs embodying novel HMT concepts mature in order to ensure that aviation development assurance standards remain relevant, sufficient, and well-grounded in both a common body of knowledge based on research and practical experience.* [See findings: 4, 5, 6]

3.3 Human–machine interaction aspects of safety analysis

Equipment failure can negatively impact crew performance, including both the creation of additional workload or by making other tasks more difficult. In some cases, crew are expected to react to notifications and participate in the mitigation of the failure. For example, if pitot tubes ice up and airspeed data is lost, pilots are notified of the failure by aural and visual means, autopilot is disconnected, and the crew must hand-fly the aircraft at pitch and power settings calculated to avoid stall and overspeed conditions.

Safety assessment and development assurance practices like those described in Section 3.2 take into account both the possibility of crew mitigation of failure conditions and the deleterious effects of those failures on humans aboard the aircraft. AFHA and SFHA as defined in ARP4761A require analysts to (a) describe the expected crew response to failure conditions and (b) describe the impact of failure conditions on cockpit crew and other aircraft occupants as well as impact on the aircraft itself as a machine [39]. Both kinds of assessments impact the functional and development assurance requirements assigned to equipment, with the expectation that crew will intervene to effectively mitigate a failure condition acting to lower development assurance requirements aimed at preventing the machine from exhibiting that failure condition [38, 39]. ARP4761A does not describe how to obtain or validate crew response or impact information, and it is assumed that a separate community of practitioners outside of safety assessment perform these functions.

Following the 737 MAX accidents, SAE's S-18 committee decided to address such human considerations in the aircraft development and the safety assessment process [44, 45]. Discussion between SAE S-18's development-assurance-focused membership and SAE G-10's human-factors-focused membership started to characterize the interaction between the relevant practitioners [36, 37, 46]. Because there is no well-established guidance document describing the interactions between human-factors practitioners and the development assurance and safety assessment practices, the S-18H subcommittee of S-18 has begun work on this topic [46]. S-18H has drafted an aerospace information report detailing the interfaces between the human factors team and the AFHA and SFHA processes described in ARP4754B and ARP4761A [38, 39, 47].

Finding 8: *The recommended development assurance and safety assessment practices defined in ARP4754B, ARP4761A, and their predecessors treat the human–machine interaction aspects of these processes mainly as matters addressed by a separate community of practitioners.*

Finding 9: *While airframers address both (a) expected crew responses to and (b) the impact on crew of the failure conditions of aircraft and equipment, they do so according to their own internal processes rather than in conformance with an industry-standard consensus document.*

Finding 10: *The lack of commonly accepted standards that address the interaction between aircraft failures and the crew leads to a gap in being able to describe aspects of human–machine interactions where history is not an adequate guide and what kinds of justification for assessments would be needed in those cases.*

Novel human–machine teaming concepts represent a departure from traditional cockpit allocation of tasks to humans and machines. To the degree that they do, we can expect an increase in the number of crew response and impact inputs to AFHA and SFHA that cannot be established solely by reference to long-established history. That is, there will no longer be historical data that establishes a crew's response to known failures or the impact of those failures on the crew. As in other cases where history is not a sufficient guide, the inputs in these cases will need to come from fresh empirical evidence. This will, in turn, create an increased need for appropriate studies, including, for example, human-in-the-loop studies, to provide this evidence base.

Finding 11: *As aircraft designs begin to implement novel human–machine teaming concepts, there will be an increased need for studies to establish well-evidenced understandings of both how human crew will react to failure conditions and how they will be impacted by failure conditions.*

Recommendation 3: *Organizations responsible for the research, development, deployment, certification, and operation of aircraft should perform reproducible research studies on the crew's ability to perform their roles under failure conditions, in order to create a body of evidence that can be used to create standards and/or best practices for such assessments, especially when historical data is not available or relevant.* [See findings: 8, 9, 10, 11]

4 The missing knowledge needed to build resilient HMTs

Human–machine teams are currently central to performing complex, safety-critical functions on aircraft. Traditional design paradigms with well-developed roles and responsibilities for human crew help to engineer resilience into the system for when "unknown unknowns" are encountered. However, there is a lack of understanding as to why these paradigms work as well as they do, and there is no well-evidenced body of research that enables designers to alter the roles and responsibilities in human–machine teaming paradigms such that they can fully predict the behavior of the new system under all conditions, specifically those that are unforeseen, unplanned, or unanticipated. While the field of crew resource management studies the interactions between human-human teams, the lessons learned from this field may not be directly applicable to human–machine teams. This section explores factors that will be necessary to design human–machine teaming paradigms that exhibit resilience as an emergent property.

4.1 Interdisciplinarity

HMTs represent the union of humans, data, and algorithms. Each of these domains influences the other in both well-understood and unknown ways [5]. Data, which are filtered through algorithms created by humans, influence machine behavior; human interactions are altered by the introduction of machine systems; and in HMTs, agents collectively interact with and influence one another [14]. As humans increasingly interact with machines, those interactions will mediate work and social interactions that can materially impact efficiency, safety, and well-being. These IA machine systems will only grow in complexity, and their increasingly sophisticated interactions with—and impacts on—humans adds the complexities of human individual and social behaviors to the mix. Because of these complexities, "these hybrid human–machine systems pose one of the most technically difficult yet simultaneously most important areas of study for machine behavior" [14, p. 483].

Machine behavior cannot be fully understood without integrated study of algorithms within the human social and societal environments in which the algorithm operates [48]. To study the behavior of "black box" algorithms in real-world settings requires integrating knowledge from across a variety of scientific disciplines. On the one hand, the computer scientists and engineers who create these systems possess the deep expertise necessary to mathematically describe the properties of their algorithms and evaluate their underlying quality and appropriateness of algorithmic techniques used. Social scientists, on the other hand, possess the deep expertise in experimental methods and statistics particular to the study of behavior. While these approaches can be useful to the study of machine behavior (even though the behaviors evinced by the machines may be qualitatively different than human behaviors), the expertise of social scientists is essential to understand the impacts of the IA machine system on humans and society.

> **Finding 12:** *Human machine teaming is an emerging interdisciplinary field of study, requiring specialized expertise in IA machine systems and in human behavior.*

> **Finding 13:** *The agents in human–machine teams will affect each other's behavior in both immediate and long-term ways that are not yet fully understood.*

Recommendation 4: *Organizations responsible for the research, development, deployment, and operation of HMT systems should actively engage interdisciplinary teams with expertise in machine systems, human behavior, system safety, and societal impacts.* [See findings: 12, 13]

4.2 Resilience

Systems that operate successfully in dynamic environments for which the range of possible states or conditions is not fully defined must possess properties that enable sustained operations as conditions change. Hollnagel [49] describes these properties as *resilient performance*: the capability of a system to sustain operations by anticipating, monitoring for, responding to, and learning from expected and unexpected perturbations. To date, function allocations between humans and automation in fielded operational systems have largely relied upon humans to provide this capability for resilient performance as a protective measure in the event of automation failure or if a situation is encountered that was not anticipated or accounted for by automation designers. Indeed, in an analysis of aviation "major incidents" in which pilot interventions successfully mitigated potentially catastrophic events, equipment malfunctions were identified as a threat in 55% of those events, and there was no defined checklist or defined procedure to address those equipment malfunctions in over 45% of those cases (i.e., the failure was not anticipated or accounted for by designers) [50].

As IA systems becoming increasingly more capable of performing tasks previously performed by humans, considerations with regard to allocation of the capability for resilient performance arise. Either the IA system must be designed with a capability to safely perform in conditions that the designers themselves did not anticipate or account for, or that capability must remain with a human user.

The study of human performance in the behavioral sciences has focused predominantly on the study of human failures, limitations, and errors. The identification of systematic errors can inform understanding of the logic by which human cognitive systems work [51]. Indeed, the human performance literature is rich with findings of cognitive failures, paradigms for studying them, and methods to identify, label, and measure them. As a result, well-intentioned efforts to be "data driven" in system design are limited by systematic biases (i.e., favoring "performance errors" over "performance") in which data are collected and analyzed. Human error is increasingly emphasized in product and system design [52], and the field of safety science has largely become the study of human error prevention: "Modern technology has now reached a point where improved safety can only be achieved through a better understanding of human error mechanisms" [53].

While the study of human error provides some information about underlying cognitive mechanisms and boundary conditions of human performance, there is limited direct data about how humans succeed, and the systematic study of human resilient performance is an emerging but growing area of research. A recent analysis of operational aviation data suggests that pilots intervene to address aircraft malfunctions on normal flights over 157,000 times for every time that a human error is implicated in an accident, suggesting, ironically, that behaviors that are far more common in real-world contexts are less studied (and thus less well understood) than less frequent behaviors [54]. Furthermore, because the human's role in preventing failures is rarely acknowledged, it is often assumed that "nominal" performance is free of failure and free of human intervention—an assumption that may be

indicative of selectivity in the data that are collected and analyzed as an imperfect proxy for work-as-done.

Since humans are the predominant provider of resilient performance in today's safety critical operations, the relative lack of systematic study of human resilient performance means that limited data from actual exemplars are available to inform development of such capabilities in IA systems. Challenges associated with designing IA systems to operate beyond what is anticipated or accommodated by the designer is evident in current-day implementations such as Tesla's Autopilot and Full Self-Driving capabilities, for which Tesla notes on their website: "Autopilot and Full Self-Driving capability are intended for use with a fully attentive driver, who has their hands on the wheel and is prepared to take over at any moment" [55]. No credible timeline has been offered by IA system researchers or designers as to when IA systems might be able to reliably demonstrate resilient performance in complex real-world settings, so in contexts that are unbounded, high consequence, and defined by uncertainty, HMTs will rely upon humans' capability for resilient performance for the foreseeable future.

A critical challenge, therefore, in the design of HMTs is to determine which tasks should or should not be allocated to which agents, as well as how and when to make those allocations. With respect to sustaining the capability for human resilient performance, it is necessary to consider interdependencies across tasks, such that task allocations to the machine agent enhance, or at least preserve, the capability for human resilient performance rather than reduce it. Such task allocations might conflict with those derived using substitution-based function allocation strategies (e.g., "Humans-Are-Better-At/Machines-Are-Better-At" [22]), which can paint a design-invariant picture of human capabilities and limitations without an underlying theory of why human performance might be "good" or "bad." Furthermore, such function allocation strategies put the focus on human *or* machine, without consideration of human *and* machine, and thus the processes that support multi-agent interaction.

Conventional wisdom suggesting that humans are inherently "bad at" or "good at" certain cognitive tasks is often not well-supported by the literature on human cognition. For example, evidence for "false memories," in which someone confidently "remembers" an event that did not actually occur, is often viewed as a memory "glitch" and used to support the fragility and fallibility of human memory. However, Howe et al. [56] found that associative problem solving was faster and more accurate when critical information was falsely recalled than when it was not, suggesting that what determines whether the processes by which human memory work have positive or negative consequences depends on how those processes are put to use. Likewise, Cosmides and Tooby [57] demonstrated that simply rewording a probability problem could elicit correct Bayesian reason and eliminate base-rate neglect for a majority of subjects (i.e., from 56% to 4%), despite the widely held view that people are "bad at" intuitive statistics.

What is emerging is a view of human cognition that is highly sensitive to the context in which it is situated. Cosmides and Tooby [58] have posited a theory of "evolutionary psychology" suggesting that human cognitive systems evolved to be functionally specialized, and thus are finely tuned to solve adaptive problems in particular ways and environments. Such a theory infers that environments and tools that are poorly aligned with what human cognitive systems evolved to do can limit cognitive performance, whereas well-aligned environment and tools can enable cognitive performance. Thus, successful function allocation

to support design and implementation of HMTs will likely depend on a nuanced understanding of the adaptive nature of human cognition rather than simplified invariant assumptions about what humans are "good at" or "bad at." This approach is particularly important with regard to human cognitive capabilities that are critical to preserving system safety, such as the capability to demonstrate resilient performance.

> **Finding 14:** *Human–machine teams will rely on humans' capability for resilient performance (i.e., the capability to sustain operations by anticipating, monitoring for, responding to, and learning from expected and unexpected change) for the foreseeable future.*

> **Finding 15:** *Interdependencies across tasks in terms of how they impact human cognitive processing can affect outcomes in ways that can run counter to "conventional wisdom" about human performance.*

> **Finding 16:** *The factors that determine whether human cognitive processes have positive or negative outcomes depend on how those processes are put to use, rather than from an invariant list of what humans are "good at" and "bad at."*

> **Recommendation 5:** *Organizations responsible for the research, development, deployment, certification, and operation of HMT systems should identify systematic and repeatable means for exploring the function allocation trade space for HMTs and justifying resulting tradeoffs.* [See findings: 14, 15, 16]

4.3 Human–machine teaming requires a new design paradigm

IA systems have the potential to tremendously benefit society through efficiencies and enhanced decision-making that these systems can support or provide. At the same time, these benefits may falter without recognition of and addressing the potential pitfalls of incorporating IA systems into everyday human life and work. HMT requires juxtaposing the most complex algorithmic models in existence with the most complex natural systems in existence, and accounting for how they affect and are affected by each other in contexts that are unbounded, high consequence, and defined by uncertainty. Unprecedented challenges such as this require unprecedented solutions—new paradigms of design and development.

> Automate everything that one can ... has been the default strategy used in most systems that have been automated to date, often because increasing efficiency or reducing costs are major driving forces for automation. However, a problem with this strategy is that the human operator is left with functions that the designer finds hard, expensive, or impossible to automate (until a cleverer designer comes around). This approach therefore defines the human operator's roles and responsibilities in terms of the automation. Designers automate every subsystem that leads to an economic benefit for that subsystem and leave the operator to manage the rest [59].

For the reasons outlined extensively in this report, such design strategies are unlikely to work for HMTs, which are prominently featured in the transformation toward increasing levels of digitalization and automation. Success in this endeavor will likely depend

upon neither a "human-centric" nor "machine-centric" focus, but rather a focus centered on the processes by which humans and machines will interact, coordinate, cooperate, and collaborate—that is, a "human-*and*-machine-centric" focus. Just as HMTs represent a union of humans and machines, the research and development necessary to develop successful HMTs will require a union of interdisciplinary experts working together—integrating knowledge across a variety of scientific disciplines [14].

5 AI/ML as a catalyst for novel HMT paradigms

There is substantial interest in "artificial intelligence" (AI) in aviation applications. While this term is somewhat nebulous, it is often applied to machine learning (ML) software: software that is not implemented by algorithms crafted by humans to implement requirements but rather through models derived from data. This interest has resulted in the publication of AI roadmap documents by two civil aviation authorities: the European Union Aviation Safety Agency (EASA) and the US FAA. In this section, we examine a roadmap and guidance from EASA and a roadmap from the FAA. While there are questions about how to provide development assurance for aircraft items implemented with ML, our focus here is exclusively on the HMT aspects of these documents. That is, we are concerned with the potential changes to HMT paradigms that may be catalyzed by the use of AI/ML in aviation systems and how this will impact the safety of the overall system.

5.1 EASA's AI roadmap and guidance

In May of 2023, EASA released its *Artificial Intelligence Roadmap 2.0: Human-Centric Approach to AI in Aviation* [60]. This document lays out EASA's vision for AI in aviation, which includes vehicles and operations that embody novel HMT concepts. In 2024, EASA followed their roadmap up with a concept paper giving *Guidance for Level 1 & 2 Machine Learning Applications* [61]. The guidance is meant to apply only to Level 1 and 2 AI as defined in their AI roadmap, and then only to aircraft items no more critical than ARP4754B/ARP4761A IDAL C [38, 39, 61]. It takes the form of a series of objectives that applicants should meet. Many of these are associated with anticipated means of compliance.

These documents introduce matters that are not directly relevant to our discussion, including both matters of ethics and technical assurance: before deploying ML, it is necessary to ensure (a) that it does not result in harms such as inequitable impact across racial, gender, or similar social lines; (b) that ML-based components perform according to their requirements; and (c) that they do not also exhibit dangerous unintended behavior. While satisfactorily addressing these matters might be a prerequisite to deployment of ML components that would enable novel HMT, they are not our focus here.

> **Observation 1:** *Appropriate ethical analyses, development assurance techniques, and regulatory guidance will be required to satisfy equity-related objectives for ML, especially in light of failures in aviation-related systems [62].*
>
> **Observation 2:** *Development assurance of newer forms of machine learning will require both appropriate techniques and appropriate evidence of the efficacy of these techniques.*
>
> **Observation 3:** *It may be difficult to provide sufficient evidence of the efficacy of development assurance of newer forms of machine learning given the lack of a long operating history in high-integrity applications in diverse operating environments.*

Our focus here is on the HMT-related aspects of these documents. As we discussed in Section 3, the overall aircraft safety assessment and development processes require understanding how the crew interact with the aircraft. This must be clearly articulated, and

engineering staff must have good predicates for predicting behaviors and responses. Accordingly, we focus here on what these documents say about the interactions between the crew and the systems that these documents foresee.

5.1.1 Levels of aviation AI

In their roadmap, EASA define three levels of aviation AI. These are further fleshed out in their concept paper. The levels are:

Level 1. Level 1 AI provides "assistance to the human," and would include things like AI-powered versions of traditional instruments and synthetic vision aids [61].

Level 2. Level 2 AI enables HMT and is distinguished from Level 1 AI by performing "automatic selection and implementation of actions" [61].

Level 3. Level 3 AI performs "advanced automation" and is distinguished from Level 2 AI by having "full authority to make and implement decisions" [61].

> **Finding 17:** *EASA's roadmap outlines 3 levels of aviation AI where Level 1 AI "provides assistance to the human", Level 2 AI "enables automatic selection and implementation of actions" on its part, and Level 3 AI has "full authority to make and implement decisions."*

These levels are each split into A and B categories. The guidance describes the human-automation teaming that characterizes Level 2A and 2B AI as follows:

> The new concept of [human–automation teaming (HAT)] refers to the cooperation and collaboration between the end user and the AI-based system to achieve goals. This HAT concept, depending on the maturity of the AI-based system, may involve a shared understanding of goals, roles and processes (decision-making/problem solving) between the HAT members. It also implies the development of trust and an effective interaction. With this evolution of the AI-based system towards Level 2 AI applications, there is a growing need for guidance on how to effectively introduce and use this concept of HAT.
>
> To this purpose, the guidance makes a clear distinction between the notions of cooperation and collaboration to clarify the definition of the AI leveling as well as to provide novel [means of compliance] (C.4.2):
>
> - *Human-AI cooperation* (Level 2A AI): 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 between the end user and the AI-based system. Communication is not a paramount capability for cooperation.

- *Human-AI collaboration* (Level 2B AI): 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 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.

These are patterns of HMT that do not exist in currently certificated civil aviation aircraft. While similar patterns may exist between human pilots, these may not directly predict patterns between human and machine as described in Section 2. These new patterns of HMT might raise issues of (a) real-time reallocation of tasks between humans and machines, (b) shared human–machine situation awareness, and (c) the potential for AI agents, rather than human pilots, to be the final maker of some safety-critical decisions. We explore issues of shared situational awareness in greater depth in Section 5.1.3.

> **Finding 18:** *EASA's Level 2A AI performs a predefined task allocation with feedback to the end user on decision and action implementation. Level 2B AI works jointly with the end user to achieve predefined shared goals, which implies the capability for a shared situational awareness and real-time task reallocation.*

> **Observation 4:** *The HMT patterns described in conjunction with Level 2A and Level 2B AI have no analogue in existing civil aviation aircraft.*

> **Finding 19:** *Applicants to EASA who seek to develop, deploy, assure, and/or operate AI-based human–machine-teaming systems would benefit from research in areas beyond current engineering practice. Research is needed to develop a means of specifying, modeling, assessing, analyzing, and assuring HMT concepts such as real-time task reallocation, human–machine shared situational awareness, and AI agents making safety-critical decisions unchecked by humans, among others.*

5.1.2 A linear path through the levels

EASA's documents describe a linear path through deployment of AI at these levels, with Level 1 applications expected in 2025 and "consist[ing] in the gradual ramp up to more automated solutions (Level 2 AI)" [60]. This, they predict, will include "single-pilot operations (SPO) in large commercial air transport ... around 2035." Between 2035 and 2050, "in a progressive way," the next steps beyond Level 2 AI will be "advanced automation with human supervision (Level 3A AI) and ultimately without human oversight (Level 3B AI)." Additionally, they speak of the AI "evolving" to a Level 2A application when they define their AI Levels.

> **Finding 20:** *EASA describes a "ramp up" between Level 1 AI and more automated solutions in Level 2 AI. EASA further portrays a "progressive way" by which Level 2 AI will become "advanced automation with human supervision" and eventually no longer requiring "human oversight" in Level 3B.*

It is an open research question as to whether it is a path or a series of diverse end states that exists at the EASA AI Levels. At Level 2, the human-AI teaming concept foresees a partial release of authority to the machine, but under full oversight of the end user. By contrast, Level 3 AI-systems are given authority to make and implement decisions without direct, real-time end user involvement or supervision. As described herein, there are many challenges and currently unanswered questions about how to develop effective and successful HMTs, in no small part because "the human mind is less computer-like than originally realized, and AI is less human-like than originally hoped" [63]. Meeting these challenges will require significant investments that may not apply once there is no longer a human on the team. Furthermore, pressure from investors to show rapid returns on investment may drive network operators to "be interested in the path that realizes full autonomy as quickly as possible" [64]. This raises some questions as to how HMT paradigms might be viewed by system designers: Is a safe and successful HMT desired as a sustained operational goal in its own right, or as a temporary state that should be passed through as rapidly as possible while awaiting approval for full machine autonomy? If the latter, then developers may be less motivated to make large resource investments in HMTs, particularly if they prognosticate that realization of full machine autonomy is rapidly approaching. A detailed accounting of the value of what can be learned from developing effective HMTs, and the potential applicability of what is learned to full machine autonomy, is critical for ensuring appropriate investments in both HMTs and full machine autonomy. Depending on the task(s) or function(s) in question, the end goal for the human–machine paradigm with an AI component may be full automation or it may be some form of human–machine teaming in order to trade off between a variety of properties, qualities, or concerns. Similarly, the overall safety of any two human–machine teaming paradigms with an AI component may be different depending on the environmental and/or operational context. Furthermore, there will be a different burden of evidence and degree of difficulty associated with each human–machine teaming paradigm, all of which should be considered carefully when making design decisions.

> **Finding 21:** *Applicants to EASA who seek to develop, deploy, assure, and/or operate AI-based human–machine-teaming systems would benefit from research to characterize which task allocation paradigms provide the necessary data and experience to enable any other paradigm.*

> **Recommendation 6:** *Organizations responsible for the research, development, deployment, certification, and operation of aircraft should perform reproducible research studies to characterize which task allocation paradigms provide the necessary data and experience to enable any other paradigm.* [See findings: 17, 18, 19, 20, 21]

5.1.3 Shared situational awareness, negotiation, and explainability

The kind of teaming associated with Level 2 AI requires both "situational awareness" shared between the human and machine and the ability to negotiate [61, HF-01 through HF-09]. At Level 2A, the AI must "build its own individual situational representation," "reinforce the end-user individual situational awareness," "identify a suboptimal strategy and propose

through argumentation an improved solution," and "process and act upon a proposal rejection from the end user" [61, HF-01, HF-02, HF-04, & HF-05 and its corollary]. At Level 2B, these objectives are extended with requirements for:

- "The ability to enable and support a shared situational awareness" [61, HF-03]
- "For complex situations under abnormal operations ... identify the problem, share the diagnosis including the root cause, the resolution strategy and the anticipated operational consequences" [61, HF-06]
- "Process and act upon arguments shared by the end user" [61, HF-06 corollary]
- "Detect poor decision-making by the end user in a time-critical situation, alert and assist the end user" [61, HF-07]
- "The ability to propose alternative solution and support its positions" [61, HF-08]
- "The ability to modify and/or to accept the modification of task allocation pattern (instantaneous/short term)" [61, HF-09]

> **Finding 22:** *EASA Level 2 AI requires both "shared situation awareness" between the human and the machine and the ability to negotiate between these two agents.*

To support these and other interactions, EASA's guidance explains that users will require different kinds of explanations of the behavior of different levels of AI. At Level 1, the end user—the pilot—need only know how and how far to take the system's outputs into account in decision-making [61]. But at Level 2A, the AI system is "capable of teaming with an end user" [61]. Because the AI is limited to a

> predefined task allocation pattern, ... communication is not a paramount capability for cooperation. However, informative feedback on the decision and/or action implementation taken by the AI-based system is expected. ... The end user will require explanations in order to cooperate to help the end user accomplish their end goal. A trade-off is expected at design level between the operational needs, the level of abstraction of an explanation and the end-user cognitive cost to process the information received.

At level 2B, "the human and the AI-based system will both communicate and share strategies/ideas to achieve a common goal" [61]. This is the level of a system meant to replace one crew member in what is now a two-pilot cockpit. This kind of teaming

> will require explanations in order to collaborate, negotiate or argument towards a common goals [sic]. A trade-off is expected at the design level between the operational needs, the level of abstraction of an explanation and the end-user cognitive cost to process the information received.

At Level 3A, there is "no permanent oversight from the end user" [61]. But "in order for the end user to override the AI/ML systems' decision, the appropriate level of explanation or information is going to be needed for the good operation of the system." At Level 3B, "the end user is effectively removed from the process" [61]. Thus, EASA sees no need for explainability at the level of the end user.

To provide the needed explainability at Levels 1B, 2A, and 2B, EASA defines a number of objectives. One requires applicants to define the explainability that is needed [61, EXP-12]. This includes defining both the timing and level of abstraction of the explanation [61, EXP-13 & EXP-15]. It includes designing the system to "presen[t] explanations to the end user in a clear and unambiguous form" [61, EXP-11]. And it includes designing the system to "enable the end user to get upon request explanation or additional details on the explanation when needed" [61, EXP-16].

Providing useful explanations to a human co-pilot might be particularly challenging. As has been observed in relation to so-called explainable AI, explanations can be particularly brittle with respect to (a) subtle differences in exactly what is being explained, (b) level of detail in the explanation, (c) multiple factors that interact in producing the effect being explained, and (d) the purpose to which an explanation will be put [65]. An explanation that might help a software developer to understand how software produced a particular output might be unsuited to helping a pilot understand how a machine's capabilities might have changed due to failure or where the pilot's own situational awareness might be inaccurate.

> **Finding 23:** *For an AI-based machine agent to cooperate or collaborate with a human pilot or operator in the manner described by EASA, the machine will need to be able to explain its decisions and recommendations in a manner that is clear and timely, and provides the right information at the right level of abstraction.*

> **Recommendation 7:** *Organizations responsible for the research, development, deployment, assurance, and operation of AI-based HMT systems should perform research to identify what characteristics are necessary for an explanation of an AI-made decision to be clear and provide the correct information at a valuable level of abstraction in a timely fashion to the necessary set of end users or stakeholders and to enhance situation awareness.* [See findings: 22, 23]

It is worth noting that, for speed reasons, the dialogue implied by collaboration in Level 2B applications will likely need to take place in spoken natural language. In the context of the equity issues discussed in Section 5.1, this raises the need for computer speech generation and interpretation issues that perform well enough for safety critical use across a broad range of speakers. Experience with contemporary speech-to-text software suggests that the required performance may be beyond the current state of the art [66].

> **Observation 5:** *Some forms of HMT would require speech generation and interpretation in the machine agent that must provide dependably excellent performance across a broad demographic range of speakers. Experience with contemporary accessibility software suggests that substantial improvement might be needed to provide the required accuracy while achieving equity goals.*

5.2 FAA's roadmap for AI

In July 2024, the FAA published its own *Roadmap for Artificial Intelligence Safety Assurance* [31]. This roadmap covers several matters, but two of biggest relevance to human–

machine teaming are (1) the relationship between humans, advanced automation, and aviation tasks and (2) the knowledge needed to provide effective safety analysis of aircraft and systems using advanced automation.

The FAA's roadmap stresses the need to conceptualize advanced automation realized through AI components in existing aviation terms to the extent possible. That is, we should "place AI within the aviation context rather than aviation within an AI context" [31]. Rather than personify AI as a teammate with whom a pilot can hold a conversation and negotiate responsibility, the document makes a principle of avoiding personification, advising that AI should be treated "as a tool, not a human," and "cannot be a part of crew-resource management." AI might control some flight functions, but accountability for how it behaves must remain with its developers. Indeed, the roadmap motivates its avoidance of personification by stressing the need for clarity about responsibility:

> Personifying AI can erode safety by creating ambiguity on the assignment of responsibility for safe operation. As certain operations, traditionally accomplished by people, are instead accomplished by automation, responsibility shifts from the human operator to the system designer. The system designer must delineate the responsibilities that are assigned to human beings as compared to the requirements that are assigned to systems and tools and must do so in a manner consistent with applicable aviation regulatory requirements and international standards. The responsibility for systems to meet their requirements rests with the system designer and AI developer, not the AI itself.
>
> Aviation experience with complex automation and human factors has highlighted the importance for the human operator, the pilot, to have a solid understanding of the modes, operation, and malfunction of the automation. Personifying AI applications suggests that they have human-like capabilities and potentially unexpected behavior. This contributes to the false impression that the modes, operation, and malfunction would be that of a human, and that AI is an entity which can be responsible. While the normal operation may be intended to automate something that can be performed by a human, the modes and malfunctions are notably different. The safety of future operations depends on the pilot understanding that a system containing AI is just a system and not another human with whom they can reason or negotiate.

Finding 24: *FAA's roadmap states that AI should be treated as a tool; that AI cannot be a part of crew resource management; and that it is important that the human operator have a solid understanding of the modes, operation, and potential malfunctions of the AI-implemented automated agent.*

In the roadmap's vision, AI is not an entity but an implementation of an aircraft function. When describing how it will implement a function and how humans will use it, practitioners should "emphasize clear responsibility assignment and avoid human-centric language to maintain a clear understanding of AI's role and limitations in aviation." The impact of that role should be developed, to the extent possible, in accordance with existing human factors practices and wisdom:

> Additionally, aviation regulations and guidance already address automation and the role of the pilot and other crewmembers. While AI is frequently considered a tool to develop or provide more advanced levels of automation, there is already considerable experience in human factors design principles, evaluation, and training in the aviation context that should be applied. Regulations and guidance continue to improve with experience and as new automation capabilities become feasible.
>
> Issues associated with human-automation integration should be addressed as human factors and automation issues, and not as AI issues, unless the use of AI introduces risks that might not be present with other types of automation.
>
> **Finding 25:** *FAA's roadmap for AI states that issues related to human–machine teaming should be addressed as human factors and automation issues unless the use of AI introduces novel risks that are not present with other forms of automation.*

Where AI is implemented using machine learning, development assurance practices designed for traditional software and electronic hardware may be insufficient, necessitating research into effective methods. Traditional development assurance of hardware and software items follows standards such as RTCA DO-178C and RTCA DO-254 in the context of an ARP4754B and ARP4761 development assurance and safety assessment process [38–40, 67]. These processes center on *traceability*: design proceeds through a series of integrity-preserving transformations to ever more concrete design descriptions, with data and analysis at each stage confirming the integrity of the transformation [32]. But, per the roadmap [31]:

> When AI is used to develop a learned AI implementation, those links are broken; the designer cannot derive the requirements that directly describe the AI implementation and cannot validate them by showing that they provide coverage to the next higher level. The lower-level requirements are the learned AI implementation (an algorithm). For example, it may be a large neural network (NN) with thousands of weights performing basic arithmetic functions that cannot be traced to higher level requirements.

This lack of traceability has consequences for showing both the correctness and innocuity[2] of the item created with ML [31]. Not only are software development assurance practices that stress transformational refinement, like those of DO-178C, inapplicable, but so too are techniques we might use to identify potential failure modes and analyze contributions to them:

> The properties of training datasets and the nature of NN weights are often inscrutable to human review or may deviate significantly from the properties typically analyzed in design assurance. Therefore, validation and verification methods in DO-254 fall short when applied to AI. Another challenge is performing the Failure Mode and Effects Analysis (FMEA), given that the anomalous behavior of the AI implementation may not be identified, or even identifiable.

[2] *Correctness* and *innocuity* are terms of art defined in the *Overarching Properties* [68].

The FAA identified the need for research to address this gap in methods:

- **Aberrant Behavior Review:** The largest challenge for AI is in assuring innocuity .... Research should be conducted and sustained into the types of aberrant or unexpected behavior of learned AI systems to enable system designers to account for their potential manifestation in their systems and their safety mitigations.
- **Safety Assurance Using Formal/Numerical Methods:** This research should assess the effectiveness of potential methods to analyze the performance of AI systems during training and subsequent qualification. The method should provide performance indicators that can be used for assessing the safety of an integrated complex system, in which AI is a component.
- **Safety Assurance Using Systems and Testing Methods:** This research should assess the effectiveness of potential methods to constrain and test the performance of learned AI implementations in integrated systems. For example, can a designer adequately describe the environment in which the system will operate, and can the designer define adequate performance requirements and constraints to assure the integrated system is safe?

Finding 26: *FAA's roadmap for AI identifies 3 prioritized areas of research: (1) aberrant behavior review, (2) safety assurance using formal/numerical methods, and (3) safety assurance using systems and testing methods.*

Finding 27: *Applicants to the FAA who seek to develop, deploy, assure, and/or operate AI-based systems employing HMT would benefit from research to: (1) identify novel risks posed by the introduction of AI that are not present with other forms of automation; (2) characterize AI-specific behaviors leading to unexpected modes and malfunctions of AI-implemented automated agents; and (3) evaluate the efficacy of techniques such as aberrant behavior review, formal methods, testing, and other assurance techniques that may be applicable to AI-based HMT systems.*

Recommendation 8: *Organizations responsible for the research, development, deployment, assurance, and operation of AI-based HMT systems should perform research to: (1) identify novel risks posed by the introduction of AI that are not present with other forms of automation; (2) characterize AI-specific behaviors leading to unexpected modes and malfunctions of AI-implemented automated agents; and (3) evaluate the efficacy of techniques such as aberrant behavior review, formal methods, testing, and other assurance techniques that may be applicable to AI-based HMT systems.* [See findings: 24, 25, 26, 27]

6 Relevant recent research performed under NASA auspices

In this section, we focus on several NASA sponsored research programs executed under the NASA NRA entitled, "Assuring Increasing Autonomous Systems with Non-Traditional Human-Machine Roles". There were three performers: (1) PI Pennsylvania State University (Penn State) with Co-I Iowa State University (Iowa State) (PIs: Prof. Amy Pritchett and Prof. Cody Fleming); (2) PI Collins Aerospace with Co-Is Florida Institute of Technology (FIT) and Soar Technologies, Inc. (PIs: Dr. Jennifer Davis, Prof. Siddartha Bhattacharyya, and Mr. Randall Jones); and (3) PI Massachusetts Institute of Technology (PI: Prof. Nancy Leveson). They executed research devoted to understanding how to capture the human contribution to safety in systems and how to create safety-producing behaviors in HMTs. A brief summary of the three research projects follows in this subsection, and then each project has their highlights broken out in the subsequent subsections in this chapter. For example, the work performed by Penn State and Iowa State can be found in Section 6.1 and Section 6.2. Similarly, the work performed by Collins Aerospace, FIT, and Soar Technologies is further detailed in Section 6.3 and Section 6.4. Finally, the work executed by MIT is expounded in Section 6.5 and Section 6.6. Additionally, NASA research done on the topics of diverse remote piloting paradigms and the pilot's ability to intervene when necessary is briefly summarized in the last two subsections Section 6.7 and Section 6.8.

Assuring Increasingly Autonomous Capabilities with Novel Delegations of Authority and Responsibility. From January 2021 to December 2024, Pennsylvania State University principal investigator Prof. Amy Pritchett and co-investigator Prof. Cody Fleming from Iowa State University performed the research project, *Assuring Increasingly Autonomous Capabilities with Novel Delegations of Authority and Responsibility*, under the contract award number 80NSSC20M0004. The first thread of research, performed by Prof. Pritchett, concerns the ability of well designed HMTs to generate safety producing behavior through actions like monitoring. This research is further detailed in Section 6.1. The second thread of research, performed by Prof. Fleming, focuses on the application of safety assessment to novel HMT paradigms. This thread of research is further examined in Section 6.2.

Prof. Pritchett's group took a concept of operation (ConOps) for an advanced aerial mobility (AAM) small UAS (sUAS) delivery operation and translated it into a computational model. They then further refined the ConOps through simulation utilizing the Work Models that Compute (WMC) framework [69]. The work model was analyzed for work dynamics and to identify timing constraints and information requirements of the actions comprising the work. Additionally, the impact of function allocation on the team and task work dynamics was measured by the task load and information requirements on each agent as well as the overall completion time. Thirdly, the ability to generate monitoring actions during runtime based on mismatches between authority and responsibility in the function allocation allowed the analysis of the task load and information requirements of basic, full and extended forms of monitoring. The computational model and simulation enabled the analysis of the emergent system dynamics of a wide-range of HMT structures as represented by the number of agents, their assumed capability and task limits, and the allocation of authority and responsibility across the team. This research is summarized in Section 6.1.

Secondly, Prof. Fleming's group adapted an existing hazard analysis technique to explore the hazards that a human–machine team might evince. Their research seeks to answer

the following questions.

- What are the design errors that may escape a model-based verification program for safety assurance?
- How to define a design model that is free from those design errors?

We summarize this thread of the research in Section 6.2.

Assured Human Machine Interface for Increasingly Autonomous Systems (AHMI-IAS). From February 2021 to March 2023, Collins Aerospace principal investigator Dr. Jennifer Davies, co-investigator Prof. Siddartha Bhattacharyya at the Florida Institute of Technology, and co-investigator Mr. Randall Jones at Soar Technologies Inc. performed the research project , *Assured Human Machine Interface for Increasingly Autonomous Systems (AHMIIAS)*, under the contract award number 80NSSC20M0005. The performers developed a framework for (1) specifying the roles of a human operator and autonomous co-pilot, (2) verifying that the team satisfies safety properties, and (3) verifying that the autonomous co-pilot meets its requirements. Their framework included: (1) an analyzable architectural model of the human machine team, specified in the Architecture Analysis and Design Language (AADL), (2) specification and analysis of safety properties through the use of the Assume Guarantee REasoning Environment (AGREE), (3) analysis of safety properties under fault conditions through the use of Architectural Modeling and Analysis for Safety Engineering (AMASE), (4) translation of the autonomous system implementation to a formal model through the use of a novel Soar-to-nuXmv translator, and (5) analysis of the requirements for the autonomous system through the use of model checking with nuXmv [70]. Insights related to this work are summarized in Section 6.3.

The framework was used to evaluate assurance for two selected scenarios: an unreliable sensor scenario and an abort landing scenario. The roles of the human operator and autonomous co-pilot in these scenarios were specified and their interactions modeled in AADL/AGREE. Key properties of the piloting team were established with AGREE, and fault scenarios (e.g., delayed pilot response) were reasoned about with AMASE. A prototype autonomous co-pilot agent that performs the selected scenarios was implemented in Soar and simulated in X-Plane with the AgustaWestland AW609 tilt-rotor aircraft model. The performers translated the agent to nuXmv and performed property verification. To demonstrate the robustness of the reasoning framework, the agent was enhanced with a learning function that learns the pilot preference for an early warning threshold for a potentially unreliable sensor, and each of the requirements was proved over the resulting system [71]. The implications of having a learning enabled component in a human–machine team are explored in Section 6.4.

Modeling and Analysis of Safety in New Human-Automation Teaming. From September 2021 to August 2024, Massachusetts Institute of Technology principal investigator Prof. Nancy Leveson performed the research project, *Modeling and Analysis of Safety in New Human-Automation Teaming*, under the contract award number 80NSSC20M0080. The research performed had two goals: (1) develop a framework based on systems theory for creating safe and effective teaming in IA systems and (2) develop and assess a novel new paradigm for the use of top-down modeling and analysis to assure safety, security, and

other system-level properties in the conceptual development of complex, increasingly autonomous systems. To ensure practicality and usefulness, the performers developed these new approaches using the NASA ConOps for UAM systems supporting an intra-metro shuttle mission. To this end, the performers used a new, extended accident causality model called STAMP (System-Theoretic Accident Model and Processes), which they have developed. STAMP treats accidents and losses as resulting from inadequate control over system behavior, and they are a result of a complex process involving interactions among system components, including people, societal and organizational structures, operational activities, and physical system components.

The performers investigate advanced air concepts that incorporate complex teaming and coordination among humans and automation in the context of STAMP and STPA, which identifies potential scenarios that could lead to future accidents. They introduce a novel system-theoretic analytical process to identify unsafe collaborative control actions [72]. We summarize this thread of the research in Section 6.5. This is a part of a broader set of techniques that extend novel STPA hazard analysis to systematically address collaboration. The method rigorously expresses the different ways multiple commands may be unsafe together. Using Systems Theory, it employs abstraction to manage the combinatorial complexity in enumerating control contributions from multiple collaborating components. An algorithm then integrates these concepts into an end-to-end process and is supported by automation to enumerate, refine, prune, and prioritize unsafe combinations of control actions. The output of the method feeds the specification of system requirements to implement safety-guided design starting early in concept development. The performers demonstrate this process on a crewed-uncrewed aircraft teaming case study [73]. Second, they adapted an existing hazard analysis technique to explore the hazards that a human–machine team might create. We summarize this thread of the research in Section 6.6.

6.1 The need for monitoring by both humans and machines

Monitoring is a crucial element of human-autonomy teaming in novel aviation operations. Across a distributed team of agents, cross-checking or monitoring is needed to identify safety-critical situations. The work performed by Prof. Pritchett's group examines the impact that varying information distributions across the team, and when monitoring occurs, has on monitoring performance. The PI applied an agent-based simulation software, WMC, to examine a cookie-delivery ConOps for AAM. Within a day's simulated operation involving several agents, 5 electric vehicles, and 30 missions totaling 60 flights, the case study introduced a degraded vehicle battery that would, if undetected, eventually result in the vehicle departing without sufficient energy to complete its flight.

The monitoring was varied in two ways. First, the algorithm used to assess the scenario was varied to reflect different potential information distributions across the team, in which the monitor would know the schedule, current state, requirements for upcoming flights, and/or predictions of state upon which the day's operations schedule is based. Second, the monitoring was conducted at different times: before, during, and after each flight.

Examining achievable true positive and false positive detection rates for each variation of monitoring, none of the types of monitoring were able to achieve perfect performance. Instead, monitoring accuracy in detecting a degraded battery varies significantly with the information distribution across the team and the timing of when the monitoring is conducted.

Further, monitoring increases agent task load and the amount of information needing to be transferred between agents.

In summary, the following four insights were gained by the Penn State PI during the research.

- *Coordinating taskwork.* Agents can control when they perform some of their tasks, but other actions are time-critical, and need to happen at specific times, requiring coordination. Agents must be able to communicate to coordinate, and agents must be able to interrupt current actions to perform time-critical actions.
- *Ensuring monitoring occurs.* Civil aviation CRM practice has evolved to overtly require monitoring that might be delayed or omitted under time pressure. There are immediate consequences of delaying taskwork, but if everything is going well, missed monitoring usually goes unnoticed. An analogous concept might be needed in new operations.
- *Monitoring and resources.* Monitoring requires effort and information. Systems must be designed to provide this information. When taskload increases, monitoring gets delayed beyond critical points (i.e., monitoring is missed/skipped). In current-day operations, CRM is a protection against this by overtly requiring monitoring.
- *Variable, imperfect monitoring accuracy*. There are theoretical obstacles to 'perfect' monitoring, such taskload, timing, and the operational dynamics. The efficacy of monitoring will vary based on information distribution and on the criterion for measuring monitoring accuracy (e.g., comparing against an operational objective vs against an idealized process model).

  Finding 28: *When an HMT is well constructed and operating smoothly, it can provide several safety-producing behaviors, such as monitoring. The safety-producing benefit of monitoring is enhanced when monitoring takes advantage of disparate viewpoints across the team and the agents properly distribute the necessary information to other agents so that they can perform their monitoring tasks and communicate to enable coordination. However, there are theoretical obstacles to perfect monitoring.*

6.2 Safety assessment

The work performed by the Iowa State team falls under the label of "Safety-guided Model-based Design". Prof. Fleming leverages the rapid advancement of formal methods and model-based Safety Analysis (MBSA) in order to combine them to attempt to verify whether the safety-critical scenarios are adequately addressed by the design solution of a complex HMT system. Prof. Fleming's group identifies a key gap: if specific safety-critical scenarios are not included in the given design solution (i.e., the model) in the first place, the results of MBSA cannot be trusted for safety assurance. To tackle this problem, the Iowa State team developed a new safety-guided design methodology called System Theoretic Process Analysis Plus (STPA+) to complement MBSA. Inspired by STPA, STPA+ treats a system as a control structure, which is particularly fit for systems with complex interactions between humans, machines, and automation. Three methods are developed in STPA+

that tackle the possible omission of safety-critical scenarios caused by (1) incorrectly defined safety constraints, (2) improperly constrained process models, and (3) inadequately designed controllers. In this way, STPA+ attempts to derive an adequately defined design solution as the input to an MBSA verification program and bridge the gap between current MBSA approaches and safety assurance.

6.3 Architectural frameworks for modeling and assessing HMTs

As increasingly autonomous systems take on more responsibility for safety critical decision making and the human–machine role allocation changes, new failure modes may arise. With the increasing complexity and autonomy in these HMT paradigms, traditional verification approaches such as testing will not suffice. The research conducted by Collins Aerospace, Florida Institute of Technology, and Soar Technology Inc. [70] focused on creating an environment whereby they could (1) identify requirements and procedures for safe HMT behaviors; (2) include the human in the model so that human–machine interactions can be analyzed; (3) use formal methods where possible and practical to prove safety requirements are satisfied by (the model of) the system or component; and (4) to evaluate the impact of component faults on the human–machine interactions. The performers used the Assume Guarantee REasoning Environment (AGREE) [74], an annex to the Architecture Analysis and Design Language (AADL) [75] to capture formalized requirements for selected increasingly autonomous agent examples. They also used AADL and AGREE to create a system architecture model with requirements allocated to components to show that the system requirements are satisfied given the component requirements. Behavioral analysis in the presence of faults was performed using the AADL safety annex described in [76]. To show that the system requirements are satisfied by the implementation in Soar, the cognitive architecture framework used to specify learning agents, the performers translated the Soar implementation into the formal model checking environment nuXmv, and then used the nuXmv model checker to verify select safety properties.

This architectural framework provided a centralized repository for requirements, constraints, assumptions, and models that were used throughout the design, development, and safety assessment process for the selected examples. This enabled the performers to manage change rigorously and catch conflicting assumptions as refinements to the roles and responsibilities in the HMT were made. This also allowed for "what if" scenarios to be run to examine the implications of making a change to the HMT paradigm (in terms of roles and responsibilities), and allowed teams of interdisciplinary experts (e.g., requirements engineers, cognitive architecture designers, formal methodists, etc.) to exchange information in a common framework. This enabled the downstream effects of design decisions for the HMT paradigm be considered early in terms of the difficulties that they might cause for the ensuing safety assessment.

> **Finding 29:** *Architectural frameworks enable rigorous change management and interdisciplinary teams to communicate, collaborate, and convey the implications of changes to HMT role and responsibility allocations throughout the design phase for the system.*

6.4 Learning components in HMT paradigms

Integrating a learning subsystem with an existing decision-making agent to create "learning in decision making" is a non-trivial task, even before safety is concerned. Research performed by Collins Aerospace, Florida Institute of Technology, and Soar Technology Inc., attempted to train a Soar cognitive agent to learn a pilot's preferences for receiving alerts related to positioning error, where there are three potential error signals (errors between the GPS, LIDAR, and IMU sensors). Reinforcement learning (RL) was used for this purpose, and Soar cognitive agents allow for preferences to be learned for proposed rules; however this required extra memory for the agent to store data, since RL requires several cycles to process and finalize the computation. Similarly, methods for initializing the threshold values for the rules learned needed to be incorporated into the architecture. Several learning policies associated with RL were evaluated (i.e., Boltzmann with high temperature, Boltzmann with low temperature, simulated annealing decreasing temperature, and Softmax). The policies changed the value of the reward based on the algorithm specific to that policy. The positive or negative value of a reward was based on the input received from the pilot. One of the key challenges was how to know when to stop learning. This problem was solved heuristically by executing 50 trials and observing the preference values of preferred actions. The rule for generating an alert stabilized around 1.70 and the rule about when not to alert stabilized to a value of -1.60, after about 35 trials, as the agent was trained using the scripted pilot response. This stabilization technique is sensitive to the initial values selected, and different heuristics may result in agents trained to different final thresholds.

> **Observation 6:** *ML techniques are sensitive to initial parameter selection and termination of training is more an art than a science.*

In the integrated learning agent, the user's error-alert preferences and responses were incorporated into the larger chain of decision-making, to determine when to disable a sensor that appeared to be dysfunctional. The performers found that more contextualized learning enabled better results in terms of false alarms. They discovered that it was important to make sure that learned preferences for different error signals did not interfere with each other (allowing the system to learn different preferences for different error types). They also discovered that the utility of the agent increased if the learning "decision space" allowed the system to learn different preferences and error tolerances for different mission phases, thereby reducing false alarms.

> **Finding 30:** *The performance of ML models for increasingly autonomous learning-enabled components in HMT will depend on the information shared between agents and may depend on shared contexts.*

6.5 Complex collaborative control in novel HMT paradigms

Human teams collaborate by establishing roles, changing functional authorities, maintaining team cognition, coordinating, and helping one another close control loops. These complex interactions are inspiring novel system concepts to improve human–machine and multi-machine collaboration. However, these new systems challenge existing methods to model,

analyze, design, and assure their safety. As such, few have been fielded in safety-critical domains like aerospace.

Research conducted in [72] introduces a system-theoretic framework to describe multi-controller interactions. This includes a taxonomy of seven structural dimensions that influence such interactions and nine dynamics observed in collaborative control that are defined using STAMP.

An analyzed set of controller interactions in aerospace systems demonstrates the framework and highlights how designers are trying to create more sophisticated systems. Seven classes were identified in a taxonomy describing the structure of interaction between controllers (i.e., types of controllers, hierarchical structure, behavioral intent, connectivity, information exchange, roles and responsibilities, and developmental origins). These seven classes were coupled with the nine identified collaborative control dynamics (i.e., cognitive alignment, lateral coordination, mutually closing control loops, shared authority, transfer of authority, dynamic authority, dynamic hierarchy, dynamic membership, and dynamic connectivity) to evaluate a set of 101 component interactions that are part of aerospace systems. The 101 evaluated systems represented both fielded systems and unfielded systems, e.g., systems in concept development, systems that have been prototyped but not yet fielded, etc. Most of the systems were encountered by the MIT team while reviewing the teaming literature, and the set is not necessarily representative of all possible and actual aerospace systems. However, there are two important takeaways from this analysis. First, there is evidence in the literature that systems are being designed to exhibit each of these complex collaborative control dynamics. And second, of the systems analyzed, those that have not yet been fielded tend to exhibit more of these complex interactions. These points support the argument that causal factors associated with these dynamics must be considered in safety analysis and design.

> **Finding 31:** *HMT paradigms with complex collaborative control structures are frequently found in the design of emerging operations, but they are not yet realized in currently fielded systems.*

6.6 Systematic approaches to analyze safety

A rigorous and systematic approach to analyze safety and enable the safety-guided design of systems that exhibit collaborative control is required to enable systems with complex, collaborative control paradigms. The system-theoretic taxonomy described in Section 6.5 consists of (1) a taxonomy of the structure of interactions between multiple controllers and (2) a set of dynamics observed in collaborative control. It creates the necessary foundation to extend STPA methods needed to systematically identify causal factors associated with these interactions. Researchers at MIT have begun developing such a process as follows. First, a mechanism is developed to incorporate the nine collaborative control dynamics into STAMP control structure models so that they are explicitly considered in hazard analysis. Second, a process is derived from STPA to identify unsafe combinations of control actions between multiple controllers. The procedure systematically considers potential issues involving gaps, overlaps, transfers, and mismatches in authority that are found in teams. It is executed using an abstraction-based algorithm that manages combinatorial complexity and provides automation support. Third, a method is introduced to identify causal factors from

these unsafe control combinations that relate to the collaborative dynamics. This extension to STPA is referred to as STPA-Teaming.

The STPA-Teaming extension was used [73] to analyze several crewed-uncrewed teaming applications that focused on the collaborative interactions between the human pilot and the UAS in the execution of shared mission tasks. This research employed both scoping and abstraction to enumerate, refine, prune, and prioritize the sets of collaborative commands that together can lead the system to enter a hazardous state. The ability to find unsafe combinations of control actions can then be aligned to one (or more) of the nine collaborative control dynamics outlined in the taxonomy. As a result, new causal factors can be found that were previously not systematically considered, as the refinement of the nine factors would expose unsafe control combinations that would have been abstracted away. Currently, heuristics can be used to expose novel causal scenarios, but the researchers have noted that a template for a generic collaborative control structure which is able to express these causal relationships explicitly in the system model would greatly enhance the user's ability to derive such scenarios.

> **Finding 32:** *Structured approaches that identify causal relationships between collaborative control agents enhance the ability of safety analyses to expose novel causal factors that lead to unsafe states in novel collaborative HMT paradigms.*

6.7 Remote piloting and other diverse piloting paradigms

With the advent of autonomous and uncrewed operations, diverse piloting paradigms have emerged. Small UAS vehicles, by necessity, have no humans onboard, and are either remotely piloted or are potentially autonomous. For the remote piloting of these vehicles, the traditional role of a pilot in command along with the number of pilots associated with a vehicle may deviate from convention. Thus, there may be m pilots associated with N vehicles. Given that the transition from 1:1 to m:N inherently changes the role of the human operator and shifts much of their traditional tasking and execution to the supporting systems, the technologies that advance the capabilities, robustness, and resiliency of automation and autonomy are critical for the ability to conduct the envisioned operations safely and dynamically.

For m:N operations to occur in a safe and viable manner, a few key considerations must be addressed. Firstly, the human role and their interaction with the system of aircraft must be understood and designed for. Secondly, the implications of current policy and regulation on the feasibility of such operations must be properly comprehended. The current approval efforts with small UAS and beyond-visual-line-of-sight (BVLOS) operations provide a potential blueprint for the path that lays ahead [77]. Finally, an ability to apply performance- and risk-based criteria to system approvals at the multi-aircraft systems level could provide a more holistic understanding of the m:N operational landscape.

If the deployment of sUAS are seen as a roadmap to the potential adoption of m:N HMT paradigms, then it is important to assess the current limitations and assumptions inherent in such operations today. Non-recreational operations of sUAS encompass various activities, including but not limited to property photography, roof inspections, and aiding nonprofit organizations. Such operations are governed by regulations specified in 14 CFR Part 107,

commonly known as the Small UAS Rule. However, waivers may be sought for operations contrary to Part 107 regulations, subject to FAA approval.

Data integration and visualization methodologies must ensure that operators are able to maintain appropriate levels of situational awareness for all assigned vehicles without negative effects on their workload. To accomplish this, interfaces must provide human operators with enough information and stimulation to understand the status of their vehicles and avoid underload states that may lead to slow response times, but not so much as to overload users and reduce performance [78]. Furthermore, *m*:*N* operations require careful coordination across teams of human operators. Technologies for all teams should be developed to support effective communications, shared situation awareness, and efficient task sharing among the team. Communication between teammates can either be voice-based or text-based, but the approach should ensure compatibility with other task demands [79]. Individuals within a team should be provided with synchronized access to the same data sources [80], and capabilities should be developed to enable task sharing among team members.

Ultimately, gaining approval for novel concepts like *m*:*N* operations that inherently have elements inconsistent with existing regulations and their underlying motivations and assumptions will likely require significant regulatory expertise and time, along with technical innovation, to gain the needed accommodations and approvals. It may be necessary to progress through a series of operations with increasing levels of complexity and potential risk to help motivate and inform the development and implementation of regulatory changes and supporting policies. A key observation is that approval for *m*:*N* operations with lower inherent risk will be easier to obtain, indeed they are already being obtained for sUAS operations conducted under Part 107, albeit still under case-by-case review and waiver issued by the FAA rather than following requirements prescribed directly in the regulations. Approval of more advanced *m*:*N* operations will likely follow a progression of operations with progressively higher levels of complexity and potential risk that support CAAs in the development of appropriate confidence in the enabling technologies, as well inform the development of regulatory updates recognizing the role and requirements of these technologies.

> **Finding 33:** *For human–machine teaming paradigms that have elements inconsistent with existing aviation regulations, a progression of operations with increasing levels of complexity and potential risk may support the achievement of necessary waivers and enable the eventual development and implementation of regulatory change and supporting policy.*

6.8 Pilot's ability to intervene

The pilot's ability to intervene in some of the novel HMT paradigms being proposed will be severely limited in comparison to how aircraft are currently flown. For instance, if the flight controls system overseeing the inner loop control rotor control of an eVTOL octocopter experiences a catastrophic fault, it is unclear what, if anything, a pilot may be able to do to recover the vehicle. Additionally, the response time of the pilot should be taken into account for any action proposed as intervention in contingency scenarios. If there is insufficient time for the pilot to respond, or if the pilot cannot maintain sufficient situation awareness to respond, or if the pilot's skills have degraded to the extent that the response is no longer

tenable, it may not be viable to have the pilot function as a mitigation to the associated fault.

There is an inability to anticipate every future possibility, and out of necessity, resilient humans are relied upon to deal with unforeseen, unplanned, and unanticipated hazards and circumstances. No one has yet found a way to guarantee absence of unpleasant surprises in fielding novel human–machine teaming paradigms with new airborne capabilities, and for some HMT paradigms, the likelihood of unpleasant surprises is high.

> **Finding 34:** *If pilots are relied upon to deal with unforeseen, unplanned, and unanticipated events in HMT paradigms, it is necessary that the pilot have sufficient situation awareness, time to respond, and skill level to execute the intervention.*

Implicitly, the value of pilot flexibility is more valuable than is readily apparent. The airframing OEM typically goes to considerable effort and expense to minimize crew workload. It is true that sometimes in real-life service, some things simply cannot be done if all the rules are constantly followed, but ad hoc workarounds can easily be preludes to missing or misleading information. More broadly, novel human–machine teaming paradigms coupled with exotic new functions and systems in aircraft, even in highly automated functions (perhaps especially in highly automated functions), are disproportionately associated with operator blunders, confusion, and poor awareness of actual system states, leading to expensive losses of one sort or another.

6.9 Summary of research efforts

In summary, research conducted under NASA auspices has significantly added to the body of work surrounding HMT. Significant results include discovering how the use of monitoring in HMTs can foster safety-producing behavior when the monitoring is constructed in ways to enhance HMT performance. For instance, the efficiency of monitoring varies based on how information is distributed across agents in the team and on whether the criterion for measuring monitoring accuracy is based against an operational objective or against an idealized process model. Similarly, it was discovered that learning-enabled components performed better when they were contextualized, that is, when the learning-enabled component was allowed to learn different preferences for different error types in different phases of flight, the false alarm rate for alerting to an error threshold went down significantly. Additionally, a taxonomy of the structure of interactions between multiple controllers was developed, and a set of dynamics observed in collaborative control patterns was documented. This documentation of collaborative control patterns enabled analysts to find unsafe combinations of control actions in a systematic fashion.This type of research is necessary in order to assemble the body of knowledge that will be required to field novel HMTs, as it will provide a solid foundation for any safety assessment that will occur around these systems.

> **Recommendation 9:** *NASA should leverage, through targeted funding such as NRAs, expertise in organizations such as academia, industry and others, to create a body of knowledge that enables the understanding, design, deployment, and operation of HMTs that not only support individual taskwork needs but also support teamwork and help create safety-producing behaviors.* [See findings: 28, 29, 30, 31, 32, 33, 34]

7 Discussion, findings, and recommendation

In this report, we've examined the issue of novel HMT in proposed aircraft and operations, exploring the meaning of "team" and its implications (Section 2), current aviation standardization efforts (Section 3), the knowledge needed to engineer and provide assurance of effective teaming (Section 4), AI/ML roadmaps and guidance from civil aviation regulators (Section 5), and NASA-sponsored research (Section 6). Table 1 lists our findings from this review. Table 2 gives the recommendations we make on the basis of those findings.

In focusing on HMT, we focus on an issue that will be fundamental to future aircraft design. That is, one cannot even begin the safety assessment of an aircraft (as described in Section 3.2) without first knowing its functions, which depend on whether and in which form parts of it will team with a human pilot or operator. And the implications of that choice echoes further: while experience with a form of teaming might help build a knowledge base for assuring similar systems in the future, subtle differences might keep the experience with one form of team from being a useful basis for predictions about another.

In focusing on HMT, however, we have left important matters by the wayside. For example, as noted in Section 5.2, there are important questions to be asked about implementations of AI components of aviation systems:

- *How do we know that the AI system behaves as its designers and the safety assessment team intend it to?*
- *How do we know what unintended behavior the AI system might exhibit?*
- *How do we know the safety implications of that unintended behavior?*
- *What kind of monitoring will be needed for forensic analysis of these systems after an accident or incident?*

As the FAA has observed, these questions highlight the need for research. Once satisfactory answers have been found and demonstrated, standardization activities such as those described in Section 3 can help to codify recommended practice.

Similarly, there are questions related to the sensitivity and fragility of AI or ML components trained on limited sets of data that may not cover the entire operational design domain of the system. Since ML techniques are sensitive to initial parameter selection and termination of training is more an art than a science, the reproducibility and explainability of such components may create a great deal of difficulty as they are integrated into HMTs. Additionally, if these AI or ML components are used outside of their training domain or in different contexts (e.g., users who have different thresholds for alerting, different phases of flight, etc.) the results can be such that the component no longer meets its specification. This will create enormous issues in HMT paradigms where the human is subject to decisions that are opaque and unexplained.

There are many other issues that were deemed out of scope to this document such as ethics, economics, governance, and non-civil aviation platforms. We recognize that these issues inform and influence the civil aviation HMT ecosystem, but we do not consider them materially in our findings or recommendations.

7.1 Findings

This report produced multiple findings based on a survey of literature on HMT, standards related to safety analysis for civil aviation applications, evaluation of current civil aviation authority roadmaps related to HMT, and research produced by multiple performers under a NASA Research Announcement targeted at capturing the human contribution to safety in novel HMT paradigms (with new vehicles in emerging operations). The findings are summarized in the table below.

Table 1: Summary table of findings

1	Designers must design flight deck controls and information intended for the flight crew's use to be clear, unambiguous, and accessible, and to enable awareness of the flight crew's actions.
2	Behavior of installed equipment must be (a) predictable and unambiguous and (b) designed to enable the flight crew to intervene in a manner appropriate to the task.
3	Designers must design equipment so that appropriately skilled flight crew can manage errors resulting from foreseeable, non-malicious use of that equipment.
4	Current standards for the development and safety assessment of civil aircraft (e.g., ARP4754B, ARP4761A etc.) are written to be applied to aircraft which are operated according to traditional piloting paradigms (i.e., directly piloted aircraft from within the aircraft cockpit).
5	SAE International's S-18 committee is analyzing its recommended practices for aircraft development and safety assessment (ARP4754B and ARP4761A) to identify gaps between those practices and the development assurance needs of novel air vehicles, some of which would employ novel forms of human–machine teaming.
6	SAE International's S-18 committee is analyzing hazard assessment techniques that might be used to assess aircraft embodying novel human–machine teaming concepts and preparing information reports on their use in an aviation context.
7	NASA participation in the making of relevant development assurance standards, such as those published by SAE International's S-18 committee and ASTM International's F44 committee, has resulted in improvements to those documents.
8	The recommended development assurance and safety assessment practices defined in ARP4754B, ARP4761A, and their predecessors treat the human–machine interaction aspects of these processes mainly as matters addressed by a separate community of practitioners.
9	While airframers address both (a) expected crew responses to and (b) the impact on crew of the failure conditions of aircraft and equipment, they do so according to their own internal processes rather than in conformance with an industry-standard consensus document.

Table 1 Continued: Summary table of findings

10	The lack of commonly accepted standards that address the interaction between aircraft failures and the crew leads to a gap in being able to describe aspects of human–machine interactions where history is not an adequate guide and what kinds of justification for assessments would be needed in those cases.
11	As aircraft designs begin to implement novel human–machine teaming concepts, there will be an increased need for studies to establish well-evidenced understandings of both how human crew will react to failure conditions and how they will be impacted by failure conditions.
12	Human machine teaming is an emerging interdisciplinary field of study, requiring specialized expertise in IA machine systems and in human behavior.
13	The agents in human–machine teams will affect each other's behavior in both immediate and long-term ways that are not yet fully understood.
14	Human–machine teams will rely on humans' capability for resilient performance (i.e., the capability to sustain operations by anticipating, monitoring for, responding to, and learning from expected and unexpected change) for the foreseeable future.
15	Interdependencies across tasks in terms of how they impact human cognitive processing can affect outcomes in ways that can run counter to "conventional wisdom" about human performance.
16	The factors that determine whether human cognitive processes have positive or negative outcomes depend on how those processes are put to use, rather than from an invariant list of what humans are "good at" and "bad at."
17	EASA's roadmap outlines 3 levels of aviation AI where Level 1 AI "provides assistance to the human", Level 2 AI "enables automatic selection and implementation of actions" on its part, and Level 3 AI has "full authority to make and implement decisions."
18	EASA's Level 2A AI performs a predefined task allocation with feedback to the end user on decision and action implementation. Level 2B AI works jointly with the end user to achieve predefined shared goals, which implies the capability for a shared situational awareness and real-time task reallocation.
19	Applicants to EASA who seek to develop, deploy, assure, and/or operate AI-based human–machine-teaming systems would benefit from research in areas beyond current engineering practice. Research is needed to develop a means of specifying, modeling, assessing, analyzing, and assuring HMT concepts such as real-time task reallocation, human–machine shared situational awareness, and AI agents making safety-critical decisions unchecked by humans, among others.

Table 1 Continued: Summary table of findings

20	EASA describes a "ramp up" between Level 1 AI and more automated solutions in Level 2 AI. EASA further portrays a "progressive way" by which Level 2 AI will become "advanced automation with human supervision" and eventually no longer requiring "human oversight" in Level 3B.
21	Applicants to EASA who seek to develop, deploy, assure, and/or operate AI-based human–machine-teaming systems would benefit from research to characterize which task allocation paradigms provide the necessary data and experience to enable any other paradigm.
22	EASA Level 2 AI requires both "shared situation awareness" between the human and the machine and the ability to negotiate between these two agents.
23	For an AI-based machine agent to cooperate or collaborate with a human pilot or operator in the manner described by EASA, the machine will need to be able to explain its decisions and recommendations in a manner that is clear and timely, and provides the right information at the right level of abstraction.
24	FAA's roadmap states that AI should be treated as a tool; that AI cannot be a part of crew resource management; and that it is important that the human operator have a solid understanding of the modes, operation, and potential malfunctions of the AI-implemented automated agent.
25	FAA's roadmap for AI states that issues related to human–machine teaming should be addressed as human factors and automation issues unless the use of AI introduces novel risks that are not present with other forms of automation.
26	FAA's roadmap for AI identifies 3 prioritized areas of research: (1) aberrant behavior review, (2) safety assurance using formal/numerical methods, and (3) safety assurance using systems and testing methods.
27	Applicants to the FAA who seek to develop, deploy, assure, and/or operate AI-based systems employing HMT would benefit from research to: (1) identify novel risks posed by the introduction of AI that are not present with other forms of automation; (2) characterize AI-specific behaviors leading to unexpected modes and malfunctions of AI-implemented automated agents; and (3) evaluate the efficacy of techniques such as aberrant behavior review, formal methods, testing, and other assurance techniques that may be applicable to AI-based HMT systems.
28	When an HMT is well constructed and operating smoothly, it can provide several safety-producing behaviors, such as monitoring. The safety-producing benefit of monitoring is enhanced when monitoring takes advantage of disparate viewpoints across the team and the agents properly distribute the necessary information to other agents so that they can perform their monitoring tasks and communicate to enable coordination. However, there are theoretical obstacles to perfect monitoring.

Table 1 Continued: Summary table of findings

29	Architectural frameworks enable rigorous change management and interdisciplinary teams to communicate, collaborate, and convey the implications of changes to HMT role and responsibility allocations throughout the design phase for the system.
30	The performance of ML models for increasingly autonomous learning-enabled components in HMT will depend on the information shared between agents and may depend on shared contexts.
31	HMT paradigms with complex collaborative control structures are frequently found in the design of emerging operations, but they are not yet realized in currently fielded systems.
32	Structured approaches that identify causal relationships between collaborative control agents enhance the ability of safety analyses to expose novel causal factors that lead to unsafe states in novel collaborative HMT paradigms.
33	For human–machine teaming paradigms that have elements inconsistent with existing aviation regulations, a progression of operations with increasing levels of complexity and potential risk may support the achievement of necessary waivers and enable the eventual development and implementation of regulatory change and supporting policy.
34	If pilots are relied upon to deal with unforeseen, unplanned, and unanticipated events in HMT paradigms, it is necessary that the pilot have sufficient situation awareness, time to respond, and skill level to execute the intervention.

7.2 Recommendations

The recommendations in this document were carefully constructed and curated to have greatest impact on the human machine teaming community. They lay out a research agenda that will create a body of knowledge that will serve as the foundation on which the design, development, and safety assessment of HMT paradigms can be enabled. The recommendations are targeted at agencies and organizations in government, industry, and academia that are responsible for the research, development, deployment, certification, and operation of human–machine–teaming systems in aviation contexts. Several recommendations are focused specifically on NASA and its ability to maintain a leadership role in this community. Each recommendation is supported by one or more findings. We believe that enacting these recommendations are necessary (but not automatically sufficient) to enable the safe design, development, deployment, and operation of novel HMT paradigms in civil aviation.

Table 2: Summary table of recommendations

	Recommendation	Predicate findings
1	Organizations responsible for the research, development, deployment, certification, and operation of aircraft embodying novel HMT concepts should perform research to develop methodical and reproducible means of evaluating (a) the predictability of the systems and equipment under the novel teaming paradigm and (b) the ability of the aircraft crew to intervene and/or perform their safety role.	1, 2, 3
2	NASA should continue to participate in the standards-making process as aircraft designs embodying novel HMT concepts mature in order to ensure that aviation development assurance standards remain relevant, sufficient, and well-grounded in both a common body of knowledge based on research and practical experience.	4, 5, 6
3	Organizations responsible for the research, development, deployment, certification, and operation of aircraft should perform reproducible research studies on the crew's ability to perform their roles under failure conditions, in order to create a body of evidence that can be used to create standards and/or best practices for such assessments, especially when historical data is not available or relevant.	8, 9, 10, 11
4	Organizations responsible for the research, development, deployment, and operation of HMT systems should actively engage interdisciplinary teams with expertise in machine systems, human behavior, system safety, and societal impacts.	12, 13
5	Organizations responsible for the research, development, deployment, certification, and operation of HMT systems should identify systematic and repeatable means for exploring the function allocation trade space for HMTs and justifying resulting tradeoffs.	14, 15, 16
6	Organizations responsible for the research, development, deployment, certification, and operation of aircraft should perform reproducible research studies to characterize which task allocation paradigms provide the necessary data and experience to enable any other paradigm.	17, 18, 19, 20, 21
7	Organizations responsible for the research, development, deployment, assurance, and operation of AI-based HMT systems should perform research to identify what characteristics are necessary for an explanation of an AI-made decision to be clear and provide the correct information at a valuable level of abstraction in a timely fashion to the necessary set of end users or stakeholders and to enhance situation awareness.	22, 23

Table 2 Continued: Summary table of recommendations

	Recommendation	Predicate findings
8	Organizations responsible for the research, development, deployment, assurance, and operation of AI-based HMT systems should perform research to: (1) identify novel risks posed by the introduction of AI that are not present with other forms of automation; (2) characterize AI-specific behaviors leading to unexpected modes and malfunctions of AI-implemented automated agents; and (3) evaluate the efficacy of techniques such as aberrant behavior review, formal methods, testing, and other assurance techniques that may be applicable to AI-based HMT systems.	24, 25, 26, 27
9	NASA should leverage, through targeted funding such as NRAs, expertise in organizations such as academia, industry and others, to create a body of knowledge that enables the understanding, design, deployment, and operation of HMTs that not only support individual taskwork needs but also support teamwork and help create safety-producing behaviors.	28, 29, 30, 31, 32, 33, 34

8 Conclusion

Novel human machine teaming paradigms in civil aviation promise to create a sea change in the way safety and safety-producing behaviors are designed, developed, evinced, and assessed in the U.S. national airspace system and abroad. The advent of increasingly autonomous systems into safety critical decision making contexts in aviation—piloting, air traffic control, or other aspects of aircraft or airspace design, operation, management, maintenance, and retirement—will act to disrupt the conventional means by which safety is assessed and assured. Specifically, it is unclear in these novel role and responsibility allocations how safety producing behaviors are generated for these complex, interconnected systems.

No design will ever identify *all* of the hazards in any complex unbounded dynamic system, so there will *always* be hazards that are unforeseen or unanticipated. Because the existence of these hazards, which can be referred to as "predictably unavoidable" [81], is known *a priori*, planning should occur to mitigate the situations when those hazards inevitably evince themselves. Currently, our safety processes for the management of these "unknown unknowns" include preparing for and recovering from the resulting failures—that ensue from these "unknown unknown" hazards manifesting themselves. Humans currently represent the primary source of preparation for and recovery from—and prevention of—hazards that are unanticipated during design.

Additionally, early enthusiasm for the adoption of novel, rapidly evolving technologies and innovative role and responsibility allocations between humans and machines may lead to an overestimation of their potential benefits without a due consideration of their attendant costs, such as time, money, and most importantly, unintended consequences to this or other interrelated systems. Several fundamental questions must be answered before novel human–machine paradigms—which may be enabled by emerging technologies such as machine learning for increasingly autonomous systems—can be fielded in safety-critical aviation contexts.

It is important to be able to clearly characterize the nature of the human–machine interaction paradigm being advocated for in the implementation of a given function or task. Without this, it will be impossible to specify the required behaviors that will be necessary to produce the requisite properties (i.e., safety, security, availability, etc.). Additionally, it is critical that the behaviors of the novel human–machine paradigm required for the allotted task or function be predictable. It is very easy to allocate all of the tasks that machines are good at performing to that component, thereby leaving the human without the necessary situational awareness to perform their tasks or intervene when required. This form of human–machine interaction sets the human up for failure, and will degrade the safety of the system. Fundamental research is required on the subject of how human–machine teaming paradigms are characterized, implemented, and evaluated, with respect to internal agent performance and overall emergent properties. It is important to note that these frameworks for characterizing, implementing, and evaluating human–machine interactions should yield methodical and reproducible analysis.

The ability to characterize the human contribution to safety is a key topic of research that must be addressed before current human–machine paradigms can be altered, as otherwise there is no way to predict the effect of the paradigm change on the safety of the system as a whole. Similarly, it is critical that a systematic and repeatable way be developed to ex-

plore, characterize, allocate, and assess potential functional allocations in human–machine teaming paradigms, thereby allowing rational documentation of any tradeoffs to be made. This research has been deemed to be a "whole of community" effort, engaging academia, industry, and relevant government agencies. It is critical that NASA maintain a leadership role in this activity to help guide the community through the initiation of relevant research activities such as NRAs, which complement internal NASA strengths and leverage the community's expertise to access resources not readily available (such as workforce and experimental laboratory facilities).

Any novel paradigm for human–machine interaction or teaming crosses a boundary in knowledge between that which is known (and potentially well understood) to that which is likely unknown. This will rapidly lead to encounters with the unforeseen and the unexpected. Even incremental changes to traditional human–machine teaming paradigms—by deployment in a novel environment or in an emerging operation; or through use of a new technology (e.g., AI/ML) or design—may cause a fundamental change in the knowledge required to design, operate, maintain, and decommission those systems safely. New hazards may emerge, and old hazards may gain new prominence or paths to realization: gaps in knowledge may be impossible to detect until the system has been in service for an extended period of time. Due to the multi-disciplinary nature of human–machine teaming and the potentially shifting locus of control for the making of safety critical decisions, it is vital that interdisciplinary teams of experts across IA systems, human behavior, data sciences and others, be formed to design, develop, deploy, assess, and perform research on systems with novel human–machine teaming paradigms. After all, it may not be readily apparent what is relevant to test or monitor until a new safety concern emerges.

But even the most well-thought-through research will fail to ask *all* of the right questions. The deployment of aircraft or operations employing novel human–machine teaming concepts will encounter hazards, or reveal contributions to hazards, that no one thought of, much less took as seriously as reality warrants. Edge cases and unknown unknowns are a fact of life. Responsible deployment thus both balances the resulting unknown risk and monitors carefully for signs of trouble before these result in accidents or serious incidents. Deployment should prioritize situations where the potential benefits for those who might be put at this unknown risk benefit substantially from taking it. A staged deployment in progressively more risk-tolerant public—good operations (e.g., wildland firefighting, hurricane relief and recovery, etc.)—will enable initial operations with novel HMT paradigms in a viable manner. And the data collection and analysis that surrounds these novel aircraft and operations in progressively more risk-tolerant environments should be planned so as, to the extent practicable, to detect potential issues before they manifest in losses or even near misses. While it is impossible to deploy novel HMT paradigms with IAS without increasing risk, it is important not to damage paths to transition for novel HMT and IAS.

Finally, it is critical to note that the capture of the fundamental research results and insights in a body of standards is crucial to the enabling of novel human machine paradigms in safety critical systems. However, it is impossible to create standards before the fundamental research is performed and that knowledge is gained, as there will remain a great uncertainty surrounding what knowledge will need to be codified. Therefore, engagement in standards is a key part of NASA's portfolio going forward, but it is equally important for NASA to serve as a steward is such activities by raising questions around notions of timeliness and knowledge availability.

References

1. Bhattacharyya, S.; Eskridge, T. C.; Neogi, N. A.; Carvalho, M.; and Milton, S.: Formal Assurance for Cooperative Intelligent Autonomous Agents. *Proceedings of NASA Formal Methods (NFM)*, no. 10811 in Lecture Notes in Computer Science (LNPSE), 2018, pp. 20–36. URL https://dx.doi.org/10.1007/978-3-319-77935-5_2.

2. Neogi, N. A.: Capturing Safety Requirements to Enable Effective Task Allocation Between Humans and Automation in Increasingly Autonomous Systems. *Proceedings of AIAA Aviation*, Washington, DC, USA, June 2016. URL https://ntrs.nasa.gov/citations/20160010169.

3. Bass, E. J.; Bolton, M. L.; Feigh, K.; Griffith, D.; Gunter, E.; Mansky, W.; and Rushby, J.: Toward a Multi-Method Approach to Formalizing Human-Automation Interaction and Human-Human Communications. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, Anchorage, AK, USA, Oct. 2011, pp. 1817–1824. URL https://dx.doi.org/10.1109/ICSMC.2011.6083935.

4. Pritchett, A. R.; and Bhattacharyya, R. P.: Modeling the Monitoring Inherent Within Aviation Function Allocations. *Proceedings of the International Conference on Human-Computer Interaction in Aerospace*, Sept. 2016. URL https://dx.doi.org/10.1145/2950112.296459.

5. Committee on Autonomy Research for Civil Aviation; Aeronautics and Space Engineering Board; Division on Engineering and Physical Sciences; and National Research Council: *Autonomy Research for Civil Aviation: Toward a New Era of Flight*. The National Academies Press, 2014. URL http://www.nap.edu/openbook.php?record_id=18815.

6. Degani, A.; and Heymann, M.: Formal verification of human-automation interaction. *Human Factors*, vol. 44, no. 1, 2002, pp. 28–43. URL https://dx.doi.org/10.1518/0018720024494838.

7. Leveson, N.: An STPA Primer. Available at: http://sunnyday.mit.edu/STPA-Primer-v0.pdf, 2013.

8. Sarter, N. B.; Woods, D. D.; and Billings, C. E.: *Human Factors & Ergonomics*, Wiley, Automation Surprises. 1997.

9. Woods, D. D.: The risks of autonomy: Doyle's catch. *Journal of Cognitive Engineering and Decision Making*, vol. 10, no. 2, June 2016, pp. 131–133. URL https://dx.doi.org/10.1177/1555343416653562.

10. O'Neill, T. A.; Flathmann, C.; McNeese, N. J.; and Salas, E.: Human-Autonomy Teaming: Need for a Guiding Team-Based Framework? *Computers in Human Behavior*, vol. 146, no. 107762, Sept. 2023. URL https://doi.org/10.1016/j.chb.2023.107762.

11. National Academies of Sciences, Engineering, and Medicine: *Human-AI Teaming: State of the Art and Research Needs*. The National Academies Press, Washington, DC, USA, 2021. URL https://dx.doi.org/10.17226/26355.

12. Lyons, J. B.; Sycara, K.; Lewis, M.; and Capiola, A.: Human–Autonomy Teaming: Definitions, Debates, and Directions. *Frontiers in Psychology*, vol. 12, 2021. URL https://dx.doi.org/10.3389/fpsyg.2021.589585.

13. Haque, M. D. R.; and Rubya, S.: An Overview of Chatbot-Based Mobile Mental Health Apps: A Review and Analysis of the Empirical Literature. *Journal of Medical Internet Research mHealth and uHealth*, vol. 11, May 2023. URL https://dx.doi.org/10.2196/44838.

14. Rahwan, I.; et al.: Machine Behaviour. *Nature*, vol. 568, no. 7753, Apr. 2019, pp. 477–486. URL https://dx.doi.org/10.1038/s41586-019-1138-y.

15. Nisbett, R. E.; and Wilson, T. D.: Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review*, vol. 84, no. 3, May 1977, pp. 231–259. URL https://dx.doi.org/10.1037/0033-295X.84.3.231.

16. Salas, E.; Sims, D. E.; and Burke, C. S.: Is there a "Big Five" in Teamwork? *Small Group Research*, vol. 36, no. 5, 2005, pp. 555–599. URL https://dx.doi.org/10.1177/104649640527713.

17. Lencioni, P.: *Overcoming the Five Dysfunctions of a Team: A Field Guide For Leaders, Managers, and Facilitators*. Jossey-Bass, San Francisco, CA, USA, 2005.

18. Hughes, R. L.; and Jones, S. K.: Developing and Assessing College Student Teamwork Skills. *New Directions for Institutional Research*, vol. 149, Spring 2011, pp. 53–64. URL https://dx.doi.org/10.1002/ir.380.

19. Ilgen, D. R.; Hollenbeck, J. R.; Johnson, M.; and Jundt, D.: Teams in Organizations: From Input–Process–Output Models to IMOI Models. *Annual Review of Psychology*, vol. 56, Feb. 2005, pp. 517–543. URL https://dx.doi.org/10.1146/annurev.psych.56.091103.070250.

20. O'Neill, T. A.; McNeese, N. J.; Barron, A.; and Schelble, B.: Human-Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 64, no. 5, 2022, pp. 904–938. URL https://dx.doi.org/10.1177/0018720820960865.

21. Ijtsma, M.; Pritchett, A. R.; and Bhattacharyya, R. P.: Computational Simulation of Authority-Responsibility Mismatches in Air-Ground Function Allocation. *Proceedings of the International Symposium on Aviation Psychology*, Dayton, OH, USA, 2015, pp. 306–311. URL https://corescholar.libraries.wright.edu/isap_2015/55/.

22. Fitts, P. M.; et al.: *Human Engineering for an Effective Air-Navigation and Traffic-Control System.*. National Research Council, Division of Anthropology and Psychology, Committee on Aviation Psychology, Washington, DC, USA, Mar. 1951. URL https://apps.dtic.mil/sti/pdfs/ADB815893.pdf.

23. Shinoda, H.; Hayhoe, M. M.; and Shrivastava, A.: What controls attention in natural environments? *Vision Research*, vol. 41, no. 25-26, 2001, pp. 3535–3545.

24. Werner, S.; and Thies, B.: Is "change blindness" attenuated by domain-specific expertise? An expert-novices comparison of change detection in football images. *Visual Cognition*, vol. 7, no. 1-3, 2000, pp. 163–173.

25. Myles-Worsley, M.; Johnston, W. A.; and Simons, M. A.: The influence of expertise on X-ray image processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 14, no. 3, 1988, p. 553.

26. Yogo, M.; and Fujihara, S.: Working memory capacity can be improved by expressive writing: A randomized experiment in a Japanese sample. *British Journal of Health Psychology*, vol. 13, no. 1, 2008, pp. 77–80.

27. MacLeod, C. M.; Gopie, N.; Hourihan, K. L.; Neary, K. R.; and Ozubko, J. D.: The production effect: delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 36, no. 3, 2010, p. 671.

28. Bellezza, F. S.: Mnemonic devices: Classification, characteristics, and criteria. *Review of Educational Research*, vol. 51, no. 2, 1981, pp. 247–275.

29. Stanger-Hall, K. F.: Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. *CBE—Life Sciences Education*, vol. 11, no. 3, 2012, pp. 294–306.

30. New, J.; Cosmides, L.; and Tooby, J.: Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences*, vol. 104, no. 42, 2007, pp. 16598–16603.

31. Federal Aviation Administration (FAA): Roadmap for Artificial Intelligence Safety Assurance. , FAA, July 2024. URL https://www.faa.gov/aircraft/air_cert/step/roadmap_for_AI_safety_assurance, version I.

32. Wasson, K. S.; and Voros, R.: Deobfuscating Machine Learning Assurance and Approval. *Proceedings of the Digital Avionics Systems Conference (DASC)*, 2024. To appear.

33. FAA: System Design and Analysis. Advisory Circular AC 25.1309-1A, Federal Aviation Administration, June 1988. URL http://rgl.faa.gov/Regulatory_and_Guidance_Library/rgAdvisoryCircular.nsf/0/50BFE03B65AF9EA3862569D100733174?OpenDocument.

34. FAA: System Safety Analysis and Assessment for Part 23 Airplanes. Advisory Circular AC 23.1309-1E, Federal Aviation Administration, Washington, DC, USA, Nov. 2011. URL http://www.faa.gov/documentLibrary/media/Advisory_Circular/AC%2023.1309-1E.pdf.

35. FAA: System Design and Analysis. Advisory Circular AC 25.1309-1B, Federal Aviation Administration, Aug. 2024. URL https://www.faa.gov/

regulations_policies/advisory_circulars/index.cfm/go/document.information/documentID/1043037.

36. G-10 Aerospace Behavioral Engineering Technology. SAE International, 2024. URL https://standardsworks.sae.org/standards-committees/g-10-aerospace-behavioral-engineering-technology.

37. S-18 Aircraft and Sys Dev and Safety Assessment Committee. SAE International, 2024. URL https://standardsworks.sae.org/standards-committees/s-18-aircraft-sys-dev-safety-assessment-committee.

38. ARP4754B: *Guidelines for Development of Civil Aircraft and Systems*. SAE International, Dec. 2023. URL https://www.sae.org/standards/content/arp4754b/.

39. ARP4761A: *Guidelines for Conducting the Safety Assessment Process on Civil Aircraft. Systems, and Equipment*. SAE International, Dec. 2023. URL https://www.sae.org/standards/content/arp4761a/.

40. RTCA DO-178C: *Software Considerations in Airborne Systems and Equipment Certification*. RTCA, Inc., Washington, DC, USA, Dec. 2011. URL https://my.rtca.org/NC__Product?id=a1B36000001IcmqEAC.

41. S-18A Autonomy Working Group. SAE International, 2024. URL https://standardsworks.sae.org/standards-committees/s-18a-autonomy-working-group.

42. AIR7121: *Challenges in the Application of Development Assurance Systems Safety Practices to New and Emerging Aviation Transportation Technology*. SAE International, 2024. Draft.

43. Leveson, N. G.; and Thomas, J. P.: STPA Handbook. Electronic document, Mar. 2018. URL http://psas.scripts.mit.edu/home/get_file.php?name=STPA_handbook.pdf.

44. Aircraft Accident Investigation Bureau: Ethiopian Airlines Group, B737-8 (MAX) Registered ET-AVJ, 28 NM South East of Addis Ababa, Bole International Airport, March 10, 2019. Aircraft Accident Investigation Preliminary Report AI-01/19, Federal Democratic Republic of Ethiopia, Ministry of Transport, Mar. 2019. URL http://www.ecaa.gov.et/documents/20435/0/Preliminary+Report+B737-800MAX+%2C%28ET-AVJ%29.pdf/4c65422d-5e4f-4689-9c58-d7af1ee17f3e.

45. Komite Nasional Keselamatan Transportasi: PT. Lion Mentari Airlines, Boeing 737-8 (MAX); PK-LQP Tanjung Karawang, West Java, Republic of Indonesia, 29 October 2018. Aircraft Accident Investigation Report KNKT.18.10.35.04, Republic of Indonesia, Jakarta, Indonesia, 2019. URL https://www.flightradar24.com/blog/wp-content/uploads/2019/10/JT610-PK-LQP-Final-Report.pdf.

46. S-18H Human Considerations for Safety Assessment Commiittee. SAE International, 2024. URL https://standardsworks.sae.org/standards-committees/s-18h-human-considerations-safety-assessment-committee.

47. AIR7127: *Human Considerations in Functional Hazard Assessments*. SAE International, 2024. Draft.

48. Milner, R.: A Modal Charactierization of Observable Machine-Behaviour. *Proceedings of the 6th Colloquium on Trees in Algebra and Programming (CAAP)*, E. Astesiano and C. Böhm, eds., Lille, France, vol. 112 of *Lecture Notes in Computer Science (LNCS)*, 1981, pp. 25–34. URL https://dx.doi.org/10.1007/3-540-10828-9_52.

49. Hollnagel, E.: The Resilience Analysis Grid. *Resilience Engineering in Practice: A Guidebook*, E. Hollnagel, J. Pariès, D. D. Woods, and J. Wreathall, eds., Ashgate, Farnham, UK, 2011.

50. Performance-based Operations Aviation Rulemaking Committee (PARC)/Commercial Aviation Safety Team (CAST) Flight Deck Automation Working Group: Operational Use of Flight Path Management Systems. Technical report, Federal Aviation Administration, Sept. 2013. URL http://www.faa.gov/about/office_org/headquarters_offices/avs/offices/afs/afs400/parc/parc_reco/media/2013/130908_PARC_FltDAWG_Final_Report_Recommendations.pdf.

51. Gilchrist, A.: *Seeing Black and White*. Oxford University Press, 2006.

52. Sharit, J.: Human Error and Human Reliability Analysis. *Handbook of Human Factors and Ergonomics*, G. Salvendy, ed., Wiley, fourth ed., Mar. 2012.

53. Reason, J.: *Human Error*. Cambridge University Press, 1990.

54. Holbrook, J.: Exploring Methods to Collect and Analyze Data on Human Contributions to Aviation Safety. *Proceedings of the 21st International Symposium on Aviation Psychology*, 2021, pp. 110–115. URL https://corescholar.libraries.wright.edu/isap_2021/19.

55. Autopilot and Full Self-Driving Capability. Retrieved 13 Aug. 2024. URL https://tesla.com/support/autopilot.

56. Howe, M. L.; Garner, S. R.; and Patel, M.: Positive Consequences of False Memories. *Bheavioral Sciences & the Law*, vol. 31, no. 5, 2013, pp. 652–665. URL https://dx.doi.org/10.1002/bsl.2078.

57. Cosmides, L.; and Tooby, J.: Are Humans Good Intuitive Statisticians After All? Rethinking Some Good Conclusions From the Literature on Judgment Under Uncertainty. *Cognition*, vol. 58, 1996, pp. 1–73. URL https://doi.org/10.1016/0010-0277(95)00664-8.

58. Cosmides, L.; and Tooby, J.: Evolutionary Psychology: New Perspectives on Cognition and Motivation. *Annual Review of Psychology*, vol. 64, Jan. 2013, pp. 201–229. URL https://dx.doi.org/10.1146/annurev.psych.121208.131628.

59. Parasuraman, R.; Sheridan, T. B.; and Wickens, C. D.: A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 30, no. 3, 2000, pp. 286–297. URL https://dx.doi.org/10.1109/3468.844354.

60. European Union Aviation Safety Agency (EASA): Artificial Intelligence Roadmap 2.0: Human-Centric Approach to AI in Aviation. , EASA, May 2023. URL https://easa.europa.eu/ai.

61. European Union Aviation Safety Agency (EASA): Guidance for Level 1 & 2 Machine Learning Applications. Easa concept paper, EASA, Mar. 2024. URL https://easa.europa.eu/ai, issue 2.

62. Waldron, L.; and Medina, B.: When Transgender Travelers Walk Into Scanners, Invasive Searches Sometimes Wait on the Other Side. *ProPublica*, 26 Aug. 2019. URL https://www.propublica.org/article/tsa-transgender-travelers-scanners-invasive-searches-often-wait-on-the-other-side.

63. Guszcza, J.; Lewis, H. H.; and Evans-Greenwood, P.: Cognitive Collaboration: Why Humans and Computers Think Better Together. *Deloitte Review*, 2017, pp. 1–24.

64. Uber: Fast-Forwarding to a Future of On-Demand Urban Air Transportation. 2016. URL https://evtol.news/__media/PDFs/UberElevateWhitePaperOct2016.pdf.

65. McDonnell, N.: The Philosophy of X in XAI. *Proceedings of the ACM IUI Workshops*, Sydney, Australia, Mar. 2023. URL https://ceur-ws.org/Vol-3359/paper21.pdf.

66. Berube, C.: Craptions. *99% Invisible Podcast*, May 2023. URL https://99percentinvisible.org/episode/craptions/.

67. RTCA DO-254: *Design Assurance Guidance for Airborne Electronic Hardware*. RTCA, Inc., Washington, DC, USA, Apr. 2000. URL https://my.rtca.org/NC__Product?id=a1B36000001IcjTEAS.

68. Holloway, C. M.: The Overarching Properities and Overarching Properties Related Arguments. Presentation, NASA, July 2022. URL https://ntrs.nasa.gov/api/citations/20220009308/downloads/OPT01-2022-06-14-1331.pdf.

69. Pritchett, A. R.; Feigh, K. M.; Kim, S. Y.; and Kannan, S. K.: Work Models that Compute to Describe Multiagent Concepts of Operation: Part 1. *Journal of Aerospace Information Systems*, vol. 11, no. 10, 2014, pp. 610–622. URL https://doi.org/10.2514/1.I010146.

70. Davis, J.; Matessa, M.; Rollini, S. F.; Bhattacharyya, S.; Gupta, A.; Narayan, N.; Ganeriwala, P.; Purohit, H.; and Jones, R. M.: Assured Human Machine Interface for Increasingly Autonomous Systems (AHMIIAS). Contractor Report NASA/CR-2023-000000, National Aeronautics and Space Administration (NASA), Mar. 2023. In press.

71. Narayan, N.; Ganeriwala; Jones, R. M.; Matessa, M.; Bhattacharyya, S.; Davis, J.; Purohit, H.; and Rollini, S. F.: Assuring Learning-Enabled Increasingly Autonomous Systems. *International Systems Conference (SysCon)*, Vancouver, BC, Canada, Apr. 2023. URL https://dx.doi.org/10.1109/SysCon53073.2023.10131227.

72. Kopeikin, A.: System-Theoretic Safety Analysis for Teams of Collaborative Controllers. Ph.D. Thesis, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, Sept. 2023.

73. Kopeikin, A.; Leveson, N.; and Neogi, N. A.: System-Theoretic Analysis of Unsafe Collaborative Control in Teaming Systems. *Proceedings of the AIAA SciTech Forum and Exposition*, Orlando, FL, USA, Jan. 2024. URL https://ntrs.nasa.gov/api/citations/20230017753/downloads/Kopeikin_AIAA_UnsafeCollabControl_v5.pdf.

74. Whalen, M. W.; Gacek, A.; Cofer, D.; Murugesan, A.; Heimdahl, M. P. E.; and Rayadurgam, S.: Your "What" Is My "How": Iteration and Hierarchy in System Design. *IEEE Software*, vol. 30, no. 2, 2013, pp. 54–60. URL https://dx.doi.org/10.1109/MS.2012.173.

75. Feiler, P. H.; and Gluch, D. P.: *Model-Based Engineering with AADL: An Introduction to the SAE Architecture Analysis & Design Language*. Addison-Wesley Professional, 2012.

76. Stewart, D.; Liu, J. J.; Cofer, D.; Heimdahl, M.; Whalen, M. W.; and Peterson, M.: AADL-Based Safety Analysis Using Formal Methods Applied to Aircraft Digital Systems. *Reliability Engineering & System Safety*, vol. 213, 2021. URL https://dx.doi.org/10.1016/j.ress.2021.107649.

77. Federal Aviation Administration (FAA): Unmanned Aircraft System Traffic Management (UTM). Web page: https://www.faa.gov/uas/advanced_operations/traffic_management, Last checked 13 Sept. 2024.

78. Warm, J. S.; Dember, W. N.; and Hancock, P. A.: Vigilance and workload in automated systems. *Automation and Human Performance*, CRC Press, 2018, pp. 183–200.

79. Wickens, C. D.; Goh, J.; Helleberg, J.; Horrey, W. J.; and Talleur, D. A.: Attentional Models of Multitask Pilot Performance Using Advanced Display Technology. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 45, no. 3, 2003, pp. 360—380. URL https://doi.org/10.1518/hfes.45.3.360.27250.

80. Endsley, M. R.; and Jones, W. M.: A Model of Inter- and Intra-Team Situation Awareness: Implications for Design, Training and Measurement. *New Trends in Cooperative Activities: Understanding System Dynamics in Complex Environments*, M. McNeese, E. Salas, and M. Endsley, eds., Human Factors and Ergonomics Society, 2001, pp. 46–67. URL https://www.researchgate.net/publication/285745823_A_model_of_inter_and_intra_team_situation_awareness_Implications_for_design_training_and_measurement_New_trends_in_cooperative_activities_Understanding_system_dynamics_in_complex_environments.

81. van der Schaaf, T. W.; and Kanse, L.: Error Recovery in Socio-Technical Systems. *Proceedings of the 7th European Conference on Cognitive Science Approaches to Process Control (CSAPC)*, Villeneuve d'Ascq, France, Sept. 1999.

Acronyms, contractions, and initialisms

AADL	Architectural Analysis and Design Language (AADL)
AAM	Advanced aerial mobility
AFHA	Aircraft functional hazard assessment
AGREE	Assume Guarantee Reasoning Environment
AHMIIAS	Assured Human Machine Interface for Increasingly Autonomous Systems
AI	Artificial intelligence
AMASE	Architectural Modeling and Analysis for Safety Engineering
ASA	Aircraft safety assessment
ATM	Air traffic management
BVLOS	Beyond visual line-of-sight
CAA	Civil Aviation Authority
CFR	Code of Federal Regulations
CRM	Crew resource management
CMA	Common mode analysis
Co-I	Co-investigator
ConOps	Concept of operations
EASA	European Union Aviation Safety Agency
eVTOL	Electric vertical take-off and landing (aircraft)
FAA	Federal Aviation Administration
FDAL	Function development assurance level
FMEA	Failure modes and effects analysis
FTA	Fault tree analysis
GPS	Global Positioning System
HAT	Human–automation teaming
HMT	Human–machine team(int)
IA	Increasingly autonomous
IAS	Increasingly autonomous systems
IMU	Inertial measurement unit
IDAL	Item development assurance level
LIDAR	Light detection and ranging
MBSA	Model-based safety analysis
ML	Machine learning
NAS	National Airspace System
NRA	NASA research announcement
NRC	National Research Council
PASA	Preliminary aircraft safety assessment
PI	Principle investigator

PSSA	Preliminary system safety assessment
RL	Reinforcement learning
SFHA	System functional hazard assessment
STAMP	System-Theoretic Accident Model and Process
STPA	System-Theoretic Process Analysis
SSA	System safety assessment
sUAS	Small uncrewed aircraft system
UAM	Urban air mobility
UAS	Uncrewed aircraft system
V&V	Verification and validation
WMC	Work Models that Compute
ZSA	Zonal safety analysis