Artificial Intelligence as an Aviation Safety Decision-Maker: Adapting Safety Management Systems for Human–AI Operations

Albert N. Clark

Independent Author

Published: August 29, 2026

ASX Research Journal and Database

ISSN 3068-3351 (Online)

Place of Publication: Cadiz City, Philippines

Publisher: ASXResearch.org

Author Note

Albert N. Clark

Department of Aerospace Sciences, ASXResearch.org

ORCID iD: https://orcid.org/0009-0002-7348-4395

The author reports no conflicts of interest.

Correspondence concerning this article should be addressed to Albert N. Clark,

Email: aclark@asxresearch.org

Abstract

Artificial intelligence is rapidly evolving from an analytical support technology into an active participant in aviation safety decision-making, creating challenges that extend beyond traditional approaches to automation and software assurance. This paper examines how Safety Management Systems (SMS) must evolve as artificial intelligence increasingly recommends or executes safety-critical operational decisions within aircraft, air traffic management, maintenance, dispatch, and organizational safety processes. Particular attention is given to automation bias, explainability, human override authority, accountability, AI-generated hazards, operational design domains, and the continuous assurance of machine-learning systems after deployment. The analysis traces the historical progression from conventional automation to human–AI teaming and examines contemporary developments involving the Federal Aviation Administration, European Union Aviation Safety Agency, Airbus, the United States military, and DARPA. Legal and civil implications are also considered, particularly the difficulty of assigning responsibility when safety decisions emerge from interactions among humans, operators, manufacturers, software developers, training data, and artificial intelligence. The paper argues that future SMS frameworks must treat AI simultaneously as a safety tool, decision-making participant, potential hazard source, and safety control requiring continuous operational monitoring. Ultimately, successful integration will depend not upon eliminating human authority or preventing artificial intelligence from making consequential decisions, but upon creating human–AI systems in which authority is explicit, decisions are explainable and reconstructable, operational boundaries are enforceable, failures remain detectable, and accountability is preserved throughout the aviation safety system.

Keywords: artificial intelligence, Safety Management Systems, human–AI teaming

Certifying the Uncertifiable: Aviation Safety Regulation for Learning Artificial Intelligence

Figure 1

Certifying the Uncertifiable: Aviation Safety Regulation for Learning Artificial Intelligence

Note. Click image for full-size image or click here.

Aviation did not arrive at artificial intelligence by suddenly deciding that computers should make safety decisions. It arrived there through a century-long transfer of increasingly sophisticated tasks from humans to machines: stabilization, navigation, flight guidance, engine control, collision avoidance, envelope protection, predictive maintenance, and eventually the interpretation of enormous streams of operational data. The intellectual warning accompanied that progress almost from the beginning. Bainbridge (1983) observed that automation can create an irony: the more successfully a system removes humans from routine control, the more

difficult the remaining human role can become when the automation encounters something it cannot manage. Parasuraman et al. (2000) later provided a particularly useful framework for understanding the transition now occurring in aviation, distinguishing automation of information acquisition and analysis from automation of decision selection and action implementation. Artificial intelligence crosses an important boundary when it moves from telling a pilot, dispatcher, controller, mechanic, or safety manager what the data show to telling that person what should be done—or doing it itself. That distinction is central to the future of Safety Management Systems (SMS), because SMS was constructed around organizations in which hazards may originate in technology but consequential safety judgments ultimately belong to identifiable human beings.

The historical human-factors literature makes clear that this is not merely a software problem. Endsley and Kiris (1995) demonstrated the "out-of-the-loop" performance problem: automation can reduce an operator's situation awareness and degrade the ability to intervene effectively after an automated system fails. Aviation adds another danger because the human supervising AI may be confronted not with an obvious mechanical failure but with a perfectly functioning algorithm producing a confidently wrong recommendation. Automation bias can therefore become more dangerous as automation becomes more intelligent. A pilot who challenges an erroneous airspeed indication understands that an instrument can fail; challenging an AI that has simultaneously analyzed weather, traffic, aircraft state, historical outcomes, terrain, and thousands of other variables creates a different psychological problem. The human may reasonably assume that the machine knows something the human does not. Grindley et al. (2026), studying operators of uncrewed air vehicles, reinforce the importance of appropriately calibrated trust: too little trust encourages unnecessary intervention, while too much can

encourage reliance when intervention is required. The safety problem is consequently not maximizing trust in AI. It is producing the correct amount of trust at precisely the correct moment.

Modern SMS is built around four familiar pillars—safety policy and objectives, safety risk management, safety assurance, and safety promotion—but AI pressures every one of them. Traditional safety risk management asks organizations to identify hazards, evaluate risk, implement controls, and monitor whether those controls work. An AI decision-maker creates an additional category: the safety control itself may generate hazards. Demir et al. (2024), in their systematic and bibliometric examination of 224 aviation-safety studies, found AI increasingly embedded in accident analysis, pilot behavior, safety assessment, optimization, and predictive applications. Yiu et al. (2026), reviewing 175 studies involving AI and large language models in aviation safety, similarly found rapidly expanding applications while identifying trustworthy, certifiable AI and human–AI teaming as fundamental issues. An airline of the near future could therefore possess an AI system that correctly identifies ninety-nine hazards humans miss and simultaneously introduces a hundredth hazard that humans would never have created. SMS must evolve from asking only "What hazards does our operation contain?" to asking "What hazards can our safety intelligence create, amplify, conceal, or normalize?" That means AI model behavior, training data, distribution shifts, false-positive and false-negative patterns, anomalous recommendations, rejected recommendations, overrides, and post-deployment performance must themselves become safety data.

The regulatory community is already moving toward this problem, although certification and SMS should not be confused. The FAA's AI Safety Assurance Roadmap explicitly separates the "safety of AI" from the "use of AI for safety," an extraordinarily important distinction: AI can

become both an object of safety assurance and an instrument used to produce safety. EASA has gone further in describing Level 2 AI as human–AI teaming in which AI systems may automatically take decisions under human oversight, while emphasizing learning assurance, explainability, ethics, and human–AI interaction. Pérez-Castán et al. (2022) tested EASA's W-shaped learning-assurance methodology using an ML-based air-traffic conflict predictor and found that performance could vary with the time at which a prediction was made—an elegant demonstration of why conventional aggregate accuracy metrics can be inadequate for safety-critical AI. Luettig et al. (2024) likewise found important gaps between conventional aerospace certification standards and machine-learning systems, particularly in requirements specification, uncertainty, generalization, verification, and unintended behavior. The lesson for SMS is profound: "the model passed validation" cannot become the AI equivalent of "the component passed inspection." Safety assurance must ask whether the model remains trustworthy in the operational environment in which it is actually making decisions.

Explainability becomes critical precisely because a safety decision without an intelligible rationale can be operationally useless even when statistically impressive. Vidot et al. (2024), whose authors include researchers affiliated with Airbus, identify robustness, provability, and explainability among the central challenges in qualifying machine-learning-based avionics. An explanation, however, must do more than satisfy an engineer after an event. In operational SMS, it has to reach the right human at the right level of abstraction and at the right time. "Reject takeoff" is an instruction; "reject takeoff because predicted compressor instability exceeds the validated threshold" is a reason; displaying the internal activation pattern of a neural network is neither, from a pilot's perspective. Human override authority therefore cannot merely mean installing an OFF button. The operator must possess sufficient information, time, authority,

training, and confidence to contradict the AI. Organizations must also protect people who exercise that authority in good faith. If every human override is subsequently treated as an error whenever the AI proves correct, crews will quickly learn that override authority exists on paper but is professionally dangerous in practice. A mature human–AI SMS should therefore treat overrides, disagreements, and near-misses as valuable safety observations rather than automatic evidence that either the person or the machine failed.

A specific manufacturer is tackling this frontier aggressively: Airbus. The company publicly identifies decision-making and autonomous flight among its principal AI domains and states that computer vision and machine learning are important to self-piloted aircraft capable of autonomous takeoff, landing, navigation, and ground-obstacle detection. In June 2026 Airbus described embedded AI applications for trajectory management, navigation, surveillance, crew decision support, obstacle detection, and vision-based automatic landing, while stressing that aerospace AI requires industrial assurance radically different from consumer cloud AI. Its autonomous-flight portfolio includes work inherited from Acubed's Wayfinder program, extended Minimum Crew Operations, and Optimate; Airbus says the former Acubed work developed technologies and processes aimed at highly reliable, certifiable autonomy and machine-learning solutions. The significance is not that Airbus is about to remove pilots from airliners. It is that one of the world's largest transport-aircraft manufacturers is actively building the technological bridge from automation that executes predetermined logic toward systems that perceive complex environments and provide increasingly consequential decision support. SMS must accompany that transition before authority migrates, not after an accident reveals that it already did.

The military is moving considerably faster because its risk calculus, mission requirements, test authorities, and certification environment differ from civil transport. DARPA's Air Combat Evolution (ACE) program used the X-62A VISTA, a modified F-16, to demonstrate AI-controlled flight and ultimately autonomous within-visual-range combat against a human-piloted F-16. The program was explicitly designed not only around autonomous performance but around measuring and calibrating human trust in combat autonomy. That work has now advanced beyond a unique research aircraft. In July 2026 DARPA and the U.S. Air Force announced flights of VENOM-modified F-16s using an autonomy kit capable of integrating AI while leaving the aircraft's core software unchanged. DARPA's Artificial Intelligence Reinforcements program is extending the problem into multi-ship, beyond-visual-range operations in uncertain environments, including adaptive autonomy, predictive models, integrated sensors, and expert human feedback. The military question is therefore rapidly becoming the same question civil SMS will eventually confront: when a human and an AI disagree under time pressure, who is commanding whom? DARPA's earlier Explainable AI program recognized that autonomous systems that perceive, learn, decide, and act require humans who can appropriately understand and manage their artificial partners.

That question exposes the central accountability problem. SMS depends upon responsibility being assignable. A chief pilot, director of safety, maintenance organization, dispatcher, manufacturer, controller, or certificate holder occupies a recognizable place in an accountability structure. An algorithm does not. Fadhil et al. (2025), examining explainable AI and legal accountability in UAV fault diagnosis, identify precisely this difficulty: AI failures can distribute potential responsibility among operators, manufacturers, software developers, and other actors, while opaque decision processes make attribution harder. If an AI recommends

continuation into deteriorating weather, the captain accepts the recommendation, and an accident follows, investigators will need to ask whether the failure was operational judgment, deficient training, a model defect, inappropriate training data, a poorly specified operational design domain, inadequate interface design, organizational pressure to follow AI recommendations, or some combination. Explainability therefore has legal as well as engineering significance. Accident investigators, regulators, insurers, courts, and litigants may need to reconstruct not merely what the AI output was, but what information it received, which model version produced the output, how confident the system was, what alternatives it considered, what explanation the human received, and whether the human had meaningful authority to reject it.

Civil ramifications extend beyond conventional product liability. AI decision-making could change the evidentiary architecture of an accident investigation. Today investigators recover flight data, cockpit voice recordings, maintenance records, dispatch information, weather, ATC communications, and organizational evidence. A mature AI-enabled aircraft or airline may require another kind of recorder: an auditable decision history preserving model identity, software and dataset configuration, inputs, outputs, confidence or uncertainty measures where meaningful, safety-envelope interventions, human acknowledgments, rejected recommendations, overrides, and subsequent system responses. Such records raise difficult questions of proprietary algorithms, cybersecurity, privacy, discoverability, international data transfer, and intellectual property. Yet allowing a safety-critical AI to make consequential decisions without preserving sufficient evidence to reconstruct those decisions would create an unacceptable investigative blind spot. SMS therefore needs an AI equivalent of configuration control and flight-data preservation. Werner et al. (2026) place operational domain and AI/ML operational design domain definition at the center of learning assurance, demonstrating how

precisely the permitted operating environment must be described. That logic belongs in SMS as well: when AI encounters circumstances outside its validated knowledge boundary, uncertainty itself should become a reportable safety condition rather than an invitation for the algorithm to improvise.

Elon Musk is relevant to the broader technological conversation but, as of August 29, 2026, there is no credible evidence that Musk, Tesla, xAI, or SpaceX is leading a program to certify AI as a safety-critical decision-maker aboard civil transport aircraft. That distinction matters. Tesla's experience with neural-network-driven vehicle automation makes Musk relevant to debates about autonomy, supervision, and human reliance, while SpaceX unquestionably operates highly automated aerospace systems. But automation, artificial intelligence, machine learning, and adaptive safety-critical decision authority are not interchangeable terms. The current aviation work that can be documented directly is being driven by regulators such as FAA and EASA, manufacturers including Airbus, specialist developers and research institutions, and defense organizations including DARPA and the Air Force. Airbus, for example, now explicitly describes future high-compute aircraft architectures capable of running onboard AI applications that support pilots and shift some human activity from tactical execution toward strategic management. Attaching Musk's name to civil aviation AI certification without evidence would make the subject more sensational and the research less accurate. He belongs in the discussion as an adjacent technological influence, not as a participant whose involvement can presently be demonstrated.

The greatest operational danger may ultimately be neither an AI that fails dramatically nor one that becomes uncontrollably intelligent, but an AI that is correct often enough that humans stop asking whether it is correct this time. That is where automation bias becomes an

SMS problem. Mosier et al. (1998) distinguished omission errors, in which people fail to act because automation fails to identify a problem, from commission errors, in which people follow an automated recommendation despite contradictory information. AI can intensify both. An AI may suppress a hazard by failing to flag it, or create an apparent consensus by presenting a recommendation with enough authority that a crew discounts conflicting sensory evidence. The traditional SMS concept of safety culture therefore needs an additional dimension: disagreement culture. Pilots, controllers, mechanics, dispatchers, and safety analysts must be trained not merely to operate AI but to challenge it intelligently. Recurrent training should include deliberately erroneous AI recommendations, uncertainty, conflicting evidence, degraded sensors, adversarial inputs, and circumstances outside the operational design domain. A human who is designated the final authority but has never practiced overruling an apparently superior machine is not meaningful redundancy. Human oversight without practiced human independence is theater.

The future SMS architecture should consequently treat AI as something aviation has rarely encountered before: simultaneously a tool, a teammate, a hazard source, a safety control, and eventually perhaps an operational decision authority. Safety policy must define exactly which decisions AI may recommend, which it may execute, and which remain inherently human. Safety risk management must assess AI-generated hazards and human–AI interaction failures rather than considering only component failure. Safety assurance must continuously monitor operational performance, distribution shifts, anomalous outputs, override patterns, model changes, and whether safety assumptions remain valid. Safety promotion must teach personnel how the AI behaves, where its competence ends, how uncertainty is communicated, and when disagreement is expected. The objective should not be to keep a human nominally “in the loop”

regardless of whether that human contributes anything. It should be to design an accountable human–AI team in which authority follows demonstrated competence while immutable safety boundaries prevent either member from silently moving outside the approved system. The certification research reviewed by Luettig et al. (2024), Vidot et al. (2024), and Yiu et al. (2026) points toward exactly this larger assurance problem: safety cannot rest on raw AI performance alone. It must emerge from the engineered system surrounding the intelligence.

Aviation has spent more than a century learning that safety does not come from eliminating human error; it comes from designing systems capable of surviving it. Artificial intelligence changes the noun, not the principle. The next generation of SMS must be capable of surviving machine error as deliberately as today's system attempts to survive human error, while recognizing that the two will increasingly interact. The most dangerous future accident may be impossible to classify neatly as "pilot error" or "automation failure" because the causal chain will run through a human who misunderstood an AI, an AI that misunderstood the environment, an organization that misunderstood both, and a regulatory structure that assumed somebody was still unquestionably in command. The solution is not to fear artificial intelligence or to insist nostalgically that humans must make every important decision forever. It is to preserve what aviation safety has learned at enormous cost: authority must be explicit, hazards must be observable, decisions must be reconstructable, failures must be reportable, assumptions must be challengeable, and responsibility cannot disappear merely because the decision originated inside a neural network. AI can become an extraordinary aviation safety decision-maker. But the measure of success will not be the day an aircraft can make a safety decision without a human. It will be the day aviation can explain exactly why that decision was made, prove the boundaries

within which it was safe, intervene when it was not, learn from the outcome, and still answer the oldest question in safety management: who was responsible?

References

Bainbridge, L. (1983). *Ironies of automation*. Automatica, 19(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Demir, G., Moslem, S., & Duleba, S. (2024). *Artificial intelligence in aviation safety: Systematic review and biometric analysis*. International Journal of Computational Intelligence Systems, 17, 279. https://doi.org/10.1007/s44196-024-00671-w

Endsley, M. R., & Kiris, E. O. (1995). *The out-of-the-loop performance problem and level of control in automation.* Human Factors, 37(2), 381–394. https://doi.org/10.1518/001872095779064555

Fadhil, T. H., Al-Haddad, L. A., & Al-Karkhi, M. I. (2025). *Legal accountability and UAV fault diagnosis explainable AI in aviation safety and regulatory compliance for liability challenges.* Discover Artificial Intelligence, 5, 410. https://doi.org/10.1007/s44163-025-00690-2

Grindley, B., Cherrett, T., Scanlan, J., & Plant, K. L. (2026). *Exploring the relationship between operator experience, propensity to trust automation and perceived system trustworthiness of uncrewed air vehicles*. Ergonomics. Advance online publication. https://doi.org/10.1080/00140139.2026.2625177

Luettig, B., Akhiat, Y., & Daw, Z. (2024). *ML meets aerospace: Challenges of certifying airborne AI*. Frontiers in Aerospace Engineering, 3, 1475139. https://doi.org/10.3389/fpace.2024.1475139

Mosier, K. L., Dunbar, M., McDonnell, L., Skitka, L. J., Burdick, M., & Rosenblatt, B. (1998). *Automation bias and errors: Are teams better than individuals?*

Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 42(3), 201–205. https://doi.org/10.1177/154193129804200304

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). *A model for types and levels of human interaction with automation*. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 30(3), 286–297. https://doi.org/10.1109/3468.844354

Pérez-Castán, J. A., Pérez Sanz, L., Fernández-Castellano, M., Radišić, T., Samardžić, K., & Tukarić, I. (2022). Learning assurance analysis for further certification process of machine learning techniques: Case-study air traffic conflict detection predictor. Sensors, 22(19), 7680. https://doi.org/10.3390/s22197680

Vidot, G., Gabreau, C., Ober, I., & Ober, I. (2024). Qualification of avionic software based on machine learning: Challenges and key enabling domains. Journal of Aerospace Information Systems, 21. https://doi.org/10.2514/1.I011164

Werner, F., Christensen, J. M., Stefani, T., Köster, F., Hoemann, E., & Hallerbach, S. (2026). *Formulating a learning assurance-based framework for AI-based systems in aviation*. Aerospace, 13(2), 200. https://doi.org/10.3390/aerospace13020200

Yiu, C. Y., Li, W. C., Ng, K. K. H., Chi, C. F., & Schiefele, J. (2026). *Enhancing aviation safety with artificial intelligence: A systematic literature review on recent advances, challenges and future perspectives.* Advanced Engineering Informatics, 71, 104378. https://doi.org/10.1016/j.aei.2026.104378