Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 1095–1104
https://doi.org/10.54941/ahfe1006922

Implementation of Artificial Intelligence (AI) in Transport Accident Investigations

Dimitrios Ziakkas[1], Debra Henneberry[1], and Ioanna Lekea[2]

[1]Purdue University, School of Aviation and Transportation Technology, West Lafayette, IN 47907, USA
[2]Hellenic Air Force Academy, Department of Aeronautical Sciences, Division of Leadership-Command, Humanities and Physiology, Dekeleia Air Base - Attiki, Athens - 13671 (1010) – Greece

ABSTRACT

Transport accident investigations are crucial for understanding causal factors, improving system safety, and preventing future incidents. Traditionally, these investigations rely on a multidisciplinary process involving human expertise, manual data analysis, and narrative reconstruction. However, with the growing complexity of transportation systems and the increasing volume of operational data—from flight data recorders, cockpit voice recordings, sensor logs, to surveillance systems—the limitations of manual analysis are becoming evident. This paper explores the emerging role and potential of Artificial Intelligence (AI) in augmenting and transforming transport accident investigations across aviation, maritime, rail, and roadway domains. AI technologies such as machine learning, natural language processing (NLP), and computer vision are proving to be powerful tools in extracting patterns, identifying anomalies, and drawing correlations from large datasets that are otherwise time-consuming and error-prone for human analysts. This paper examines several case studies and research projects where AI-assisted tools have been piloted or implemented in post-accident analysis. These include automated speech recognition for cockpit voice recordings, anomaly detection in flight trajectories, and sentiment analysis of maintenance logs. Findings indicate that AI can significantly reduce investigation time frames, increase objectivity in evidence evaluation, and uncover hidden contributing factors—particularly in cases involving complex system interactions or human-machine interface failures. Despite its promise, the implementation of AI in accident investigations is not without challenges. One critical concern is transparency and explainability. Unlike traditional analytical methods, AI models—especially deep learning systems—can function as "black boxes," making it difficult for investigators, regulators, and courts to interpret how a conclusion was reached. This raises questions about the admissibility of AI-generated evidence and its alignment with legal and ethical standards in safety investigations. The paper emphasizes the need for human-in-the-loop approaches where AI augments, rather than replaces, expert judgment. Human oversight remains essential in contextual interpretation, ethical reasoning, and final decision-making. Furthermore, the integration of AI into accident investigation agencies requires cultural and organizational shifts. Investigators need training not only in technical AI tools but also in data literacy, interdisciplinary collaboration, and understanding the biases that AI models may inherit from their training data. This paper proposes a roadmap for implementation, including phased adoption, validation protocols, inter-agency cooperation, and regulatory support. In conclusion, AI has the potential to revolutionize transport accident investigations (Ziakkas & Plioutsias, 2024) by enhancing speed, depth, and predictive capability. However, its integration must be guided by principles of transparency, accountability, and collaboration between technologists and human factors experts. As transportation systems evolve toward greater automation and data dependence, leveraging AI in accident investigations is not only beneficial but essential for ensuring the continued integrity and learning capacity of safety-critical systems.

Keywords: Artificial intelligence, Accident investigation, Human-in-the-loop, Transportation safety, Explainable AI, Predictive analytics

© 2025. Published by AHFE Open Access. All rights reserved.

INTRODUCTION

Accident investigation has always been a challenge in disciplined research. The investigator's work—assembling fragments of evidence into a coherent narrative of how a system drifted into failure—depends on technical acumen and human insight in equal measure. For aviation, Annex-style principles codified this stance: independence from blame-seeking, rigorous data preservation, structured reporting, and recommendations aimed at systemic improvement rather than retribution. While modalities differ, the same ethos underwrites inquiries at sea, on rail, and on roads. Today, however, investigators confront an epistemic challenge: the data they must honor has multiplied in kind and in quantity, and the subtlety of human-automation interaction has outpaced paper-era methods. A modern event may involve highly automated control systems, opaque software states, conflicting sensor interpretations, and multinational teams whose communication practices are shaped by culture. The result is a data landscape that is both richer and more refractory to manual analysis.

Longstanding human factors frameworks provide a compass. Reason's organizational accident theory (Reason, 1997) and the Swiss cheese (Reason, 1997) metaphor remind us that failure is rarely the property of a single act; it is the alignment of latent conditions and active breakdowns. Human Factors Analysis and Classification System (HFACS) added hierarchical structure - unsafe acts, preconditions, supervision, organizational influences - allowing investigators to code how system and interpersonal factors link (Wiegmann & Shappell, 2003). Line Operations Safety Audit (LOSA), as a non-jeopardy observational method, taught us to see threats, errors, and undesired states in normal line operations (Kanki, Anca, & Helmreich, 2010). Crew Resource Management (CRM) (Kanki, Anca, & Helmreich, 2010; Helmreich, Merritt, & Wilhelm, 1999) and, more recently, Competency-Based Training & Assessment (CBTA) (ICAO, 2013)/Evidence-Based Training (EBT) moved training and assessment toward observable behavior under real constraints. These frameworks are not obsolete in an AI era; they are the grammar through which algorithmic signals can become operational meaning.

At the same time, cultural dynamics in multinational operations complicate what is "in the data." Speech acts captured on cockpit voice recordings are not mere words; they carry pragmatics—directness, hedging, facework (Gudykunst, 2002)—that vary across cultural communities. Silence can be deference rather than agreement; formal politeness may mark dissent as much as assent. Without cultural fluency, algorithmic summaries of audio transcripts or sentiment signals from narratives risk misclassification. Cultural Intelligence (CQ) (Ang & Van Dyne, 2008), introduced to aviation as a teachable competence, provides a lens for interpreting communication and authority gradients (Helmreich & Merritt, 1998; Altemeyer, 1996) without stereotyping. It is therefore as relevant to AI assisted investigations as it is to training: outputs must be read with sensitivity to the people who produced them.

The case for AI in investigations is pragmatic. First, time is a critical controlling factor. Public accountability, regulatory deadlines, and the opportunity to prevent recurrence hinge on reducing time-to-insight without sacrificing depth. Second, the kinds of structure investigators seek—temporal sequences, outliers, recurring patterns of interaction—are precisely the sorts of patterns statistical learning can surface rapidly when data are well curated. Third, the fidelity and diversity of modern data (high-rate FDR parameters, multi-channel audio, surveillance video, digital maintenance logs) suit multi-modal models that can synchronize evidence across streams. Yet the promise is bounded by legitimate concerns: transparency, replicability, chain-of-custody, and admissibility. If an algorithm cannot show its working, it cannot earn the trust of investigators, regulators, or courts. The way forward, then, is not to mechanize judgment but to instrument judgment—to place defensible AI components inside an investigative workflow that remains human-led and just-culture aligned.

This paper proceeds in that spirit. It offers a methodology for integrating AI into investigation without abandoning the conceptual rigor of established frameworks. It then reports practice-grounded findings on where AI delivers, where it distorts, and how to design guardrails. Finally, it sets out a roadmap for organizational adoption that emphasizes investigator competence, cross-agency collaboration, and cultural literacy as conditions for success. The argument is intentionally cross-modal: while examples lean on aviation, the methodological seams are shared across maritime, rail, and roadway domains.

METHODOLOGY

The study adopts an interpretivist, translational methodology: the goal is not to build one monolithic model but to specify how AI can be responsibly inserted into the investigator's craft. The approach braided three strands—framework alignment, AI task mapping, and governance design—each iterated with reference to established safety methods and regulatory expectations (Table 1).

Table 1: Research methodology overview.

Methodology	Purpose / Techniques	Key Findings	Implications for Practice
Framework alignment	Map AI outputs to trusted HF frameworks (Reason/Swiss-cheese; HFACS; LOSA; CRM/CBTA); identify analysis junctures where data volume/complexity slows human reasoning.	Faster convergence on shared mental models under time pressure; earlier visibility of threat–error patterns aligned to existing categories.	Use explicit paraphrase/summary markers in CBTA; brief to surface cultural expectations; use AI outputs as structured evidence linked to HFACS/LOSA codes.

Continued

Table 1: Continued

Methodology	Purpose / Techniques	Key Findings	Implications for Practice
AI task mapping	ASR + speaker diarization for CVR; temporal alignment with FDR; unsupervised anomaly detection for trajectories/-parameters; NLP on logs/statements; computer vision for video; multimodal fusion.	Reduced time-to-insight; earlier recognition of automation surprise & workload spikes; contradictions surfaced across narratives; video indexing clarified task/sequence disputes.	Treat model outputs as triage and hypothesis generators; maintain human review at all decision points.
Governance design	Chain-of-custody artifacts (versioned code, data hashes, configs, inference logs); explainability (attention/salien-cy/exemplars); qualified human reviewer gates.	Increased trust, reproducibility, and legal defensibility; smoother regulator/court interactions.	Require model cards and replayable analyses; make audit artifacts part of the evidence file.
CQ guardrails for NLP/audio	CQ-informed prompts; validate assertive-ness/indirectness metrics against cultural context; bias checks on prosody and stance models.	Reduced misinterpretation and bias; fewer authority-gradient failures without flattening leadership.	Add CBTA markers that assess inviting dissent and culturally adaptive debriefing; train reviewers on CQ for transcript interpretation.
Knowledge-graph fusion	Entity/event graph with timestamps, typed edges, and confidence; supports narrative rehearsal and Bayesian updating.	Shared situational awareness; transparent enumeration of evidence supporting/-contradicting hypotheses.	Use graphs to organize competing narratives and maintain provenance with click-through to sources.
Organizational upskilling	CBTA-based data-literacy curriculum; just-culture protections; competence assessment via realistic cases.	Higher adoption; reduced resistance to AI; improved investigator ability to interrogate models.	Define observable competencies (read model cards, interpret confidence, spot overfitting); assess transfer on live files.

Continued

Table 1: Continued

Methodology	Purpose / Techniques	Key Findings	Implications for Practice
Inter-agency collaboration	Shared model repositories; anonymized benchmarks; common validation protocols; regulator guidance.	Improved generalization and reduced duplication; clearer admissibility thresholds for AI-assisted evidence.	Endorse documentation standards; fund cross-modal pilots; include sociolinguists/cultural-psychology in debiasing work.

Framework alignment began with a close reading of mature investigative grammars—Reason's organizational view, HFACS' layered taxonomy, LOSA's threat–error logic, and CRM/CBTA's behaviorally anchored assessment. For each, we identified analysis junctures where data volume or complexity typically slows human reasoning: spanning, for example, the transcription and diarization of cockpit audio, the sifting of high-rate parameter streams for salient anomalies, and the aggregation of unstructured narratives across witness statements and maintenance records. These pressure points became candidates for AI assistance precisely because they are repetitive, pattern-centric, and auditable. The alignment step ensured that any algorithmic output could "attach" meaningfully to categories investigators already trust (e.g., HFACS preconditions; LOSA threat management; CRM communication markers).

AI task mapping then articulated specific model classes to investigative tasks. Automated speech recognition and speaker diarization were mapped to CVR/bridge audio, with domain-tuned language models trained on aviation/maritime lexicons to reduce out-of-vocabulary error; temporal alignment reconciled transcripts with flight data recorders (FDR) and surveillance timelines. For trajectories and high-rate parameters, unsupervised anomaly detection (e.g., clustering, density estimation) and sequence models supported the identification of outlying regimes preceding undesired states. Natural language processing (NLP) pipelines (tokenization, topic modeling, contradiction detection) were tied to maintenance logs and narrative statements to surface recurrent themes, inconsistencies, and sentiment shifts. Where video existed, computer vision pipelines performed detection and re-identification to reconstruct ground movements. A multi-modal fusion step—implemented conceptually as a knowledge graph—bound entities (aircraft, systems, actors, messages) and events (mode transitions, messages, alerts) with timestamps and confidence, providing a substrate for causal reasoning, whether qualitative (narrative synthesis) or quantitative (Bayesian updating).

Governance design addressed legitimacy: chain-of-custody, reproducibility, explainability, and human-in-the-loop controls. Every model stage was paired with an audit artifact (versioned code, data hashes, configuration manifests, and inference logs) so that third parties could replay analyses. Explainability depended on choosing model forms that admitted reasons: attention maps over audio for contested utterances, saliency over parameters for flagged anomalies, exemplar retrieval for NLP classifications.

Critically, every AI output was routed to a qualified reviewer—an investigator trained in both the operational domain and data literacy—who could accept, reject, or annotate the claim. This governance was anchored in just-culture and CBTA principles: the aim is learning, not blame, and competence must be observable and assessed. Furthermore, the method was stress-tested against cross-cultural communication scenarios to ensure that AI did not smuggle cultural bias into evidentiary narratives. For example, assertiveness metrics from transcripts were validated against CQ-informed interpretations so that indirect speech or polite mitigation strategies were not misread as absence of dissent. Investigators were supplied with prompts and checklists derived from CQ training materials to contextualize transcript snippets before drawing inferences about crew dynamics. In effect, CQ served as the interpretive guardrail for NLP-assisted audio analysis (Ziakkas et al., 2024).

The methodology thus does not claim novelty at the level of algorithms; it claims fitness for investigative purpose. It sets out where AI's pattern skills match the structure of the work, and how to keep the human investigator sovereign over meaning, ethics, and recommendations.

FINDINGS

The first finding is expected: time-to-insight can be materially reduced without diluting rigor when AI is positioned as an evidence triage and structuring companion. In several investigations, automated speech recognition with speaker diarization decreased the latency between raw audio and usable transcripts from weeks to days. More importantly, diarization exposed conversational turn-taking and interruptions that manual reviewers routinely under-code when fatigued. Coupled with temporal alignment to FDR events, investigators could see when an alert tone masked a soft-spoken challenge or when a checklist invocation overlapped with ATC transmissions. When these signals were read through CRM and CBTA lenses, evaluators could distinguish between absence of challenge and challenge unheard, a distinction with different training and design implications.

Second, unsupervised anomaly detection over trajectories and parameter streams proved especially valuable in complex events where mode transitions and pilot intent were in tension. Clustering revealed "families" of approaches or climbs that departed from a fleet's typical dispersion; density estimation highlighted sequences rarely observed in benign operations. Investigators used these patterns as hypothesis generators, not conclusions: the algorithm said, "look here," and human analysts asked, "why now?" In several cases, this led to earlier recognition of automation surprises and of workload spikes where standard calls compressed and cross-monitoring thinned. These are the very seams HFACS would later code as preconditions for unsafe acts; AI simply made them easier to see early (Ziakkas et al., 2024).

Third, NLP over maintenance logs and narrative statements uncovered latent contradictions at scale. Topic models clustered repeated complaints about intermittent sensors that individual case readers had treated as noise. Contradiction detection flagged witness statements that diverged sharply on timing or content, prompting targeted follow-up.

While sentiment analysis is often caricatured, in safety narratives its value lies less in emotion detection than in stance mapping: who minimizes, who problematizes, and how these stances correlate with organizational role. When paired with organizational-influence layers in HFACS, such patterns sharpened recommendations from the vague ("improve reporting culture") to the actionable ("revise acceptance test criteria; add fatigue-aware sign-off windows").

Fourth, computer vision contributed meaningfully where surveillance footage existed—on ramps, crossings, or bridge wings. Object detection and re-identification reconstructed vehicle and person flows; pose estimation suggested task allocation and potential procedural deviations. In a rail yard incident, video-derived tracks clarified how an assumed hand signal never occurred, resolving a dispute where two memory-based testimonies were irreconcilable. The evidence was not "what the model said," but the underlying frames the model helped index. That distinction proved critical for legal defensibility: the machine accelerated retrieval; the human interpreted.

Fifth, a knowledge-graph fusion layer—even as a conceptual data structure—helped investigators maintain coherence as they integrated multi-modal outputs. By treating actors, systems, and events as nodes connected by typed, time-stamped edges with confidence scores, teams could rehearse competing narratives without losing data provenance. Such graphs integrated naturally with Bayesian reasoning for updating belief in hypotheses as new evidence accrued. The lived benefit was conversational: teams could ask, "What supports the 'automation surprise' hypothesis?" and the system could enumerate audio, parameter, and narrative links, each with a click-through to source. This did not declare causation; it scaffolded shared situational awareness among investigators.

These benefits hinged on explainability and reproducibility. Where models were opaque or their outputs could not be re-run with frozen data and code, investigators' trust waned—and rightly so. Attention visualizations over transcripts, saliency maps over parameters, and exemplar retrieval for text classifications were the minimal currency of credibility. Just as importantly, the audit trail—hashes of inputs, model versions, configuration files, and action logs—was essential to chain-of-custody and to withstand judicial scrutiny. Teams that treated these artifacts as part of the evidence file found downstream interactions with regulators and courts markedly smoother.

The principal risks clustered around misinterpretation, bias, and organizational culture. NLP systems trained on general English misread mitigated speech from high context cultures as indecisive or "negative." Automated prosody-based stress inferences drifted across accents. Image models struggled in adverse weather or with occlusions; trajectory clusters sometimes reified rare but benign practices as suspicious. These were not reasons to abandon AI; they were reasons to buffer it with CQ-aware human review and to improve domain tuning. Culturally intelligent investigators were better at catching algorithmic category errors because they recognized that directness and dissent wear different clothes across communities. Embedding CQ prompts in the transcript review interface—"Could this be deference rather than agreement?"—reduced over-confident misreadings.

Finally, the organizational challenge was less technical than pedagogical. Investigators needed data literacy—not to become data scientists, but to interrogate models with the same skepticism they apply to human testimony. CBTA principles provided a ready scaffold for upskilling: define observable competencies (e.g., ability to interpret confidence intervals; to spot overfitting symptoms; to demand and read model cards); teach against realistic cases; assess transfer on live files. Where agencies coupled training with just-culture commitments, adoption prospered; where AI was perceived as a surveillance tool to police investigator performance, resistance hardened. As with CRM, legitimacy flowed from a clear line of sight to safer outcomes and from leaders who modeled curiosity over certainty.

To summarize, AI did not change what an investigation is. It changed how quickly and consistently certain kinds of structure become visible, and how disciplined teams could test narratives against richer, synchronized evidence. The craft remained human: weighing plausibility, interpreting culture, and issuing recommendations with moral clarity.

CONCLUSION

The question is no longer whether AI has a role in transport accident investigations; it is how to give it a legitimate one. The answer resides in design choices that respect the moral architecture of investigation—independence, transparency, fairness—and the epistemic humility of complex systems work. The most important choice is to keep humans in the loop and in charge, using AI to instrument rather than substitute judgment. That stance honors the history of human-factors inquiry, from Reason's systemic lens to HFACS' layered coding and LOSA's observational discipline: we learn reliably when we can see patterns, challenge them, and connect them to defensible categories that guide prevention.

A practical roadmap begins with phased adoption. Agencies need not field a monolith; they can start where returns are immediate and risk is low: audio transcription and diarization tuned to operational lexicons; trajectory clustering for exploratory analysis; contradiction detection over narratives. Each insertion should be paired with validation protocols, model cards, and audit artifacts from day one. As confidence grows, more ambitious steps—multi-modal fusion, knowledge graphs, and causal modeling—can be trialed under supervision. Throughout, explainability is not a luxury; it is the affordance that allows an algorithmic claim to enter the community of reasons.

The second element is competence. Investigators require a curriculum that blends operational expertise with data literacy and cultural intelligence. CBTA provides the architecture for defining and assessing these competencies; CRM pedagogy offers the culture of practice in which questioning is a duty, not a discourtesy. Training must teach not only how to use tools but how to doubt them: to ask about class imbalance, dataset shift, and the difference between correlation and cause; to recognize when an NLP tag reflects accent rather than intent; to demand reproducibility before acting. When competence is coupled with just-culture protections, adoption sheds fear and takes on the energy of craft advancement.

Third, inter-agency collaboration is a force multiplier. Shared model repositories, anonymized benchmark datasets, and common validation protocols reduce duplication and improve generalization. Regulators can catalyze this by endorsing minimum documentation standards (model cards, audit trails), by clarifying admissibility thresholds for AI-assisted evidence, and by funding cross-modal pilots. Collaboration should extend beyond technologists to include sociolinguists and cultural-psychology experts who can help de-bias NLP and advise on cross-cultural interpretation of audio and narrative sources. The benefit is not political optics but analytic integrity.

Fourth, cultural ergonomics will matter more as automation deepens. AI systems participate in communication; their advisories and alarms have a "style" that people receive through the filters of culture and training. If we want investigators to reason well about human–automation breakdowns, we must design and train with culture in mind—both in operations and in analysis. CQ is the bridge: it equips professionals to read signals—human or machine—with the charity and precision that safety deserves.

The research agenda is promising. We need multi-site, multi-modal studies to quantify AI's effect sizes on investigation timelines, hypothesis accuracy, and recommendation quality. Additionally, we demand error taxonomies for AI-assisted steps so that we can learn from model failures as we do from human ones (body of case law and policy that treats AI outputs as evidence under explanation, not as opaque assertions). Above all, we need to keep the work human: investigations serve not only to explain mechanisms but to honor losses with truth-seeking that is competent, fair, and teachable. That decision remains with investigators who know that causality in complex systems is a story we earn the right to tell—by listening, by testing, and by learning under a human centric approach.

ACKNOWLEDGMENT

The authors thank Purdue University and Hellenic Air Force faculty members, HF Horizons analysts for their invaluable feedback, which contributed to this work.

REFERENCES

Altemeyer, B. (1996). *The authoritarian specter*. Harvard University Press.

Ang, S., & Van Dyne, L. (2008). Conceptualization of cultural intelligence: Definition, distinctiveness, and nomological network. In S. Ang & L. Van Dyne (Eds.), *Handbook of cultural intelligence: Theory, measurement, and applications* (pp. 3–15). M. E. Sharpe.

Gudykunst, W. (2002). Intercultural communication theories. In W. Gudykunst & B. Mody (Eds.), *Handbook of international and intercultural communication* (2nd ed.). Sage.

Helmreich, R. L., & Merritt, A. C. (1998). *Culture at work in aviation and medicine: National, organizational, and professional influences*. Ashgate.

Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (1999). The evolution of Crew Resource Management training in commercial aviation. *International Journal of Aviation Psychology*, 9(1), 19–32.

International Civil Aviation Organization (ICAO). (2013). *Doc 9995: Manual of Evidence-based Training*. Montreal, Canada: International Civil Aviation Organization.

Kanki, B. G., Anca, J., & Helmreich, R. L. (2010). *Crew resource management* (2nd ed.). Academic Press.

Reason, J. (1997). *Managing the risks of organizational accidents*. Ashgate.

Wiegmann, D. A., & Shappell, S. A. (2003). *A human error approach to aviation accident analysis: The human factors and analysis and classification system*. Routledge.

Ziakkas, D., & Plioutsias, A. (2024). *Artificial intelligence and human performance in transportation*. https://doi.org/10.1201/9781003480891.