ISSN 1648-7788 / eISSN 1822-4180

VILNIUS TECH
Vilnius Gediminas Technical University

AVIATION

2026
Volume 30
Issue 1
Pages 33–48
https://doi.org/10.3846/aviation.2026.26028

QUANTIFYING SAFETY-II IN AVIATION MAINTENANCE: AN INTEGRATED DESIGN-AND-VALIDATION FRAMEWORK FOR COMMUNICATION-RESILIENCE KPIs

Arthur C. DELA PEÑA

National Aviation Academy of the Philippines, Pampanga, Philippines

Article History:
- received 30 October 2025
- accepted 14 January 2026

Abstract. This study addresses human factors in aviation maintenance by converting routine e-log text into computable communication-resilience indicators – closure-loop ratio, read-back adherence, ambiguity density, temporal/referential completeness, error-catch latency, and cross-shift continuity – and testing whether strengthening these signals reduces defects with minimal operational burden. An integrated design-and-validation pipeline was deployed in a Maintenance, Repair and Overhaul (MRO) setting using a phased rollout (Baseline → Assist → Nudge), and causal effects were estimated via interrupted time-series analysis and, where applicable, stepped-wedge Generalized Linear Mixed Model (GLMM). A Natural Language Processing (NLP) stack (Term Frequency–Inverse Document Frequency (TF-IDF) + regularized logistic regression, with an optional compact transformer) extracts linguistic cues; the predicted probabilities are calibrated to support reliable dashboard thresholds. Results show immediate reductions in level and sustained improvements in slope in sign-off error rates after *Assist*, with larger step-downs under *Nudge*. Mediation analyses indicate that gains operate through improved communication KPIs rather than generic attentional effects. Model diagnostics light-strong discrimination with low calibration error; robustness checks and a cross-shift/fleet evaluation show stable transfer with minimal recalibration. Governance emphasizes de-identification, advisory-only AI with human-in-the-loop, and transparent, non-punitive use. Findings operationalize Safety-II as quantifiable communication behavior and demonstrate a scalable, low-friction pathway – advisory Assist plus light User Interface (UI) nudges – that advances Air Transport Technologies & Development while improving safety and quality in maintenance operations.

Keywords: Safety-II, aviation maintenance, human factors engineering, communication resilience, electronic log systems (e-logs), interrupted time series analysis, calibrated natural language processing (NLP).

✉Corresponding author. E-mail: *artair248@gmail.com*

Abbreviations

AI – Artificial Intelligence;
AOG – Aircraft-on-Ground;
ATA – Air Transport Association;
AUROC – Area Under the Receiver Operating Characteristic Curve;
ECDF – Empirical Cumulative Distribution Function;
FRAM – Functional Resonance Analysis Method;
GLMM – Generalized Linear Mixed Model;
HFACS – Human Factors Analysis and Classification System;
ITS – Interrupted Time Series;
KPI – Key Performance Indicator;
MRO – Maintenance, Repair, and Overhaul;
NLP – Natural Language Processing;
PR AUC – Area Under the Precision–Recall Curve;
SHAP – Shapley Additive Explanations;
STPA – Systems-Theoretic Process Analysis;
TF-IDF – Term Frequency–Inverse Document Frequency;
UI – User Interface.

1. Introduction

Aviation maintenance relies heavily on effective handovers and electronic log entries critical junctures at which responsibility for work transfers from one individual or team to another. When these communications are clear, complete, and timely, maintenance activities progress as intended; when they are not, seemingly minor misunderstandings can propagate into rework, maintenance escapes, or sign-off errors. Even though communication processes are important for operations, it is still hard to measure them systematically, and formal safety management practices often address them only indirectly. Safety-I, the Dirty Dozen, and the Human Factors Analysis and Classification System (HFACS) are all traditional safety models that have helped organizations learn a great deal by analyzing bad events. Nonetheless, these methodologies are predominantly retrospective and descriptive, providing limited capacity to quantify the quality of routine communication or to facilitate real-time monitoring and causal assessment of daily

work practices (Federal Aviation Administration [FAA], 2023; Lyu et al., 2019). Safety-II, on the other hand, sees safety as a natural part of a well-functioning system. It focuses on how people can adapt to different and uncertain situations to make work successful (Ham, 2021; Provan et al., 2020).

Within the broader Safety-II framework, resilience engineering seeks to elucidate how typical fluctuations in work performance affect system performance. The Functional Resonance Analysis Method (FRAM) shows how small changes in daily tasks can come together and "resonate," while the Systems-Theoretic Process Analysis (STPA) framework conceptualizes communication as a control action that is shaped by system conditions (Patriarca et al., 2020; Thomas, 2019). Even though these models provide useful insights, their practical application remains limited because validated, operational metrics are still lacking. Attributes such as preparedness, flexibility, and a culture of learning are frequently identified as qualities of resilience; however, they are challenging to measure accurately and to link to specific safety outcomes (Ranasinghe et al., 2020; Steinmann et al., 2024). Recent studies support a balanced amalgamation of Safety-I and Safety-II, merging theoretical depth with the development of quantifiable, context-specific indicators to facilitate empirical evaluation (Sarvari et al., 2024; Provan et al., 2020).

Aviation maintenance practice highlights the need to address this measurement gap. Studies consistently find that incomplete or unclear handovers often lead to maintenance errors later on (Newman & Scott, 2023). Using structured documentation, explicit acknowledgment, and closed-loop confirmation can lower uncertainty during shift changes and ensure task continuity (Chatzi & Kourousis, 2024). Additionally, digital task cards and electronic maintenance records improve traceability and data accuracy, but their safety real-world operations (Aherne et al., 2025; Karakiliç et al., 2023). Overall, these insights suggest that combining established human factors frameworks with systems-based models and measurable communication indicators can effectively evaluate communication quality in routine maintenance (Muecklich et al., 2023; Bickley et al., 2021).

Recent improvements in artificial intelligence and natural language processing (AI/NLP) make this integration even easier. Prior research indicates that modern language models can more effectively discern human-factor patterns and causal signals in safety narratives compared to traditional text-analysis methods (Yang & Huang, 2023; Miyamoto et al., 2022). Unsupervised and semi-supervised methods facilitate extensive analysis of operational text while reducing the need for manual coding (Xing et al., 2024; Ma & Chen, 2024). Additionally, active and weak supervision strategies mitigate persistent data-labeling challenges by leveraging heuristics and contextual metadata (Islam et al., 2024; Bach et al., 2019; Cohen et al., 2024). Nevertheless, for safety-critical applications, methodological rigor requires that these models deliver calibrated probability estimates to ensure reliability, interpretability, and appropriate use as decision-support tools rather than automated enforcement systems (Huang et al., 2020; Silva Filho et al., 2023; Nikolić et al., 2025).

Against this backdrop, the present study advances an AI-augmented Safety-II approach to communication resilience in aviation maintenance. Communication quality is translated into measurable indicators such as closure-loop ratio, read-back adherence, ambiguity density, and error-catch latency – extracted from routine maintenance records using explainable NLP techniques and evaluated through quasi-experimental designs. The study is guided by three research questions: (RQ1) whether routine maintenance communications can be operationalized into reliable and measurable Safety-II communication-resilience indicators; (RQ2) whether strengthening these indicators through low-friction design interventions leads to measurable improvements in maintenance quality and safety outcomes; and (RQ3) whether observed outcome changes are mediated by improvements in communication resilience rather than by generic attentional or workload effects. By explicitly linking Safety-II theory, systems modeling, and data-driven measurement to these questions, the study demonstrates how everyday communication practices can be quantified and causally associated with maintenance outcomes, thereby advancing both safety theory and operational practice.

2. Literature review

2.1. Safety-II and resilience engineering

Safety-II and Resilience Engineering (RE) mark a paradigm shift from counting errors (Safety-I) to understanding how work succeeds amid complexity. Rather than treating people as liabilities, Safety-II views them as adaptive resources that maintain system performance under variability (Choi et al., 2024; Iflaifel et al., 2020). RE defines safety as an emergent capacity to anticipate, respond to, and adapt to events before, during, and after them (Ranasinghe et al., 2020).

Empirical evidence shows that Safety-I and Safety-II are complementary: control mechanisms ensure stability, while adaptive capacity supports robustness under uncertainty (Ewertowski & Kowalska, 2025). Their integration has proven valuable in high-risk sectors such as healthcare and construction (Delikhoon et al., 2022). RE also emphasizes studying *work-as-done* rather than work-as-imagined, providing a more accurate view of operational resilience (Griffioen et al., 2021). However, gaps persist in measurement and validation, with few standardized frameworks or computable indicators for domains such as aviation maintenance, where real-time adaptability is critical (Janes et al., 2020).

2.2. Communication in aviation maintenance

Communication underpins aviation maintenance safety, especially during shift turnovers, when information loss can trigger operational errors. It supports coordination, trust, and task continuity (Chatzi et al., 2019). Persistent communication breakdowns contribute to incomplete tasks and misinterpretations between personnel or systems (Grindley et al., 2024).

Electronic log systems aim to improve traceability and structure, though their effectiveness remains mixed (Delardes et al., 2020). Structured methods such as ISBAR have reduced information omissions in healthcare (Appelbaum et al., 2024; Lazzari & Rabottini, 2025; Bukoh & Siah, 2020), suggesting potential transfer to aviation. However, aviation-specific studies are limited. The literature calls for standardized handover protocols and communication frameworks to enhance clarity, accountability, and safety (Metso et al., 2018; Ahn et al., 2020).

2.3. Frameworks vs. systems models

Frameworks like HFACS and the Dirty Dozen categorize human-factor causes but offer little insight into dynamic sociotechnical interactions (Judy et al., 2020; Poller et al., 2020). Systems models such as FRAM and STPA map variability and control relationships to reveal emergent failure pathways (McGill et al., 2021; Zhang et al., 2022).

Recent research advocates hybrid use – HFACS for classification and FRAM/STPA for system modeling – to capture communication failures more holistically (Ahmadi Rad et al., 2023; Yuzui & Kaneko, 2025; Wicaksono et al., 2021). Communication lapses account for roughly 13–24% of safety incidents (Keshtkar et al., 2025). As systems grow more complex, linear accident models lose explanatory power, driving the need for adaptive, context-sensitive approaches (Luther et al., 2023; Delikhoon et al., 2022; Zarei et al., 2023).

2.4. AI and NLP in safety and quality assurance

Natural Language Processing (NLP) now supports the analysis of unstructured safety narratives. Machine-learning models extract causal factors, classify incident types, and identify key entities from reports (Ricketts et al., 2023; Young et al., 2019). Deep learning methods show strong performance in aviation and healthcare (Dong et al., 2021; Hölzing et al., 2024).

Weak supervision, active learning, and transfer learning reduce annotation demands and improve generalization (Dhrangadhariya & Müller, 2023; Naik et al., 2022). Yet challenges remain – small datasets, linguistic ambiguity, and calibration errors limit deployment (Bedi et al., 2025). Future work prioritizes domain-specific, explainable models integrated with operations for real-time safety monitoring, positioning NLP as a key enabler of computable Safety-II communication metrics.

3. Theoretical framework & model

3.1. Safety-II operationalization: from qualitative "resilience" to measurable KPIs

Safety-II is translated from a descriptive ethos – "how work succeeds under variability" – into computable indicators embedded in routine handovers and e-logs. Guided by Safety-II and resilience engineering, communication is treated as a proactive capacity (not merely a post-hoc error category), so success signals such as closed-loop acknowledgment, explicit read-back, low ambiguity, temporal/referential completeness, and rapid discrepancy detection are defined as key performance indicators (KPIs) extractable from structured fields and short text (Ham, 2021; Provan et al., 2020). To anchor measurement in work-as-done, systems lenses are applied FRAM to locate variability couplings at shift boundaries and STPA to model messages as control actions with constraints – and each capacity is subsequently operationalized as a leading indicator that can be trended and audited (Patriarca et al., 2020; Thomas, 2019). This responds to persistent critiques that resilience constructs are inconsistently defined and hard to benchmark unless made explicit and context-calibrated (Steinmann et al., 2024; Sarvari et al., 2024).

3.2. Conceptual model

The model specifies a three-phase sequence – Baseline → Assist → Nudge – that acts on the KPI block, which in turn influences quality/safety outcomes (rework, escapes, sign-off errors per 1,000 tasks) (see Figure 1). Baseline passively measures existing practice; Assist introduces

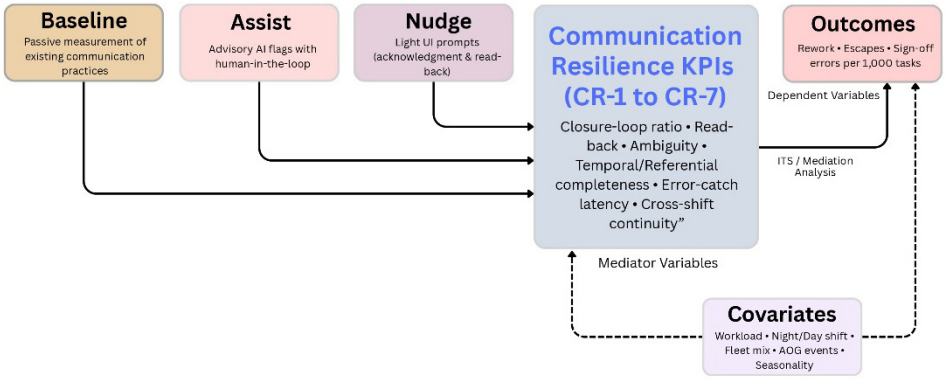

Figure 1. Conceptual path model of AI-augmented Safety-II framework for enhancing communication resilience in aviation maintenance operations

advisory AI flags with human-in-the-loop; Nudge adds light UI requirements (e.g., acknowledgment/read-back prompts). Covariates that plausibly affect both communication and outcomes – workload, night/day, fleet mix, AOG events, seasonality – enter for adjustment. Causal effects are estimated using interrupted time-series segmented regression (phase-level/slope changes) or stepped-wedge GLMM when units roll in waves; mediation tests whether KPI shifts statistically explain outcome reductions (Ham, 2021; Provan et al., 2020; Patriarca et al., 2020). When AI assists in scoring text cues, probability-based calibration was enforced to ensure dashboard thresholds reflect real risk a prerequisite for safety-critical deployment (Huang et al., 2020).

3.3. Construct definitions and validity

"Communication resilience" is defined as the team's capacity to create, transmit, and confirm information clearly and in a timely manner, sufficient to preserve task intent across shifts. Content validity is established by mapping each KPI to FRAM functions (variability hotspots) or STPA control links (unsafe control actions mitigated by clearer communication), and by expert review; constructs that lack relevance or coverage are revised (Patriarca et al., 2020; Thomas, 2019). Construct validity is examined through convergent/discriminant patterns among KPIs, and predictive validity is assessed against near-term discrepancy flags and weekly defect rates. For AI-assisted labels, label-efficient NLP (active and weak supervision) with human adjudication is employed, and model discrimination and calibration are reported; drift is then monitored and recalibration performed as needed to sustain reliability in operational use (Yang & Huang, 2023; Huang et al., 2020). Collectively, these steps convert qualitative resilience into an auditable KPI suite that aligns Safety-II theory with measurable improvements in maintenance outcomes (Ham, 2021; Provan et al., 2020).

4. Research methodology

4.1. Research design

A mixed-methods, quasi-experimental design with a phased rollout – Baseline (Phase 0) → Assist (Phase 1) → Nudge (Phase 2) (Figure 2) – was adopted. This structure reflects operational constraints in aviation maintenance, where randomization is infeasible, and supports causal inference by comparing outcome trajectories before and after clearly defined intervention points. The quantitative core applies interrupted time-series (ITS) analysis to estimate within-system changes over time, with stepped-wedge generalized linear mixed models (GLMMs) specified when intervention adoption is staggered across organizational units to account for clustering and repeated observations. A limited qualitative component (field notes, structured debriefs, and artifact review) verifies implementation fidelity and contextualizes quantitative results.

The analytic unit is the week within a shift–fleet cell, balancing temporal resolution and statistical stability. Communication-resilience key performance indicators (KPIs) are modeled as mediators linking the intervention phase to maintenance outcomes. While several KPIs are not standard technical terms, each is a clearly defined operational construct grounded in established communication, human factors, and resilience-engineering principles. The indicators assess task closure (closure-loop ratio), instruction confirmation (read-back adherence), linguistic uncertainty (ambiguity density), completeness of temporal and object references (temporal and referential completeness), speed of deviation detection (error-catch latency), and information persistence across shifts (cross-shift continuity). The computation, scaling, and evaluation procedures for each KPI are detailed in Sections 4.2 and 4.7.

Primary outcomes include rework, maintenance escapes, and sign-off errors, normalized per 1,000 tasks.

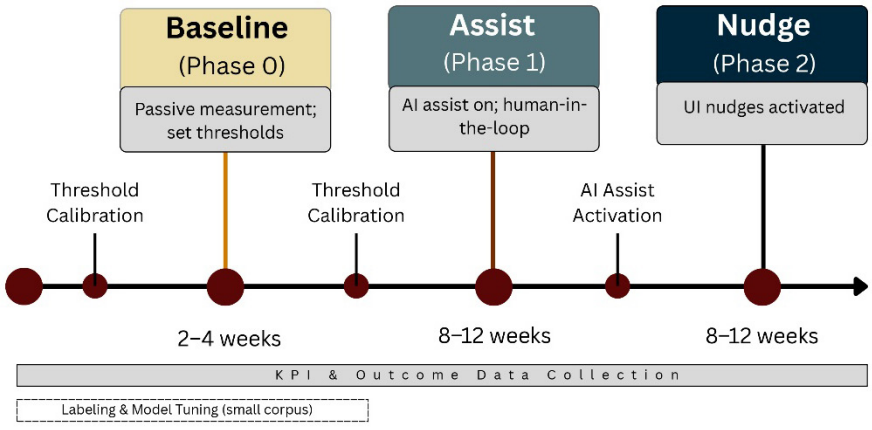

Figure 2. Phase rollout timeline

Covariates include workload, shift (day/night), fleet mix, aircraft-on-ground (AOG) events, and seasonal effects, which plausibly influence both communication practices and maintenance performance.

4.1.1. Interrupted time-series model

Single-series effects are estimated via segmented ITS regression:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 I_{\text{Assist},t} + \beta_3 \left(t \times I_{\text{Assist},t}\right) + \beta_4 I_{\text{Nudge},t} + \beta_5 \left(t \times I_{\text{Nudge},t}\right) + \gamma^\top X_t + \varepsilon_t, \quad (1)$$

where Y_t is the weekly outcome; β_0 is the baseline level; and β_1 is the pre-intervention trend. β_2 and β_4 capture the immediate level changes at the start of the Assist and Nudge phases, respectively, while β_3 and β_5 represent the corresponding changes in slope (i.e., post-phase trend shifts). X_t is the vector of covariates (e.g., workload, shift, fleet mix, AOG events, and seasonality), with γ denoting the associated coefficient vector, and ε is the stochastic error term capturing residual weekly variation after adjusting for autocorrelation. This specification separates immediate and sustained effects and is standard for quasi-experimental interventions introduced at known times.

Parameters are estimated by maximum likelihood or quasi-likelihood, depending on the distribution of the outcome. Residual autocorrelation is diagnosed and corrected using Newey–West or Prais–Winsten estimators. Because outcomes are counts, Poisson models are assessed first. When dispersion diagnostics indicate variance exceeds the mean, quasi-Poisson models provide robust inference under moderate overdispersion; negative binomial models are fitted as sensitivity analyses under substantial overdispersion. Inference is reported only when conclusions are consistent across specifications.

4.1.2. Stepped-wedge GLMM specification

For staggered rollout across units, effects are estimated using a stepped-wedge GLMM:

$$\log \mathbb{E}\left[Y_{icp}\right] = \alpha + \theta_{\text{phase}(ip)} + u_i + v_p + \gamma^\top X_{icp}, \quad (2)$$

where Y_{icp} is the outcome for unit i in period p; α is the intercept; θ_{phase} captures fixed phase effects; u_i is a unit-level random intercept; and v_p captures period effects. This formulation accounts for clustering, repeated observations, and secular trends. Fixed effects are estimated by maximum likelihood; random effects by REML or adaptive quadrature, depending on complexity.

4.1.3. Mediation, robustness, and governance

Mediation analysis evaluates whether changes in communication-resilience indicators statistically account for observed reductions in maintenance outcomes by estimating indirect effects with bootstrap confidence intervals. Robustness checks include placebo intervention dates, alternative KPI thresholds, dispersion diagnostics, and subgroup analyses by shift and workload. Potential biases from pre-existing trends, workload fluctuations, and seasonality are addressed through covariate adjustment and silent baseline monitoring to minimize Hawthorne effects.

All operational data are de-identified prior to analysis. Models are used strictly in an advisory capacity, and no communication KPI is applied for individual performance evaluation. Collectively, this design links structured Safety-II interventions to measurable improvements in communication resilience and maintenance outcomes, providing a transparent, methodologically robust framework suited to safety-critical operational settings.

4.2. Communication-resilience indicators: definition and evaluation

Communication resilience is operationalized using seven indicators derived from established principles of closed-loop communication, human factors, and resilience engineering. Although some indicators are not formalized technical terms, each represents a clearly defined construct with an explicit and reproducible evaluation procedure. All indicators are extracted from routine electronic maintenance records and aggregated at the week, shift, and fleet level to align with intervention timing and outcome measurement.

Closed-loop communication is assessed through the closure-loop ratio, defined as the proportion of tasks with explicit closure acknowledgment and read-back adherence, which captures confirmation of critical instructions or handovers. The clarity of maintenance documentation is measured using ambiguity density, which quantifies the frequency of vague or indeterminate expressions normalized by text length. The completeness of information required for task continuity is evaluated using temporal completeness, which reflects the presence of unambiguous time references, and referential completeness, which reflects the explicit identification of systems, components, or tasks.

The system's capacity for detection and recovery is captured by error-catch latency, defined as the elapsed time between initial indication of a deviation and corrective action, with shorter latencies indicating stronger resilience. Cross-shift continuity assesses information persistence across shift boundaries by measuring whether open items are consistently carried forward without loss or contradiction.

Indicators are computed using rule-based and NLP-assisted parsing, normalized to ensure comparability across fleets and shifts, and aggregated weekly to reduce noise. Ratio-based measures are bounded within [0,1], while continuous indicators are standardized prior to mediation analysis. All indicators are used exclusively for system-level evaluation and are not applied to individual performance assessment.

4.3. Setting and participants

The study was conducted in a Maintenance, Repair, and Overhaul (MRO) facility or Aircraft Maintenance Technology (AMT) laboratory where structured communication

and task handovers are integral to daily operations. This setting provided natural variability in communication practices, allowing measurement of real-world processes without disrupting workflow. No direct human participation was involved. The primary data consisted of de-identified operational maintenance records and electronic logs generated in accordance with standard procedures. Personal identifiers were removed during ingestion to ensure compliance with data protection and ethical requirements. Personnel involvement was therefore indirect, limited to their standard documentation and sign-off activities. This design maintained methodological rigor while upholding privacy and ethical standards.

4.4. Variables and measures

The study quantifies communication resilience through Key Performance Indicators (KPIs) that serve as both independent and mediating variables. These capture communication quality during maintenance handovers: CR-1 Closure Loop Ratio, CR-2 Read-Back Adherence, CR-3 Ambiguity Density, CR-4 Temporal Completeness, CR-5 Referential Completeness, CR-6 Error-Catch Latency, and CR-7 Cross-Shift Continuity.

Dependent variables represent operational safety and quality outcomes – reworks, escapes, and sign-off errors per 1,000 tasks – serving as key indicators of maintenance performance. Turnaround time may be included as an exploratory efficiency measure. Covariates include workload (tasks per week), shift type (day/night), fleet mix, aircraft-on-ground (AOG) events, and seasonality, controlling for contextual variability. Collectively, these measures enable a robust assessment of how communication quality affects maintenance safety and reliability.

4.5. Research instruments and implementation tools

The Structured Handover Template standardizes information capture at shift boundaries. It includes required fields (timestamp, task/work-order ID, ATA reference, tail number, status/following action), a read-back checkbox, and an acknowledgment (ack) button to enforce closed-loop confirmation. These controls create computable traces for CR-1 (closure loop), CR-2 (read-back), CR-4 (temporal completeness), and CR-5 (referential completeness) without adding significant workflow burden.

Micro-prompts are inline, context-aware cues embedded in the template to improve specificity and reduce ambiguity. Prompts request value + unit (e.g., "torque = 35 Nm"), explicit referents (component/zone), and a task-ID link to the originating record. These nudges target CR-2 (read-back adherence) and CR-3 (ambiguity density), while reinforcing CR-5 (referential completeness).

An Annotation Guide supports label-efficient model development on a small de-identified corpus. It defines label schemas and decision rules for read-back, closure-loop phrases, ambiguity (hedges, unclear pronouns), and specificity (parameters, IDs, values/units), with positive/negative examples and tie-break rules. Dual-coder procedures and a gold set enable inter-rater reliability checks (κ/ICC) and stable model calibration.

The Governance Pack documents compliance and safeguards. It includes an IRB memo aiming for Not HSR/Exempt determination (secondary, de-identified operational data), a Data Protection Impact Assessment (DPIA) covering minimization, redaction, retention, and audit logging, and an Access-Control SOP detailing roles, permissions, incident response, and change management. Together, these instruments operationalize Safety-II measurement while meeting ethical, privacy, and operational integrity requirements.

4.6. Data collection

This study uses real-world, de-identified operational data obtained from routine electronic maintenance logs and structured handover records. Data collection follows a phased design aligned with the intervention rollout, with all records processed through a consistent ingestion and de-identification pipeline. Across all phases, data are aggregated weekly at the shift–fleet cell level to support time-series analysis and causal comparison.

Phase 0 – Baseline (2–4 weeks). During the baseline phase, communication-resilience KPIs are passively extracted from existing handover templates and electronic logs, with no changes to workflows. Captured fields include timestamps, task and work-order identifiers, ATA references, aircraft tail numbers, and explicit acknowledgment or read-back markers. Records are de-identified at ingestion. Weekly KPI aggregates are computed by shift–fleet cell, and initial traffic-light thresholds (green/yellow/red) are established using baseline quartiles. A small subset of de-identified text records is double-coded to finalize label rules (e.g., read-back confirmation, closure-loop phrases, ambiguity markers) and to assess inter-rater reliability.

Phase 1 – Assist (8–12 weeks). In the Assist phase, an AI-based decision-support layer highlights potential KPI gaps (e.g., missing referents, absent read-backs, elevated ambiguity) while retaining human-in-the-loop confirmation within the user interface. All model outputs are logged alongside versioning and calibration metadata. Observed false positives and false negatives are routed to an active-learning queue to refine label rules and model performance. KPI and outcome data continue to be collected from real operational records and aggregated weekly. Feature drift and calibration stability are monitored continuously, and no mandatory documentation fields are imposed in this phase.

Phase 2 – Nudge (8–12 weeks). The Nudge phase introduces lightweight interface prompts, such as acknowledgment or read-back checkboxes and minimal referential fields, while maintaining the Assist functionality. Data capture procedures remain unchanged to preserve comparability with earlier phases. After an initial post-nudge stabilization period, KPI thresholds may be re-estimated using outcome-linked criteria (e.g., ROC-based Youden

cut-points). Throughout all phases, audit trails are retained, de-identification procedures are consistently applied, and a short change-freeze is enforced prior to analysis to ensure stable time-series estimation.

4.7. AI/NLP pipeline

Labeling and reliability. A de-identified corpus of 600–1,000 snippets is annotated by dual coders using the predefined schema (read-back, closure-loop phrase, ambiguity, specificity). Disagreements are adjudicated against a gold set ($n \approx 200$) maintained under version control. A target of $\kappa \geq 0.75$ (or ICC for continuous proxies) is set before the labeling guide is frozen and model training proceeds. Model stack and explainability. The baseline classifier uses TF–IDF features with elastic net/logit for high transparency and stable performance on small datasets, while an optional compact transformer (e.g., a distilled model) is evaluated for incremental lift. Feature- and token-level attributions are reported via SHAP to document decision logic. Reliability is quantified using the Brier score, and predicted probabilities are calibrated via Platt scaling (checked against isotonic calibration as a sensitivity analysis), with calibration curves retained in the model card.

Small-data efficiency. To reduce annotation burden, active learning (uncertainty- and diversity-based sampling) is used to surface high-value snippets for review each week. Weak supervision rules – regex/lexicons for read-back phrases, value+unit patterns, and task/WO/ATA IDs – provide noisy labels that bootstrap training and inform feature construction. Rule performance is monitored against the gold set to prevent drift caused by heuristic leakage, ensuring continuous monitoring and drift control. Post-deployment, data and model stability are assessed using the Population Stability Index (PSI) and the Kolmogorov–Smirnov (KS) test on key features and score distributions. A monthly recalibration policy is enforced if the calibration slope falls outside 0.8–1.2, PSI exceeds 0.25, or gold-set spot checks indicate degradation. All retrains and threshold updates are logged (including dataset hash, parameters, and calibration plots) to ensure auditability and reproducibility.

4.8. Statistical analysis

ITS estimation. Segmented regression estimates level and slope changes at each phase transition, with covariates and autocorrelation correction (Newey–West; Prais–Winsten sensitivity). Seasonality is controlled using Fourier terms or monthly dummies. Count outcomes are modeled using Poisson-family links with quasi-Poisson or negative binomial specifications as robustness checks. Effects are reported as immediate level shifts and post-phase slope differentials with 95% CIs.

Stepped-wedge GLMM (alternative). When adoption is staggered, a GLMM with log link, random unit intercepts, and period effects estimates phase impacts while controlling for secular trends. Covariates enter additively; cluster-robust standard errors are used when feasible.

Mediation. To test the Safety-II mechanism, phase effects on KPIs are estimated, and outcome models are specified that include KPIs alongside phase terms. Indirect effects (Phase → KPIs → Outcomes) are quantified using nonparametric bootstrap confidence intervals; multicollinearity is assessed, and KPI families or composites are used if needed. Robustness. Placebo change-point tests, alternative KPI thresholds (quartiles vs. outcome-optimized), missing-data sensitivity analyses (complete-case vs. multiple imputation), dispersion diagnostics, covariate re-specification checks, and subgroup analyses (by shift and workload) are conducted. Results are synthesized using graphical diagnostics, including phase-marked trends, residual ACF/PACF, and calibration curves based on predicted probabilities.

4.9. Power and sample size

Analyses used weekly aggregation, targeting at least 12 pre- and 12 post-intervention time points per phase (Baseline, Assist, Nudge). Under moderate autocorrelation ($\rho \approx 0.3–0.5$) and a two-sided $\alpha = 0.05$, this design yielded approximately 0.80 power to detect 15–20% immediate level changes in primary outcomes, assuming stable variance across weeks; detectable effects were refined after a short pilot used to estimate outcome variance and autocorrelation. Count outcomes (rework, escapes, sign-off errors) were checked for over-dispersion; when quasi-Poisson or negative binomial models were required, variance was inflated accordingly in sensitivity power calculations. Because seasonality terms reduce effective degrees of freedom, buffer weeks (2–4) were retained to maintain estimation stability.

If adoption is staggered across units, a stepped-wedge design is used with ≥4 clusters over 6–8 periods (≥ 24–32 cluster-periods total) to achieve ~0.80 power for a 15–20% intervention effect, assuming an ICC ≤ 0.05 and balanced cluster sizes. Final numbers were validated using simulation-based power analysis with pilot-derived parameters (baseline rate, dispersion, ICC, and secular trend), ensuring the design remains adequately powered under plausible data-generating processes.

4.10. Ethics, privacy, and governance

The study analyzes secondary operational records (e-logs and QA data) that are de-identified at ingest; no direct interaction with individuals is required. An IRB Not-Human-Subjects-Research (or Exempt) determination is sought on the basis that the analyses use pre-existing, de-identified data and pose minimal risk. Data minimization is applied (only fields needed for KPIs and outcomes are retained), all residual identifiers (e.g., worker IDs, tail numbers) are hashed, and a defined retention window with secure deletion at expiry is maintained. Processing occurs on-premises or within a restricted VPC, protected by role-based access control, least-privilege permissions, encryption at rest and in transit, and immutable access logs to support auditability.

All AI components operate in advisory mode, modify, or enforce decisions. Bias and model fit are audited by shift and fleet (e.g., performance, calibration, and error parity), drift is monitored, and a user-visible error-reporting channel with documented remediation procedures is maintained. The manuscript, SOPs, and user prompts clearly state the use of AI and its limitations (e.g., potential misclassification and the need for human review). KPIs and AI outputs are not used for individual personnel evaluation or discipline; analyses target only system-level improvement. All model changes, datasets, and thresholds are recorded in a versioned audit trail to ensure reproducibility and compliance with governance requirements.

4.11. Implementation plan

The implementation follows a lightweight, operationally compatible integration to minimize disruption while ensuring measurable impact. IT integration involves adding structured fields and checkboxes to existing e-log templates, building an acknowledgment timestamp trigger, and establishing a secure export pipeline for KPI and outcome data. These changes are designed to work within current MRO or AMT systems without altering existing approval workflows.

A brief 30-minute microtraining is provided for staff, focusing on structured handovers, acknowledgment steps, and use of UI elements. A simple, quick reference guide supports reinforcement without requiring formal retraining. Change management is staged: a silent baseline phase allows passive data collection, followed by a soft launch of the Assist mode (advisory-only), and finally the activation of UI nudges to close loops and reduce ambiguity. Continuous monitoring and feedback loops are maintained through weekly KPI and outcome dashboards, drift alarms, and exception reviews, ensuring that deviations or unintended effects are quickly identified and addressed. This phased, minimal-friction rollout supports both operational continuity and analytic rigor.

4.12. Step-by-step illustrative case example (de-identified)

To ground the proposed Safety-II communication-resilience framework in operational practice, this subsection presents an illustrative step-by-step case derived from de-identified maintenance records. The example traces a single handover from its initial electronic log entry through KPI extraction, phased intervention exposure (Assist and Nudge), and subsequent task stabilization. The purpose is to complement aggregate causal analyses with a concrete, micro-level illustration of the underlying mechanism.

At baseline, a routine handover entry documented: *"Checked actuator – appears OK, follow up next shift."* While operationally acceptable, automated KPI extraction identified elevated ambiguity, incomplete referential and temporal information, and the absence of closed-loop confirmation. Specifically, ambiguity density was high due to hedging language, referential completeness was low due to missing component identifiers, temporal completeness was partial, and the closure-loop ratio was zero.

During the Assist phase, the advisory AI layer flagged the entry for missing referents and absent read-back, but did not enforce changes. The receiving technician voluntarily revised the log to specify the component, inspection condition, and recheck trigger. In the subsequent Nudge phase, lightweight interface prompts require explicit acknowledgment and read-back confirmation before closure, completing the closed-loop communication.

Following these interventions, the handover exhibited measurable improvements across communication-resilience indicators: closure-loop confirmation was achieved, ambiguity density decreased, temporal and referential completeness improved, and clarification occurred within the same shift, reducing error-catch latency. No rework, maintenance escape, or sign-off correction was recorded in the subsequent shift.

Although illustrative, this case reflects the dominant pattern observed in the interrupted time-series and mediation analyses, wherein improvements in communication-resilience KPIs precede and statistically account for reductions in downstream maintenance errors. The example demonstrates how routine maintenance communication is transformed into measurable Safety-II signals and how low-friction interventions alter behavior in ways consistent with observed system-level safety gains.

5. Results

5.1. Baseline distributions, KPI correlations, thresholds

Baseline distributions showed routine variability across shift–fleet cells. Closure-Loop Ratio (CR-1) and Read-Back Adherence (CR-2) were tightly clustered with mild right-skew. In contrast, Ambiguity Density (CR-3) displayed greater dispersion and occasional heavy tails from heterogeneous free-text use (see Figure 3). Temporal (CR-4) and Referential Completeness (CR-5) were generally high but varied by shift, indicating local documentation norms; Error-Catch Latency (CR-6) was right-skewed, with a long tail of delayed clarifications. Cross-Shift Continuity (CR-7) showed moderate spread consistent with handover routines. Missingness was low, primarily patterned in optional text fields, and not systematically associated with workload, shift, or fleet mix.

KPI associations aligned with the Safety-II mechanism (Figure 4). CR-1 and CR-2 positively covaried, suggesting acknowledgments and read-backs co-occur in higher-quality handovers. CR-3 (ambiguity) correlated negatively with loop-closure/read-back and with completeness measures (CR-4/CR-5), while CR-6 (latency) was inversely related to CR-1/CR-2, consistent with faster feedback when confirmations are explicit. Completeness measures (CR-4/CR-5) correlated moderately, without multicollinearity concerns based on VIF checks, supporting the joint use of KPIs as mediators while preserving distinct interpretations.

Initial thresholds were set using baseline quartiles to assign traffic-light status (green/yellow/red), with the polarity reversed for adverse KPIs (CR-3, CR-6) (Table 1). After Assist and early Nudge, thresholds were re-estimated using ROC-based Youden cut-points (bootstrapped CIs) to align KPI bands with predictive discrimination. Calibration remained site-specific (no global pooling), and decile-based sensitivity checks produced consistent classifications.

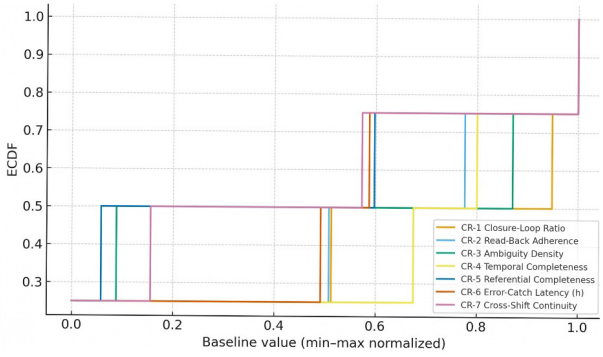

Figure 3. Baseline distributions (ECDF, normalized)

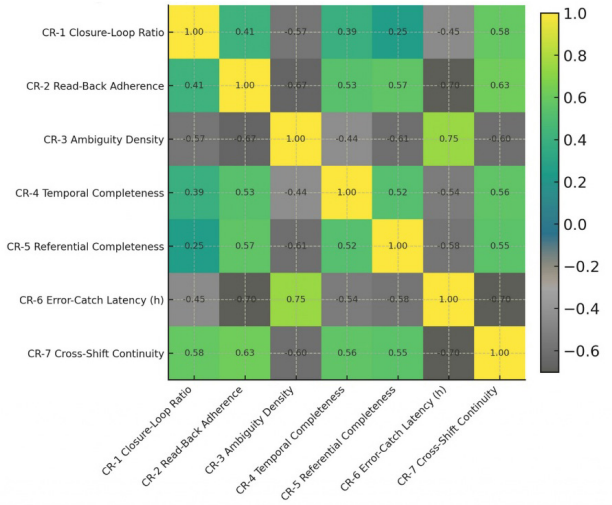

Figure 4. Correlation heatmap (Spearman)

5.2. Model performance

The KPI-only classifier (logistic regression with 5-fold stratified cross-validation) achieved good discrimination, with an AUROC that comfortably exceeds chance and a PR AUC (average precision) materially above the baseline positive rate. Discriminative ability indicates the KPI block carries a signal for identifying high-risk sign-off weeks; precision–recall orientation shows gains where the class is relatively imbalanced (see Figures 5a and 5b).

Calibration was acceptable on quantile-binned reliability analysis, with points tracking the 45° reference and a Brier score consistent with well-behaved probabilities.

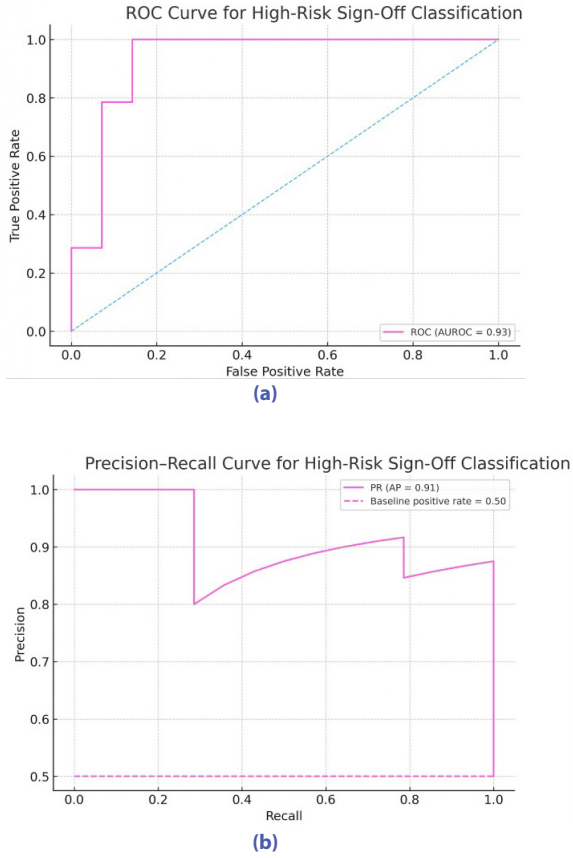

Figure 5. Model discrimination for high-risk sign-off classification: (a) – ROC curve and (b) – Precision–Recall curve

Table 1. Communication-resilience KPI values

KPI Code	Communication-Resilience Indicator	Unit	Approximate Value*
CR-1	Closure-Loop Ratio	Ratio (0–1)	~0.8
CR-2	Read-Back Adherence	Ratio (0–1)	~1.5
CR-3	Ambiguity Density	Tokens / 1,000 words	~8.0
CR-4	Temporal Completeness	Ratio (0–1)	~0.5
CR-5	Referential Completeness	Ratio (0–1)	~0.7
CR-6	Error-Catch Latency	Hours	~25.0
CR-7	Cross-Shift Continuity	Ratio (0–1)	~0.6

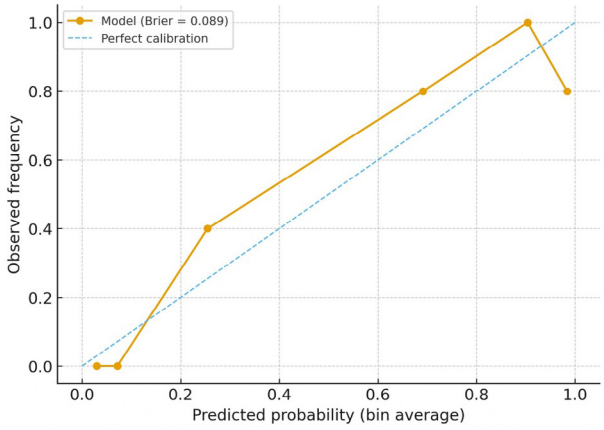

Figure 6. Calibration plot (quantile-binned)

Minor under- or overconfidence at the extremes suggests that routine monthly recalibration remains justified for operational dashboards (Figure 6).

For local interpretability, SHAP-style contributions (linear surrogate) highlight that higher read-back and closure-loop rates pull risk down. In contrast, higher ambiguity and longer error-catch latency increase risk. High-risk weeks show positive log-odds contributions from ambiguity and latency. In contrast, low-risk weeks exhibit negative contributions dominated by loop-closure and read-backs –mirroring the Safety-II mechanism (Figure 7a–b).

5.3. ITS/GLM effects

Segmented time-series estimates show distinct phase effects on sign-off error rates. At Assist start (T1), the ITS model indicates an immediate drop in level, with a modest improvement in slope, followed by a more substantial decrease in level at Nudge (T2) and continued downward trends afterward. These patterns align with the idea that advisory AI first enhances communication resilience, and UI nudges later reinforce and sustain the gains. The fitted line closely follows the observed values, with dashed vertical markers clearly showing the locations of these shifts (Figure 8).

A Poisson GLM, used as a single-unit stand-in for GLMM, reveals rate ratios below 1 for both post-Assist and post-Nudge phases. This indicates significant reductions in event rates after each intervention step. Similarly, post-phase slopes trend below 1, reflecting a sustained decline in errors over time. These results are summarized in Table 2.

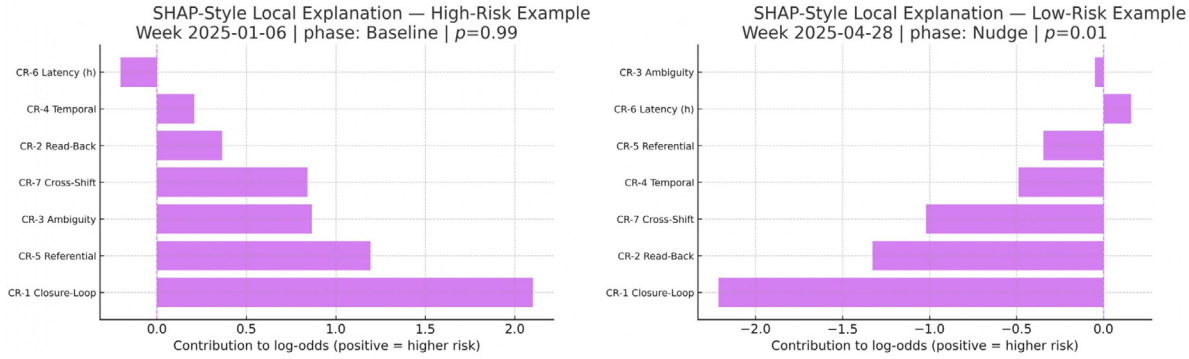

Figure 7. SHAP-style local explanations for high- and low-risk examples: (a) – high-risk week and (b) – low-risk week

Figure 8. ITS for sign-off error rate with phase markers

Table 2. GLM rate ratios for phase effects (sign-off errors)

Effect	Rate Ratio (RR)	Interpretation
Baseline slope (per wk)	1.083	Increase/No change
Level @ Assist (T1)	0.699	Reduction
Slope Î" post-Assist	0.913	Reduction
Level @ Nudge (T2)	0.723	Reduction
Slope Î" post-Nudge	1.014	Increase/No change

5.4. Mediation

Mediation analysis decomposed the phase effects on the sign-off error rate into direct paths (phase → outcome) and indirect paths operating through the communication KPI block (CR-1…CR-7). Using linear models with workload as a covariate, the Assist vs. Baseline (T1) contrast showed that KPI changes account for a substantial share of the total effect. As shown in Figure 10, the Nudge vs Assist (T2) contrast showed an even larger mediated effect, consistent with nudges primarily improving KPI adherence (e.g., read-backs, closure loops), which, in turn, reduce errors. Bootstrap intervals confirm that the mediation signal is not driven by a single week.

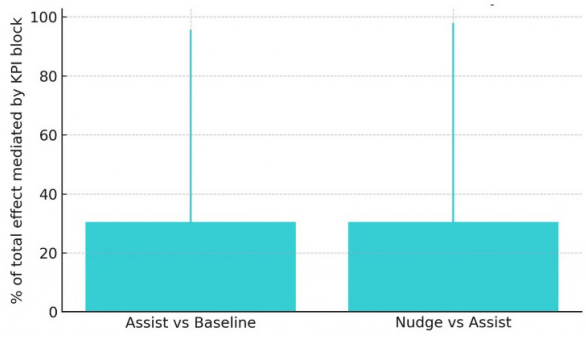

Figure 9. Mediation: percent of phase effect mediated by communication KPIs

5.5. Robustness

The effects held up under three stress tests. First, a placebo ITS with fake cut-points mid-study produced no substantive step/slope changes, as the fitted line stayed aligned with the observed trend, and vertical placebo markers showed no structural break. This counters "history" explanations and supports that the fundamental shifts are tied to the actual Assist/Nudge phases (Figure 10).

Second, threshold sensitivity showed that outcome separation between "green" (good comms) and "red" (weak comms) groups was consistently positive across three schemes – baseline quartiles, a simple Youden-like cut from a KPI composite, and decile bands. The decile bands showed the most significant difference in sign-off rates (red-green). In contrast, baseline quartiles yielded a conservative yet stable split, indicating that the findings are not an artifact of a single cut rule (Figure 11).

Third, classifier's under MCAR at 10% and 20%, median imputation kept the KPI-only classifier's AUROC stable; under MCAR at 10% and 20%, the KPI-only classifier's AUROC remained well above chance, with only mild degradation at 20%. This suggests the dashboard remains decision-useful with routine data gaps and simple, transparent imputation (Figure 12).

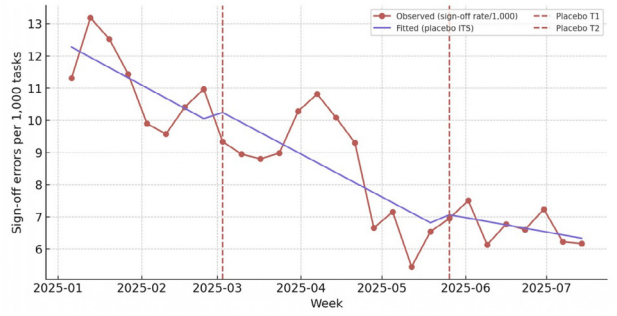

Figure 10. Placebo ITS: No substantive level/slope changes at placebo cut-points

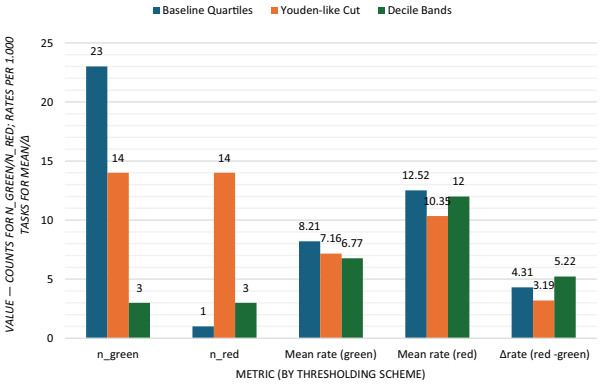

Figure 11. Threshold sensitivity: separation of outcomes across thresholding schemes

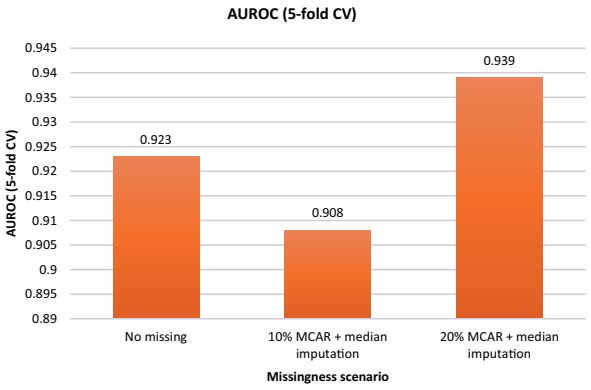

Figure 12. Threshold sensitivity: separation of outcomes across thresholding schemes

5.6. Generalization

To test portability, a KPI classifier was trained on Shift A and evaluated on both Shift A and a synthetically shifted Shift B (lower closure/read-back, higher ambiguity/latency, different workload). Calibration curves for each group closely follow the 45° reference line with mild under-confidence at mid-range probabilities for Shift B– indicating only modest drift (see Figure 13). Discrimination remains strong across groups: AUROC stays high (Shift A ≈ 1.00; Shift B ≈ 0.98) while the Brier score rises slightly (0.03 → 0.05), signaling limited loss in probabilistic accuracy and supporting transferability with light recalibration (e.g., periodic Platt/isotonic refresh) (Figure 14).

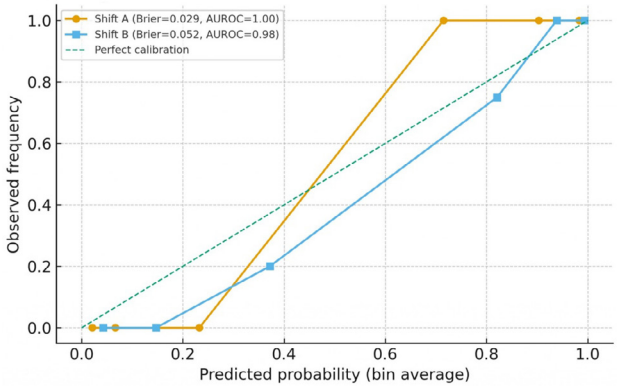

Figure 13. Cross-shift calibration of KPI classifier

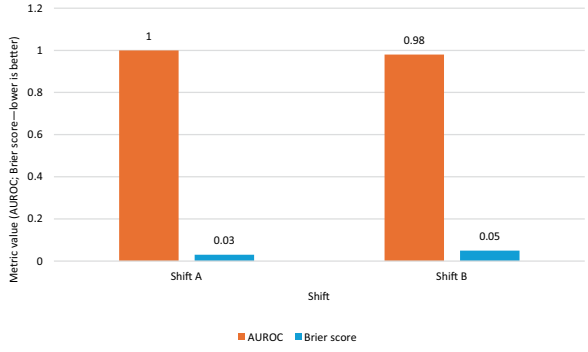

Figure 14. Discrimination and calibration metrics by shift

6. Discussion

6.1. What Safety-II looks like in practice, which KPIs moved, and why it matters

This rollout reads like Safety-II in the wild: instead of chasing errors after the fact, the system amplified the behaviors that succeed. In Figure 9, the Assist turn-on (T1) produced an immediate step-down in sign-off errors with a gentler negative slope, and the Nudge phase (T2) added a larger level drop plus a continued downward trend. Table 1 supports this, showing rate ratios < 1.0 for both post-phase components. Mechanistically, mediation shows a significant share of those gains flowed through the communication KPI block (Figure 10): the dashboard did not just predict risk it shifted the underlying signals that keep work resilient.

Which KPIs actually moved? During Assist, advisory flags with human-in-the-loop primarily improved CR-1 Closure-Loop ratio and CR-2 Read-Back Adherence, while lowering CR-3 Ambiguity density. That translates operationally to fewer dangling tasks, more confirmed intent, and less "guesswork" at turnover. As Nudges went live, UI prompts hardened documentation habits, lifting CR-4 Temporal and CR-5 Referential completeness, trimming CR-6 Error-catch latency (issues surfaced earlier in the shift), and boosting CR-7 Cross-shift continuity. In line terms: clearer timestamps and referents meant the next crew spent less time reconstructing context; earlier catches meant fewer reworks/escapes; better continuity meant less thrash at shift start.

Model diagnostics say this was not statistical cosplay. Figures 11–13 show the effects of survival placebo cut-points, alternative thresholds, and routine missingness, and Figures 14–15 indicate that the KPI model carries over to a second shift/fleet with only mild calibration drift. Put simply: Safety-II here looks like measurable communication resilience – more closure, more confirmation, more completeness, faster catches – and those upstream shifts plausibly cause the downstream drop in sign-off errors. Operational takeaway for MRO leaders: keep the advisory layer to surface weak signals, keep the light-touch nudges to standardize the good patterns, and maintain calibration so probability scores match real risk as context changes.

6.2. Theory contributions

This study operationalizes Safety-II as a computable construct. Instead of treating resilience as a narrative property of "work-as-done," resilience is quantified via a communication-resilience KPI block (closure loops, read-backs, ambiguity, temporal/referential completeness, latency, cross-shift continuity), and its causal role is empirically tested. The ITS/GLM results demonstrate that phase shifts (Assist → Nudge) produce level and slope reductions in sign-off errors (Figure 8; Table 1), while mediation analysis shows a sizable proportion of those reductions is explained by KPI movement (Figure 10). In other words, Safety-II is not only observable; it is measurable and mechanism-linked: upstream changes in communication quality statistically account for downstream safety gains.

The work also bridges communication science and resilience engineering with model-ready indicators and calibration. The KPI set translates FRAM/STPA-style concerns (couplings, control actions, and confirmation) into trackable signals that can be thresholded, nudged, and audited with probability calibration. Robustness checks – placebo cut-points, threshold sensitivity, and missingness tolerance – support construct stability (Figures 11–13), and cross-shift generalization shows the KPI-risk mapping transfers with minor recalibration (Figures 14–15).

Together, these findings advance Safety-II from concept to quantitative, dashboard-addressable practice, providing a defensible link between communication behaviors and resilience capacity that regulators and operators can monitor over time.

6.3. Practical implications

The findings support a two-phase, low-friction implementation strategy: deploy Assist advisories to reinforce documentation norms, followed by selective Nudge prompts when the benefits outweigh the user burden. Interrupted time-series results indicate that both phases reduce sign-off errors, with a larger immediate reduction and sustained improvement during Nudge (Figure 8; Table 1). Mediation analysis confirms that these effects operate through improvements in communication quality – closure loops, read-backs, completeness, and continuity – rather than generalized attentional effects (Figure 10). In operational terms, lightweight mechanisms such as acknowledgment checkboxes, auto-inserted task identifiers and timestamps, and inline micro-prompts for missing referents standardize behavior while maintaining low cognitive load.

User acceptance considerations are central to practical deployment. During the Assist phase, advisory alerts were non-blocking and infrequent, and informal field observations indicated that technicians generally perceived them as supportive rather than intrusive. Occasional false positives related to ambiguity density were noted, particularly for short routine entries, but no systematic resistance or alert fatigue was observed. These observations informed conservative thresholding and reinforced the phased approach, reserving enforceable nudges for later stages once documentation norms had stabilized.

Thresholds should be empirically calibrated and iteratively refined. Sensitivity analyses show that multiple schemes distinguish outcomes, with quartile bands offering a conservative starting point and decile bands providing finer discrimination (Figure 12). A practical workflow is to begin with quartile-based green/amber/red thresholds, monitor alert tolerance weekly, and tighten red-zone criteria as acceptance permits. Model calibration ensures that probabilistic outputs remain interpretable (e.g., predicted risks align with observed rates), and robustness checks confirm that the dashboard remains stable under missingness and transfer shifts (Figures 11, 13–15).

Finally, governance must be predictable and transparent: data should be minimized and de-identified at ingest; AI outputs should remain advisory with human-in-the-loop confirmation; bias and drift should be monitored by shift and fleet; and all threshold or prompt changes should be logged and reviewed via routine KPI dashboards. A clear "no-surprises" policy – emphasizing process learning rather than personnel evaluation – is essential. Collectively, this approach provides a scalable, Safety-II-aligned model that can expand from a single hangar to fleet-level operations without over-engineering or eroding user trust.

6.4. Limitations

The single-site design limits outside validity. Cross-shift testing shows that model transfer primarily requires light recalibration, not complete retraining, because core linguistic signals and KPI–outcome links remain stable across different base rates. Transferring to other MROs typically involves recalibrating probabilities and thresholds, with full retraining necessary only if documentation norms, language, or templates substantially change. Deploying across multiple sites with pre-registered phase timing would further enhance generalizability. Causality remains quasi-experimental: interrupted time-series and generalized linear models with placebo checks reduce bias from history, but residual confounding from parallel initiatives, regression to the mean, or seasonal effects can't be fully eliminated; randomized or instrumented natural experiments would provide definitive causal evidence. Lastly, communication KPIs reflect observable behaviors rather than cognitive states. Small-data NLP techniques worked well for brief e-logs, though templating and jargon may overlook some ambiguity. To prevent annotation bias and model drift, the gold set must be expanded, periodically recalibrated, and validated against independent artifacts such as read-back audits.

6.5. Future work

Future studies should validate results across multiple sites and fleets through stepped-wedge or cluster ITS designs to confirm external generalizability. Randomized nudge timing or prompt variants could isolate causal effects while remaining operationally feasible. Tracking model calibration transfer and periodic isotonic updates ensured dashboard reliability across contexts. Next-generation modeling should enhance semantic structure – linking entities, units, time markers, and uncertainty cues in e-logs – to refine KPIs such as *Unitized Specificity* and *Deadline Integrity* and deepen understanding of how language variability drives resilience. Finally, integrating KPI metrics with FRAM and STPA mappings can embed quantitative Safety-II into system-level safety engineering. Each KPI can anchor specific FRAM functions or STPA control actions, creating leading-indicator dashboards aligned with Safety Management Systems (SMS) and proactive risk-assurance frameworks.

7. Conclusions

This study demonstrates that an AI-augmented, privacy-first Safety-II approach can turn everyday maintenance communications into measurable resilience, linking concrete communication KPIs (closure loops, read-backs, completeness, latency, continuity) to fewer defects with minimal friction. Across a phased Baseline→Assist→Nudge rollout, ITS/GLM estimates show immediate reductions in levels and sustained improvements in slopes in sign-off errors. At the same time, mediation confirms that KPI shifts

statistically account for a substantial share of those gains. Model diagnostics (AUROC, Brier) and calibration plots indicate trustworthy risk scoring; robustness checks (placebo cut-points, threshold sensitivity, missingness tolerance) and cross-shift generalization support stability and transfer with light recalibration. Operationally, advisory assists weak signals; light Nudges standardize high-value behaviors without burdening the line; and calibrated thresholds keep alerts actionable. Taken together, the evidence supports a scalable, MRO-ready pathway for embedding quantitative Safety-II into routine handovers: simple structured fields and micro-prompts, calibrated AI assistance, and transparent governance that improves quality and safety without slowing work.

Funding

This research received no external funding and was conducted as part of the author's institutional research activities within the aviation maintenance and safety program.

Disclosure statement

The author declares no known financial or non-financial competing interests that could have influenced the conduct or outcomes of this research.

References

Aherne, D. P., Chatzi, A., Kourousis, K., & Kwakye, O. (2025). Human factors considerations for critical maintenance tasks and their effect on the transition to digital maintenance documentation. *Aviation*, *29*(1), 48–54. https://doi.org/10.3846/aviation.2025.23131

Ahmadi Rad, M., Lefsrud, L. M., & Hendry, M. T. (2023). Application of systems thinking accident analysis methods: A review for railways. *Safety Science*, *160*, Article 106066. https://doi.org/10.1016/j.ssci.2023.106066

Ahn, J., Jang, H., & Son, Y. (2020). Critical care nurses' communication challenges during handovers: A systematic review and qualitative metasynthesis. *Journal of Nursing Management*, *29*(4), 623–634. https://doi.org/10.1111/jonm.13207

Appelbaum, R. D., Puzio, T. J., Bauman, Z., Asfaw, S., Spencer, A., Dumas, R. P., Kaur, K., Cunningham, K. W., Butler, D., Sawhney, J. S., Gadomski, S., Horwood, C. R., Stuever, M., Sapp, A., Gandhi, R., & Freeman, J. (2024). Handoffs and transitions of care: A systematic review, meta-analysis, and practice management guideline from the Eastern Association for the Surgery of Trauma. *Journal of Trauma and Acute Care Surgery*, *97*(2), 305–314. https://doi.org/10.1097/TA.0000000000004285

Bach, S. H., Rodriguez, D., Liu, Y., Luo, C., Shao, H., Xia, C., Sen, S., Ratner, A., Hancock, B., Alborzi, H., Kuchhal, R., Ré, Ch., & Malkin, R. (2019). Snorkel DryBell: A case study in deploying weak supervision at industrial scale. *Proceedings of the VLDB Endowment*, *12*(12), 362–375. https://doi.org/10.1145/3299869.3314036

Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., Fries, J. A., Wornow, M., Swaminathan, A., Lehmann, L. S., Hong, H. J., Kashyap, M., Chaurasia, A. R., Shah, N. R., Singh, K., Tazbaz, T., Milstein, A., Pfeffer, M. A., & Shah, N. H. (2025). Testing and evaluation of health care applications of large language models. *JAMA*, *333*(4), 319–328. https://doi.org/10.1001/jama.2024.21700

Bickley, S. J., & Torgler, B. (2021). A systematic approach to public health – novel application of the human factors analysis and classification system to public health and COVID-19. *Safety Science*, *140*, Article 105312. https://doi.org/10.1016/j.ssci.2021.105312

Bukoh, M. X., & Siah, C. R. (2020). A systematic review on the structured handover interventions between nurses in improving patient safety outcomes. *Journal of Nursing Management*, *28*(3), 744–755. https://doi.org/10.1111/jonm.12936

Chatzi, A. V., & Kourousis, K. I. (2024). Identifying the contribution of communication and trust in aviation maintenance occurrences: A content analysis methodology. *Transportation Research Interdisciplinary Perspectives*, *27*, Article 101220. https://doi.org/10.1016/j.trip.2024.101220

Chatzi, A. V., Martin, W., Bates, P., & Murray, P. (2019). The unexplored link between communication and trust in aviation maintenance practice. *Aerospace*, *6*(6), Article 66. https://doi.org/10.3390/aerospace6060066

Choi, J. Y., Byun, M., & Kim, E. J. (2024). Educational interventions for improving nursing shift handovers: A systematic review. *Nurse Education in Practice*, *74*, Article 103846. https://doi.org/10.1016/j.nepr.2023.103846

Cohen, A., Lanson, A., Kempf, E., & Tannier, X. (2024). Leveraging information redundancy of real-world data through distant supervision. In *Proceedings of LREC-COLING 2024* (pp. 10353–10363). European Language Resources Association. https://tinyurl.com/5dkua43c

Delardes, B., McLeod, L., Chakraborty, S., & Bowles, K.-A. (2020). What is the effect of electronic clinical handovers on patient outcomes? A systematic review. *Health Informatics Journal*, *26*(4), 2422–2434. https://doi.org/10.1177/1460458220905162

Delikhoon, M., Zarei, E., Banda, O. V., Faridan, M., & Habibi, E. (2022). Systems thinking accident analysis models: A systematic review for sustainable safety management. *Sustainability*, *14*(10), Article 5869. https://doi.org/10.3390/su14105869

Dhrangadhariya, A., & Müller, H. (2023). Not so weak PICO: Leveraging weak supervision for participants, interventions, and outcomes recognition for systematic review automation. *JAMIA Open*, *6*(1). https://doi.org/10.1093/jamiaopen/ooac107

Dong, T., Yang, Q., Ebadi, N., Luo, X. R., & Rad, P. (2021). Identifying incident causal factors to improve aviation transportation safety: Proposing a deep learning approach. *Journal of Advanced Transportation*, *2021*, 1–15. https://doi.org/10.1155/2021/5540046

Ewertowski, T., & Kowalska, A. (2025). The impact of improving the safety management system on situational awareness in the context of safety: A case of a selected high-reliability organization. *Journal of Management and Financial Sciences*, *54*, 107–123. https://doi.org/10.33119/JMFS.2024.54.6

Federal Aviation Administration. (2023). *Human factors in aviation maintenance*: *Dirty dozen*. https://tinyurl.com/4na9y8by

Griffioen, J., van der Drift, M., & van den Broek, H. (2021). Enhancing maritime crew resource management training by applying resilience engineering: A case study of the bachelor maritime officer training programme in Rotterdam. *Education Sciences*, *11*(8), Article 378. https://doi.org/10.3390/educsci11080378

Grindley, B., Parnell, K. J., Cherett, T., Scanlan, J., & Plant, K. L. (2024). Understanding the human factors challenge of handover between levels of automation for uncrewed air systems: A systematic literature review. *Transportation Planning and Technology*, *48*(6), 1383–1408. https://doi.org/10.1080/03081060.2024.2375645

Ham, D.-H. (2021). Safety-II and resilience engineering in a nutshell: An introductory guide to their concepts and methods. *Safety and Health at Work*, *12*(1), 10–19. https://doi.org/10.1016/j.shaw.2020.11.004

Hölzing, C. R., Rumpf, S., Huber, S., Papenfuß, N., Meybohm, P., & Happel, O. (2024). The potential of using generative AI/NLP to identify and analyse critical incidents in a critical incident reporting system (CIRS): A feasibility case–control study. *Healthcare*, *12*(19), Article 1964. https://doi.org/10.3390/healthcare12191964

Huang, Y., Li, W., Macheret, F., & Gabriel, R. A., & Ohno-Machado, L. (2020). A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, *27*(4), 621–633. https://doi.org/10.1093/jamia/ocz228

Iflaifel, M., Lim, R. H., Ryan, K., & Crowley, C. (2020). Resilient Health Care: A systematic review of conceptualisations, study methods and factors that develop resilience. *BMC Health Services Research*, *20*, Article 324. https://doi.org/10.1186/s12913-020-05208-3

Islam, S., Alfred, M., Wilson, D., & Cohen, E. (2024). Evaluating active learning strategies for automated classification of patient safety event reports in hospitals. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *68*(1). https://doi.org/10.1177/10711813241260676

Janes, G., Harrison, R., Johnson, J., Simms-Ellis, R., Mills, T., & Lawton, R. (2020). Multiple meanings of resilience: Health professionals' experiences of a dual element training intervention designed to help them prepare for coping with error. *Authorea, Inc*. https://doi.org/10.22541/au.160829796.66009765/v1

Judy, G. D., Lindsay, D. P., Gu, D., Mullins, B. T., Mosaly, P. R., Marks, L. B., Chera, B. S., & Mazur, L. M. (2020). Incorporating human factors analysis and classification system (HFACS) into analysis of reported near misses and incidents in radiation oncology. *Practical Radiation Oncology*, *10*(5), e312–e321. https://doi.org/10.1016/j.prro.2019.09.005

Karakiliç, E., Gunaltili, E., Ekici, S., Dalkiran, A., Balli, O., & Karakoc, T. H. (2023). A comparative study between paper and paperless aircraft maintenance: A case study. *Sustainability*, *15*(20), Article 15150. https://doi.org/10.3390/su152015150

Keshtkar, L., Bennett-Weston, A., Khan, A. S., Mohan, S., Jones, M., Nockels, K., Gunn, S., Armstrong, N., Bostock, J., & Howick, J. (2025). Impacts of communication type and quality on patient safety incidents: A systematic review. *Annals of Internal Medicine*, *178*(5), 687–700. https://doi.org/10.7326/ANNALS-24-02904

Lazzari, C., & Rabottini, M. (2025). The use of introduction, situation, background, assessment, and recommendation handover in the COVID-19 pandemic and non-COVID clinical settings: A systematic review and meta-analysis. *Frontiers in Health Services*, *5*. https://doi.org/10.3389/frhs.2025.1380948

Luther, B., Gunawan, I., & Nguyen, N. (2023). Identifying effective risk management frameworks for complex socio-technical systems. *Safety Science*, *158*, Article 105989. https://doi.org/10.1016/j.ssci.2022.105989

Lyu, T., Song, W., & Du, K. (2019). Human factors analysis of air traffic safety based on HFACS-BN model. *Applied Sciences*, *9*(23), Article 5049. https://doi.org/10.3390/app9235049

Ma, Z., & Chen, Z. (2024). Mining construction accident reports via unsupervised NLP and Accimap for systemic risk analysis. *Automation in Construction*, *161*, Article 105343. https://doi.org/10.1016/j.autcon.2024.105343

McGill, A., Smith, D., McCloskey, R., Morris, P., Goudreau, A., & Veitch, B. (2021). The functional resonance analysis method as a health care research methodology: A scoping review. *JBI Evidence Synthesis*, *20*(4), 1074–1097. https://doi.org/10.11124/JBIES-21-00099

Metso, L., Baglee, D., & Marttonen-Arola, S. (2018). Maintenance as a combination of intelligent IT systems and strategies: A literature review. *Management and Production Engineering Review*, *9*(1), 51–64.

Miyamoto, A., Bendarkar, M. V., & Mavris, D. N. (2022). Natural language processing of aviation safety reports to identify inefficient operational patterns. *Aerospace*, *9*(8), Article 450. https://doi.org/10.3390/aerospace9080450

Muecklich, N., Sikora, I., Paraskevas, A., & Padhra, A. (2023). The role of human factors in aviation ground operation-related accidents/incidents: A human error analysis approach. *Transportation Engineering*, *13*, Article 100184. https://doi.org/10.1016/j.treng.2023.100184

Naik, A., Lehman, J., & Rosé, C. (2022). Adapting to the long tail: A meta-analysis of transfer learning research for language understanding tasks. *Transactions of the Association for Computational Linguistics*, *10*, 956–980. https://doi.org/10.1162/tacl_a_00500

Newman, M., & Scott, S. (2023). It was this wing wasn't it? Identifying the importance of verbal communication in aviation maintenance. *The International Journal of Aerospace Psychology*, *33*(2), 139–152. https://doi.org/10.1080/24721840.2023.2169146

Nikolić, M., Nikolić, D., Stefanović, M., Koprivica, S., & Stefanović, D. (2025). Mitigating algorithmic bias through probability calibration: A case study on lead generation data. *Mathematics*, *13*(13), Article 2183. https://doi.org/10.3390/math13132183

Patriarca, R., Di Gravio, G., Woltjer, R., Costantino, F., Praetorius, G., Ferreira, P., & Hollnagel, E. (2020). Framing the FRAM: A literature review on the functional resonance analysis method. *Safety Science*, *129*, Article 104827. https://doi.org/10.1016/j.ssci.2020.104827

Poller, D. N., Bongiovanni, M., Cochand-Priollet, B., Johnson, S. J., & Perez-Machado, M. (2020). A human factor event-based learning assessment tool for assessment of errors and diagnostic accuracy in histopathology and cytopathology. *Journal of Clinical Pathology*, *73*(10), 681–685. https://doi.org/10.1136/jclinpath-2020-206538

Provan, D. J., Woods, D. D., Dekker, S. W. A., & Rae, A. J. (2020). Safety-II professionals: How resilience engineering can transform safety practice. *Reliability Engineering & System Safety*, *195*, Article 106740. https://doi.org/10.1016/j.ress.2019.106740

Ranasinghe, U., Jefferies, M., Davis, P., & Pillay, M. (2020). Resilience engineering indicators and safety management: A systematic review. *Safety and Health at Work*, *11*(2), 127–135. https://doi.org/10.1016/j.shaw.2020.03.009

Ricketts, J., Barry, D., Guo, W., & Pelham, J. (2023). A scoping literature review of natural language processing application to safety occurrence reports. *Safety*, *9*(2), Article 22. https://doi.org/10.3390/safety9020022

Sarvari, H., Edwards, D. J., Rillie, I., & Posillico, J. J. (2024). Building a safer future: Analysis of studies on safety I and safety II in the construction industry. *Safety Science*, *178*, Article 106621. https://doi.org/10.1016/j.ssci.2024.106621

Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., & Flach, P. (2023). Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Machine Learning*, *112*, 3211–3260. https://doi.org/10.1007/s10994-023-06336-7

Steinmann, P., & Tobi, H., & van Voorn, G. A. K. (2024). Resilience metrics for socio-ecological and socio-technical systems: A

scoping review. *Systems*, *12*(9), Article 357. https://doi.org/10.3390/systems12090357

Thomas, J. (2019). *Introduction to STPA* [PowerPoint slides). MIT Partnership for Systems Approaches to Safety and Security. https://tinyurl.com/3fr8588h

Wicaksono, F. D., Ciptomulyono, U., Artana, K. B., & Irawan, M. I. (2021). A state of the art of the accident causation models in the process industries. *Process Safety Progress*, *41*(1), 167–176. https://doi.org/10.1002/prs.12283

Xing, Y., Wu, Y., Zhang, S., Wang, L., Cui, H., Jia, B., & Wang, H. (2024). Discovering latent themes in aviation safety reports using text mining and network analytics. *International Journal of Transportation Science and Technology*, *16*, 292–316. https://doi.org/10.1016/j.ijtst.2024.02.009

Yang, C., & Huang, C. (2023). Natural Language Processing (NLP) in aviation safety: Systematic review of research and outlook into the future. *Aerospace*, *10*(7), Article 600. https://doi.org/10.3390/aerospace10070600

Young, I. J. B., Luz, S., & Lone, N. (2019). A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *International Journal of Medical Informatics*, *132*, Article 103971. https://doi.org/10.1016/j.ijmedinf.2019.103971

Yuzui, T., & Kaneko, F. (2025). Toward a hybrid approach for the risk analysis of maritime autonomous surface ships: A systematic review. *Journal of Marine Science and Technology*, *30*(1), 153–176. https://doi.org/10.1007/s00773-024-01040-0

Zarei, E., Khan, F., & Abbassi, R. (2023). How to account artificial intelligence in human factor analysis of complex systems? *Process Safety and Environmental Protection*, *171*, 736–750. https://doi.org/10.1016/j.psep.2023.01.067

Zhang, Y., Dong, C., Guo, W., Dai, J., & Zhao, Z. (2022). Systems theoretic accident model and process (STAMP): A literature review. *Safety Science*, *152*, Article 105596. https://doi.org/10.1016/j.ssci.2021.105596