International Journal of Industrial Ergonomics 113 (2026) 103918

Contents lists available at ScienceDirect

International Journal of Industrial Ergonomics

journal homepage: www.elsevier.com/locate/ergon

# Data-driven human error accident risk control: A systematic review of data sparsity, causal explainability, and operational transformation

Chongfeng Li [a], Shenghan Zhou [a], Xing Pan [a,*], Song Ding [b], Ziyao Li [a]

[a] School of Reliability and Systems Engineering, Beihang University, Haidian, Beijing, 100191, China
[b] School of Business Administration, Northeastern University, Shenyang, 110169, China

ARTICLE INFO

*Keywords:*
Human error accidents
Data-driven
Risk control
Data sparsity
Explainability
Systematic review

ABSTRACT

With the increasing complexity of modern technological systems, data-driven human error accident risk control (DD-HEARC) faces three interrelated core challenges: data sparsity (Q1), explainability (Q2), and transformation application (Q3), which form a vicious cycle that hinders scientific progress. Following the PRISMA guidelines, this study systematically reviewed 95 high-quality journal articles from Web of Science, Scopus, and IEEE Xplore. The analysis reveals severe fragmentation: approximately 15% of studies address two or more challenges concurrently, and only 21% explicitly incorporate causal inference. Four key limitations are identified: causal blindness in data augmentation, disconnection between explainable AI and safety practice, lack of systematic transformation mechanisms, and absence of cross-dimensional integration methodologies. The study's core contribution is the systematic diagnosis of the Q1→Q2→Q3→Q1 constraint loop and the proposal of an innovative three-dimensional DD-HEARC framework integrating temporal, logical, and collaborative dimensions. This framework provides a unified cognitive tool for researchers and practitioners to shift from fragmented, experience-driven approaches toward systematic, science-driven risk governance.

## 1. Introduction

Human error accidents (HEA) refers to unexpected events primarily caused by human action failures that result in adverse consequences (Donaldson et al., 2000). In the era of Industry 5.0 and digital intelligence, HEA remain a critical threat to system resilience. Recent ergonomics research has shifted from retrospective statistics to proactive, real-time sensing, such as using electrodermal activity (EDA) for workload assessment (Seong et al., 2025) and non-contact cardiopulmonary features for cognitive overload risk detection (Sen et al., 2026). From traditional industrial fields (Garrett and Teizer, 2009) and transportation sectors (Ashraf et al., 2024; Li et al., 2023), to modern information and communication technology (Nobles, 2018) and other domains, HEA occur frequently (Kyriakidis et al., 2019; Maternová et al., 2023). Meanwhile, new types of human error emerging from advanced technologies present new challenges for HEA analysis methods (Shin and Shin, 2023).

The HEA generation process follows the causal relationship of "PSFs → Human Error → HEA" (Reason, 1990), where Performance Shaping Factors (PSFs) induce human errors, and human errors directly or indirectly trigger accidents through system coupling and propagation (Hollnagel, 2014). This causal chain has been widely adopted in human reliability analysis (HRA) frameworks such as CREAM (Hollnagel, 1998), ATHEANA (Forester et al., 2004), and the more recent STAMP/STPA models (Leveson, 2011), which emphasize systemic interactions over linear failure sequences. While traditional HRA relies on expert judgment, the increasing complexity of human-machine systems demands objective causal evidence. Recent advances in AI-driven forensics, such as HFACS-LLM reasoning (Li and Wang, 2025), provide new pathways for accident reconstruction. However, bridging the gap between high-dimensional sensor data and explainable management actions remains a fundamental challenge. To achieve precise risk control, it is essential to identify true causal mechanisms, namely clarifying which PSFs can substantially reduce accident risk when changed, and through controlling which links accident occurrence can be prevented.

With the rapid development of technologies such as the Internet of Things, big data, and artificial intelligence, modern industrial systems generate massive operational data, providing unprecedented opportunities for deep understanding of HEA causation mechanisms. Data-driven analysis methods can automatically discover patterns, extract knowledge, and build models from large amounts of actual data, offering significant advantages in objectivity, real-time adaptability, and

* Corresponding author.
*E-mail address:* panxing@buaa.edu.cn (X. Pan).

https://doi.org/10.1016/j.ergon.2026.103918
Received 16 November 2025; Received in revised form 15 February 2026; Accepted 22 February 2026
Available online 3 March 2026

precision over traditional methods. Recent studies have demonstrated the potential of machine learning in predicting human performance degradation (Zhang et al., 2026), identifying latent PSFs from unstructured incident reports (Tanguy et al., 2016), and integrating real-time physiological data into dynamic risk assessment. Based on the complex system safety management needs driven by digitization trends, this study defines the research concept of data-driven human error accident risk control (DD-HEARC). As a specialized, human-centric extension of the broader data-driven risk analysis (DRA) paradigm, DD-HEARC emphasizes the human factor—the principal source of variability and uncertainty that traditional hardware-oriented DRA approaches often oversimplify. By fully utilizing rich data resources in the digital age, DD-HEARC seeks to achieve a transformation from experience-driven to science-driven governance, providing a targeted methodological framework for managing the dynamic and non-linear nature of human fallibility in Industry 5.0.

Despite its potential, practical applications of DD-HEARC face more stringent constraints than general DRA due to the inherent complexity of human behavior. It is currently impeded by a triadic bottleneck consisting of three core challenges, as shown in Fig. 1.

(1) Q1: Data Sparsity

Unlike mechanical failures with detectable physical signals, human-centric data suffers from the unobservability of PSFs, where critical cognitive states lack objective measurement benchmarks (Anjum and Rocca, 2019). Furthermore, label inconsistency remains a significant hurdle, as causation judgments for the same accident often diverge among experts. Finally, HEA samples are extremely scarce, resulting in a paradox of safety: stringent safety protocols in high-reliability organizations effectively suppress incident frequency, thereby limiting the availability of historical data required for deep learning. This data sparsity increases the risk of model overfitting and undermines generalization.

(2) Q2: Causality Explainability

The second challenge concerns the requirement for explainability, which stems from the epistemological tension between the complexity of safety management demands and the opaque nature of data-driven models. While general DRA often prioritizes predictive performance, DD-HEARC necessitates a high level of managerial transparency to align AI logic with the operator's situational awareness and the organization's responsibility-based protocols. The inherent opacity of advanced machine learning prevents safety managers from identifying the mechanistic "why" and "how" behind human error, which is a prerequisite for justifiable intervention (Hong et al., 2020). Although emerging frameworks such as causal inference (Pearl, 2019) and hybrid symbolic-AI (Mehra, 2024) attempt to bridge this gap, the persistent associative trap of current models continues to erode practitioner trust and hampers the formulation of logic-driven prevention measures.

(3) Q3: Risk control method.

Even when accurate and explainable insights are achieved, a critical transformation gap persists in converting scientific cognition into actionable risk control strategies (Anjum and Rocca, 2019). Within the specialized scope of DD-HEARC, this represents the final hurdle in translating data-driven intelligence into measurable safety value—a process significantly more complex than traditional DRA interventions due to the dynamic and non-linear nature of human-machine collaboration. This gap manifests as a systemic decoupling where analytical insights remain frozen in research reports rather than being thawed into real-time, proactive interventions. Bridging this gap requires achieving cybernetic closure, ensuring that data-driven insights are seamlessly integrated into adaptive control loops that can accommodate the transient fluctuations of human reliability in modern industrial environments.

Building upon the identified challenges, this study aims to systematically synthesize the current landscape of HEA research through a PRISMA-compliant literature review. The primary objective is to navigate the complexities of sparse data processing and model explainability to establish a robust DD-HEARC implementation framework. Specifically, this research is structured to: 1) construct a problem: oriented framework following the causal evolution of "PSFs → Human Error → HEA"; 2) conduct a critical review of the triadic bottleneck, namely data scarcity, explainability, and risk control transformation; 3) diagnose systemic research limitations and delineate future strategic directions; and 4) propose a comprehensive 3D analysis framework to provide methodological guidance for both theoretical development and engineering applications.

The structure of this paper is organized into seven interconnected sections to ensure a seamless transition from theoretical grounding to practical synthesis. Following the introduction, Section 2 establishes the multi-disciplinary theoretical foundations supporting the DD-HEARC paradigm, while Section 3 delineates the systematic review methodology. Section 4 provides a tri-dimensional analysis of the current research status, which informs the identification of critical gaps and future directions in Section 5. Subsequently, Section 6 constructs a conceptual application reference framework based on the review findings. Finally, Section 7 summarizes the study's core contributions and its broader significance for the field of ergonomics.

The theoretical significance of this research resides in its systematic development of a comprehensive conceptual landscape for DD-HEARC,

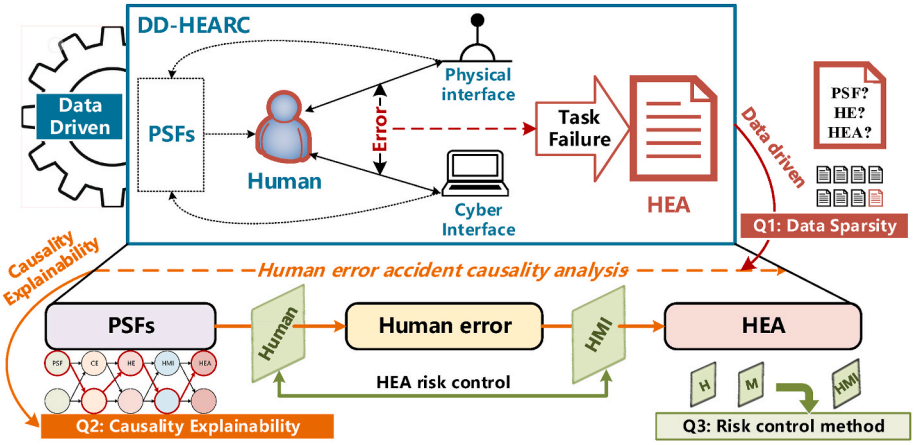


**Fig. 1.** DD-HEARC Concept and Problems to be Solved.

specifically elucidating the reciprocal dependencies between sparse data processing, causal explainability, and risk control transformation. By identifying the underlying structural pathologies within the field and introducing an innovative 3D analysis framework, this study provides a robust methodological foundation for systematic problem-solving in human factor safety. Practically, this work functions as a strategic decision-support instrument for researchers in methodology selection while offering a standardized technical roadmap for engineering practitioners.

## 2. DD-HEARC theoretical foundations

The emergence of DD-HEARC represents a paradigm shift from traditional retrospective human factors analysis to a proactive, closed-loop control system within the Industry 5.0 landscape (Belkadi and Bachiri, 2025). Theoretically, DD-HEARC synthesizes Systems Theory, Cybernetics, and Cognitive Engineering to establish a continuous operational flow of "Sensing (Q1)→Cognition (Q2)→Action (Q3)". This section elucidates the theoretical pillars that support this integrated approach in addressing the systemic challenges of data sparsity, causal opacity, and transformation gaps.

### 2.1. Theoretical pillars for DD-HEARC

DD-HEARC is grounded in a hierarchical theoretical framework that transitions from macro-systemic logic to micro-behavioral mechanisms. First, Cybernetics and Systems Thinking provide the foundational logic for DD-HEARC as a proactive safety risk control system (Chen et al., 2025). In the DD-HEARC paradigm, the "PSFs→Human Error→HEA" chain (Tang et al., 2020) serves as the causal skeleton. These skeleton transforms raw sensing data (PSFs) into mediating variables (Human Errors) and final outcomes, enabling the system to act as a feed-forward control loop where risks are mitigated before they propagate into accidents. Second, the transition from Safety-I to Safety-II provides the epistemological justification for data-driven modeling (Chen et al., 2021; Hollnagel, 2014). While traditional HRA focuses on scarce accident samples, DD-HEARC embraces the Safety-II philosophy by monitoring normal operational variability. Third, Cognitive Systems Engineering and Situational Awareness (SA) theoretical frameworks provide the bridge for model explainability (Q2). For DD-HEARC to be effective, its outputs must align with the human manager's mental model. Integrating frameworks like SAFE-AI (Sanneman and Shah, 2022) ensures that opaque model predictions are translated into transparent, causally interpretable insights, thereby facilitating the translation of scientific understanding into actionable risk control strategies (Q3). This integration ensures that intelligent forensics supplements, rather than replaces, expert judgment in complex human-machine collaborations (Zhang et al., 2026).

### 2.2. Theoretical support for Q1

In the data-driven modeling of the "PSFs → Human Error → HEA" causal chain, the primary challenge is the data sparsity. Data sparsity in the safety domain has unique characteristics, mainly manifested in three levels of problems, as shown in Fig. 2. First is PSFs unobservability: PSFs such as operators' psychological states, organizational safety culture, and informal communication patterns often lack objective measurement methods (Moura et al., 2016), stemming from the subjective and hidden characteristics of human factor variables. Second is data label quality issues: stemming from high cost and low consistency of expert annotation, different experts may have significant differences in causation judgments for the same HEA, affecting the application effectiveness of data-driven methods (Zhou et al., 2022). Third is accident sample scarcity: major HEA, as typical low-frequency high-consequence events, have extremely limited available samples for analysis due to their rarity (Parsa et al., 2019), causing serious class imbalance phenomena (Mujalli et al., 2016).

Schematic illustration of Q1 (data sparsity) manifestations along the HEA causal path "PSFs → Cognitive Error (CE) → Human Error (HE) → Human-Machine Interface (HMI) → HEA". Dotted lines indicate missing or low-quality data at three levels: ① PSFs unobservability (e.g., psychological states); ② label inconsistency due to expert subjectivity; ③ accident sample scarcity leading to class imbalance.

Addressing Q1, relevant theories provide multi-level solutions. Causal inference theory provides theoretical foundation for identifying true causal relationships under sparse data conditions through structural causal models, potential outcome frameworks, and other methods (Pearl, 2010). Small sample learning theory includes meta-learning, few-shot learning, and other methods, providing technical support for building effective prediction models under limited sample conditions (Gharoun et al., 2023). Data augmentation theory generates synthetic samples conforming to causal logic through deep learning techniques such as generative adversarial networks and variational autoencoders, combined with domain knowledge constraints (Fathy et al., 2020). Transfer learning theory improves model performance in target domains by utilizing data resources from related domains through cross-domain knowledge transfer (Chen et al., 2022). Meanwhile, professional knowledge in the HRA field provides important guidance for sparse data processing. In PSF data acquisition, HRA theory provides basis for variable classification, measurement methods, and relationship analysis (La Fata et al., 2023; Morais et al., 2022). In human error label formulation, classic skill-rule-knowledge classification (Rasmussen, 1990) provides theoretical framework for annotation systems. This combination of domain knowledge and data science technology is an important

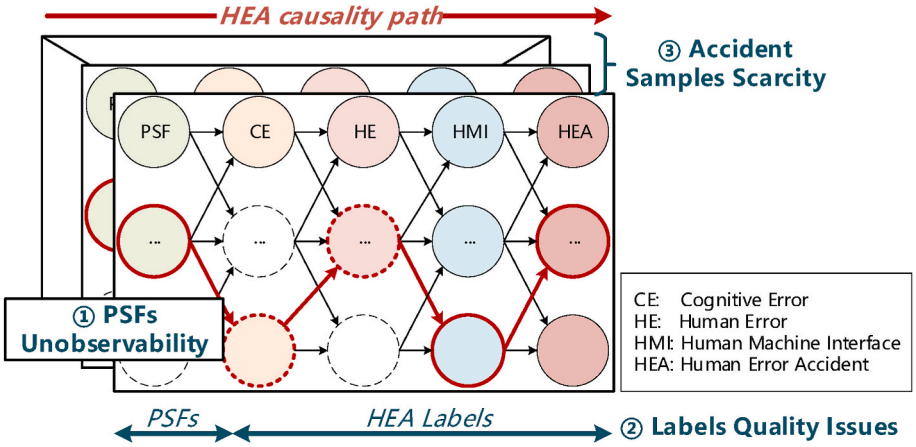


**Fig. 2.** Specific manifestations of Q1.

characteristic of DD-HEARC sparse data processing.

### 2.3. Theoretical support for Q2

After solving Q1, DD-HEARC must also ensure model explainability, which is not only a technical requirement but also an essential need of safety management (Hong et al., 2020). From a cognitive science perspective, safety managers need to understand HEA causal mechanisms to formulate effective prevention strategies. From a decision science perspective, risk control decisions involve significant safety responsibilities and resource investments, requiring explainable scientific basis; from a system engineering perspective, safety management of complex systems requires transparent causal logic to guide multi-level, multi-dimensional intervention measures. Based on the depth and mechanism of explainability, Q2 can be divided into extrinsic explainability and intrinsic explainability, as shown in Fig. 3.

Two paradigms of Q2 (explainability). Extrinsic explainability ($f_1$) treats the model as a black box and uses post-hoc methods (e.g., SHAP) to approximate associations between PSFs and HEA. Intrinsic explainability ($f_2$) embeds causal logic within a white-box structure (e.g., Bayesian networks) to reveal the internal mechanism from PSFs through HE to HEA.

Extrinsic explainability treats the HEA causation process as an opaque system, focusing on identifying statistical associations between PSFs and HEAs. Intrinsic explanation methods, in contrast, provide the theoretical underpinning for extrinsic explainability by offering mechanisms to enhance the transparency of complex models (Ribeiro et al., 2016). Although these methods can provide explanations for complex models, the fidelity and stability of explanations are controversial, making it difficult to reveal deep causal mechanisms (Barbierato and Gatti, 2024). Intrinsic explainability conceptualizes the HEA causation process as a transparent system, prioritizing the elucidation of internal mechanisms across the entire causal trajectory—from PSFs and human error to the resulting HEA. By integrating causal inference theory, the DD-HEARC framework transcends mere statistical association to facilitate a rigorous mechanistic understanding. Unlike conventional correlation-based analyses, causal inference enables the evaluation of counterfactual scenarios. This predictive capacity is essential for the design of proactive risk-control strategies. Furthermore, formalisms such as causal graph models and potential outcome frameworks provide intuitive representations of these complex interactions, simultaneously enhancing model interpretability and operational utility (Spirtes and Zhang, 2016).

### 2.4. Theoretical support for Q3

With explainable causal insights, the key issue turns to how to transform these scientific cognitions into actionable management actions. This transformation process involves decision science, human factors engineering, organizational behavior, and other theoretical fields, requiring systematic theoretical support to guide transformation from analysis to action. This study categorizes risk control measures into three types of risk control methods centered on humans, systems, and human-machine collaboration, as shown in Fig. 4.

Human-centered control relies on matching qualified individuals with appropriate operational tasks, focusing on mitigating human error by enhancing personnel capabilities. Because human errors arise from gaps between individual skills and task requirements, this strategy systematically strengthens safety performance through competency frameworks applied to selection and training (Rahman et al., 2022), learning theory for skill development (Vinodkumar and Bhasi, 2010), and motivation theory for behavioral transformation (Christian et al., 2009). System-centered control adheres to "making systems adapt to human errors," acknowledging objective limitations in human cognitive capabilities and focusing on optimizing system design to eliminate error conditions. This approach represents a shift from eliminating human errors to designing error-tolerant systems through human factors engineering for interface optimization (Larsson and Tingvall, 2013), fault-tolerant design for safe state maintenance (Wood and Kieras, 2002), and error-proofing design through physical constraints (Norman, 2013). Human-machine collaboration-centered control transcends traditional unidirectional adaptation, emphasizing bidirectional learning and mutual adjustment between humans and systems (Xu et al., 2023). This approach recognizes that optimal safety performance comes from synergistic effects rather than simple addition of human and machine performance, supported by symbiotic intelligence theory for collaborative development (Xue et al., 2024), adaptive system theory for dynamic optimization (Cao et al., 2012), and trust calibration theory for stable collaborative relationships (Ding et al., 2025b). These three theoretical modules organically unify to constitute the complete risk control transformation framework for DD-HEARC.

Finally, the above four theoretical modules support each other and organically unify, jointly constituting the complete theoretical system of DD-HEARC. Accident causation theory provides core analytical framework and causal logic, providing foundation for problem definition and research boundaries of DD-HEARC; sparse data processing theory solves data acquisition and quality assurance problems, providing guarantee for DD-HEARC data foundation; causal explainability theory ensures scientific and understandable analysis results, providing support for DD-HEARC core capabilities; risk control transformation theory realizes effective transformation from analysis to action, providing guarantee for DD-HEARC application value.

## 3. Methodology

This systematic review follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement (Page et al., 2021) and established practices from previous systematic reviews (Siddaway et al., 2019). These guidelines help systematize, comprehensify, transparentize, and reproducibilize the review process (Salmon

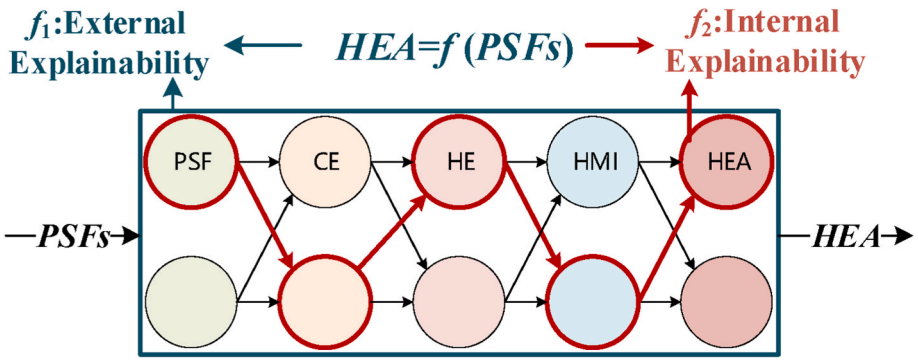


**Fig. 3.** Specific manifestations of Q2.

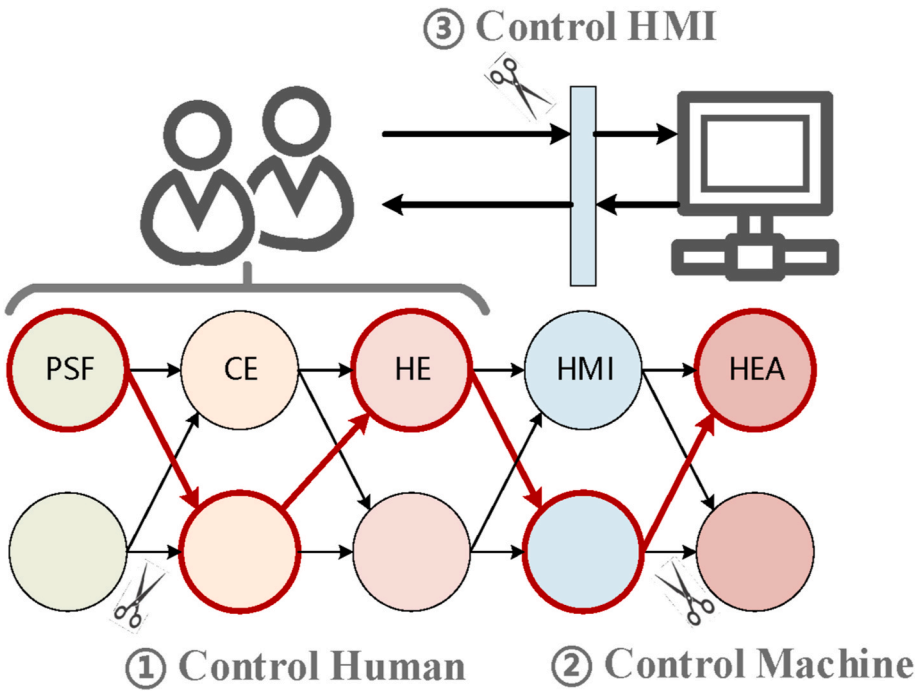


**Fig. 4.** Specific manifestations of Q3.

et al., 2022). Based on the theoretical framework of DD-HEARC and the three core challenges, this study constructed a targeted methodological system to ensure that literature search, screening, and analysis comprehensively cover the three key dimensions of sparse data processing, causal explainability, and risk control transformation.

### 3.1. Search strategy

Based on the theoretical framework established in Chapter 2, the search strategy of this study was specifically designed for the three core challenges to ensure comprehensive capture of relevant research progress. Search terms related to sparse data processing were developed through reviewing mainstream causal inference theories and small sample learning methods. Explainability-related search terms were derived from reviews of explainable Artificial Intelligence (AI) and causal explanation methods, particularly focusing on techniques such as causal inference and counterfactual explanation that align with safety management needs. Human error accident risk control-related search terms referenced review articles in the safety science field, with emphasis on transformation mechanisms from analysis to action. To reflect the data-driven characteristics of DD-HEARC, data-driven related search terms were specifically added.

In the specific search process, a combined search strategy was adopted: ((Sparse Data) OR (Causal Explainability) OR (Risk Control Transformation) OR (Data-Driven)) AND (HEA Core) (see Table 1). Web of Science Core Collection, Scopus, and ScienceDirect were selected as literature search sources because these three databases cover a wide range of journals related to the research topic. The initial search was conducted on December 31, 2025, yielding 24,845 results (8286 from Web of Science Core Collection, 10,471 from Scopus, and 6088 from ScienceDirect). The search results were subsequently imported into EndNote for deduplication processing. After removing 10,696 duplicate records, 14149 articles were retained for further screening.

**Table 1**
Search terms.

| Concept | Search Terms |
|---|---|
| Spare Data | "sparse data" OR "limited data" OR "small sample" OR "data scarcity" OR "few shot" OR "low resource" OR "data augmentation" OR "transfer learning" OR "imbalanced data" |
| Explainability | "explainable" OR "interpretable" OR "explainability" OR "interpretability" OR "causal explanation" OR "causal inference" OR "causal discovery" OR "causal model" OR "counterfactual" |
| HEA Risk Control | ("human error" OR "human factor" OR "human failure" OR "operator error" OR "cognitive failure") AND ("risk control" OR "intervention" OR "prevention" OR "mitigation" OR "safety management") |
| Data Driven | "data-driven" OR "machine learning" OR "artificial intelligence" OR "predictive model" OR "automated analysis" |

### 3.2. Screening and selection

To ensure literature quality and research relevance, this study adopted a hierarchical screening strategy based on the DD-HEARC theoretical framework. To ensure review quality, only peer-reviewed journal articles were included, excluding conference papers and book chapters. Subsequently, a two-stage review strategy was adopted: title and abstract review, and full-text review.

Inclusion criteria required literature to meet at least one of the following conditions: Q1-related research that proposes processing methods or technologies for sparse data problems in human factor accident analysis; Q2-related research that focuses on the explainability of human factor accident analysis models or causal inference methods; Q3-related research that explores transformation mechanisms or strategies from human factor accident analysis to risk control; comprehensive research that simultaneously involves two or more challenges. Exclusion criteria included: literature with research fields unrelated to human factor accidents or safety management; purely theoretical research lacking methodological innovation or application value; traditional statistical research that does not reflect data-driven characteristics; research that focuses only on statistical associations without considering causal relationships.

Quality assessment criteria were established based on the special requirements of DD-HEARC, including five key dimensions: methodological innovation, causal focus, practicality, data-driven degree, and challenge integration. In the title and abstract review stage, articles focusing on causal-enhanced human factor accident analysis in sparse

data environments or solving explainability and control transformation problems were included. In the full-text review stage, originally published articles involving data-driven human factor accident risk control methods and clearly addressing at least one core challenge were included. Subsequently, through backward and forward citation searches, the references and citations of these articles were manually checked to identify other eligible articles. Fig. 5 shows the PRISMA flow diagram for the screening of DD-HEARC-related articles included in this study.

### 3.3. Data extraction and analysis

Based on the theoretical framework of DD-HEARC, this study constructed a hierarchical data extraction system to systematically analyze the current research status of the three core challenges and their interrelationships. Content analysis method was adopted to systematically analyze the included literature, which is an objective method for describing and quantifying research phenomena that can derive reproducible and valid inferences from large amounts of literature data.

The data extraction framework adopted a three-level coding scheme consistent with the DD-HEARC theoretical framework. First-level coding corresponds to core challenge dimensions (Q1 sparse data processing, Q2 causal explainability, and Q3 risk control transformation). Second-level coding refines specific method types for each core challenge, such as data augmentation techniques, transfer learning methods, and expert knowledge integration for Q1. Third-level coding describes application characteristics, including application domains, data-driven degree, causal attention level, and integration complexity. Special attention was paid to the interrelationships among the three challenges and integrated solutions, including Q1-Q2 cross-analysis of explainability maintenance under sparse data conditions, Q2-Q3 cross-analysis of how explainability supports risk control decision-making, Q1-Q3 cross-analysis of control strategy generation with limited data, and Q1-Q2-Q3 integrated analysis of systematic solutions addressing all three challenges simultaneously.

The research team developed an initial coding scheme based on search terms and theoretical framework. Multiple authors conducted data extraction independently after establishing consistency through training coding. Inter-coder agreement coefficient reached 0.87, indicating high coding reliability. Multiple verification measures were adopted including theoretical consistency testing, expert validation, and cross-validation to ensure coding quality. The analysis strategy focused on identifying patterns, gaps, and opportunities within and across the three challenge dimensions, providing empirical basis for subsequent research gap identification and framework construction.

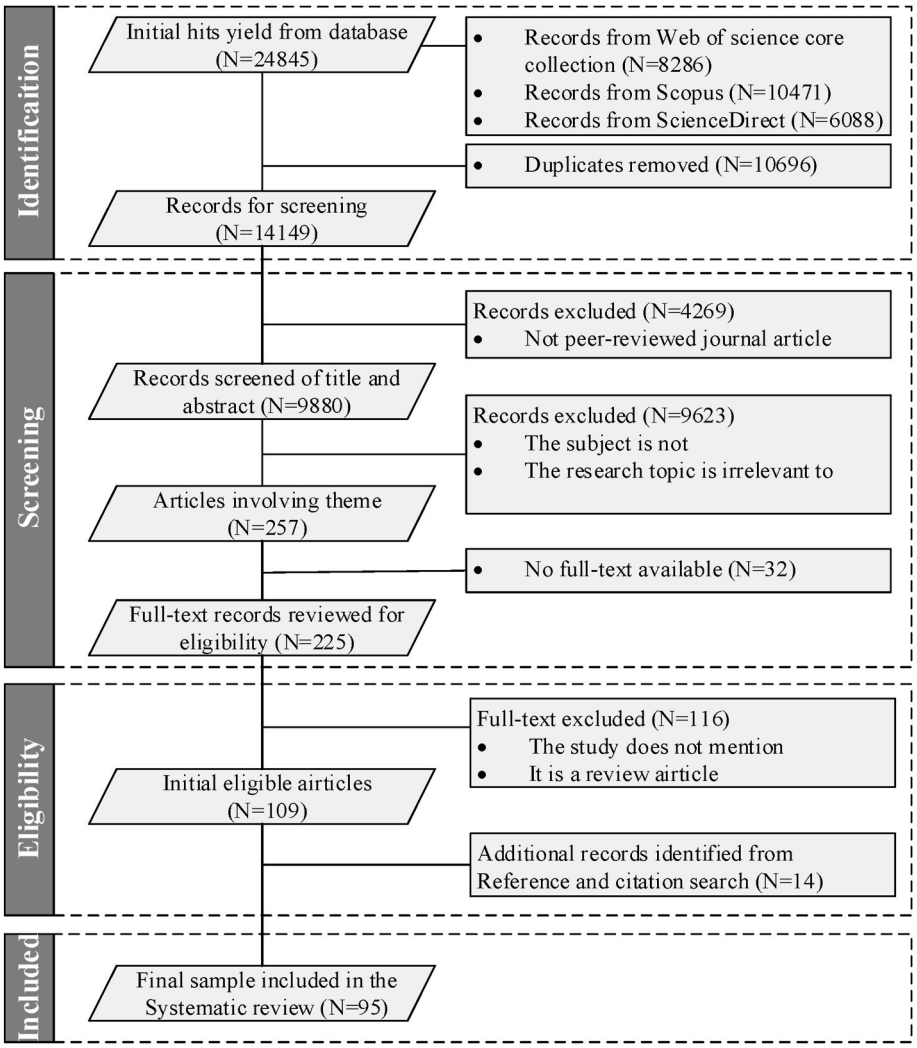


**Fig. 5.** Flow diagram of literature search and selection.

## 4. Results: current state of research

### 4.1. Overcoming the data sparsity bottleneck (Q1)

The systematic review of the selected studies indicates that the challenges of Q1 involve more than limited sample sizes; they represent a systemic crisis in the transition from data to causal evidence. Following the findings presented in Table 2, this section critically analyzes the three levels of data sparsity.

(1) PSFs Data: The Transition from Subjective Snapshots to Continuous Bio-Sensing

Existing research indicates a critical shift in PSFs acquisition. While traditional methods depend on subjective questionnaires, which are often compromised by memory inaccuracies and delayed reporting, recent trends in Table 2 demonstrate a transition toward objective, real-time sensing. This shift overcomes the limitations of passive data collection but introduces the Invasiveness-Validity Paradox: although high-fidelity physiological sensors provide objective metrics, their deployment in high-stress environments can function as a secondary performance-shaping factor, potentially distorting natural operator behavior. The primary research gap involves balancing the ecological validity of field sensing with the precision of laboratory-grade biometrics.

(2) Human Error Data: The Semantic Gap in Labeling and Annotation

The scarcity of human error data is fundamentally a problem of annotation subjectivity. Although databases like HuREX and SACADA provide a foundation, the inconsistency in expert judgments across different industries creates a semantic gap. Most studies treat human error as a binary static label. However, from a Safety-II perspective, human error is not a failure but a performance variability. The failure of current data-driven models to capture this variability stems from an over-reliance on Safety-I (accident-only) data. To bridge this, future datasets must move toward capturing near-misses and successful adaptations to provide a holistic view of human reliability.

(3) Accident Data: The Conflict between Statistical Quantity and Causal Fidelity

To address the low-frequency high-consequence nature of major HEA, data augmentation remains the primary strategy. However, a deep conflict exists between mathematical expansion and logical authenticity. Traditional oversampling methods like BL-SMOTE or MWMOTE generate data points in a high-dimensional space but often break the temporal causal chain of an accident. Emerging generative approaches attempt to preserve this logic through Large Language Models. The fundamental limitation identified here is the Causal Hallucination in AI-generated samples. While generative models can achieve statistical parity with real datasets, they often fail to maintain the strict physical and psychological constraints required for risk control. If Q1 lacks Causal Fidelity, any subsequent analysis in Q2 will merely be an explanation of noise, ultimately leading to the failure of Q3.

### 4.2. Improving the explainability (Q2)

The systematic review identifies a significant epistemological shift in Q2. As modern systems transition toward Industry 5.0, the requirement for explainability has evolved from simple post-hoc attribution to intrinsic mechanistic understanding. Based on the trends in Table 3, this section critiques the Explainability-Trust Gap through two dimensions (see Table 4).

(1) Extrinsic Explainability: Limitations of Post-hoc Interpretations

**Table 2**
Studies addressing Q1.

| Q1 | Problem | References |
|---|---|---|
| **PSFs Data Scarcity** | **PSFs Classification** | ■ ***Nuclear***: Emphasizing interrelationships and risk accumulation over time (Kim and Jung, 2003).<br>■ ***Industrial Maintenance***: Includes organizational, environmental, task, and individual factors affecting human reliability and error (Franciosi et al., 2019).<br>■ ***Chemical Process Industry***: PSFs grouped by task features, error types, and error mechanisms; focus on operator competence, environment, and task (Menezes et al., 2021).<br>■ ***Offshore***: Group factors into human, environmental, organizational, technical, and task-related categories (Wu et al., 2021).<br>■ ***Manufacturing***: Classified by internal/external factors, often tailored to discrete manufacturing (Jha et al., 2024).<br>■ ***Internet/Communication***: Communication, interaction, and social environments (Triplett, 2022) |
| | **PSFs Measurement** | ■ **Questionnaire Measurement**: PSFs questionnaire (Groth and Mosleh, 2009)<br>■ **Experimental Measurement** Cognitive workload: heart rate, facial expressions (Sharma et al., 2020); Attention: eye movement, reaction time (Groth and Mosleh, 2009), fNIRS (Asgher et al., 2020); Fatigue: HRV (Lu et al., 2022), EEG (Ding et al., 2025a); EDA (Seong et al., 2025); Non-contact (Sen et al., 2026)<br>■**Task Decomposition**: SHERPA-Based (Zhang et al., 2020), IDAC-Based (Chang and Mosleh, 2007) |
| **Human Error Data Scarcity** | **Basic Data Acquisition** | **Database Acquisition: Uses incident reports, dynamic updating**<br>■ Nuclear:*HuREX* (Jung et al., 2020), *SACADA* (Chang et al., 2022)<br>■ Aviation: *NTSB* (Burns and Bonaceto, 2020), *AIDS*<br>■ Offshore: *CORE-DATA* (Basra and Kirwan, 1998)<br>■ Industry: *eMRAS*, *FACTS*, *MHIDAS*, *CAIRS* (Arun et al., 2022)<br>■ Transportation: *CISS*, *GIDAS*, *CEDATU*, *RAIDS* (Bauder et al., 2022)<br>**Simulation Data: Realistic scenario modeling, validation/calibration**<br>■ *Monte Carlo* (Cho et al., 2020)<br>■ *System dynamics* (Emroozi et al., 2024a)<br>■ *Modified CREAM/THERP* (Ramezani et al., 2020; Sun et al., 2012)<br>**Experimental Data: Structured expert input, physiological/psychological tests** |
| | **PSFs Modification** | ■ *AHP–SLIM* (Park and Lee, 2008), *HEART-BWM* (Aliabadi et al., 2022), *CREAM with HIFs* (He et al., 2021)<br>■ Multipliers set based on expert consensus or mapped from prior models, like *SPRA-H Based* (Boring and Blackman, 2007), *CREAM-Based* (Wu et al., 2017) |

(*continued on next page*)

**Table 2** (*continued*)

| Q1 | Problem | References |
|---|---|---|
| | | ■ Use data statistic to estimate multipliers and their uncertainties, like *Historical Data* (Kim et al., 2015), *simulation data* (Griffith and Mahadevan, 2015), *empirical data* (Kim and Park, 2019), *Bayes inference* (Takeda and Kitada, 2024)<br>■ Analyze and weight interdependencies among PSFs, then adjust multipliers accordingly, like *DEMATEL/Fuzzy Bayes Network* (Liu et al., 2022; Xu et al., 2023); *Anchored PSF* (Park, 2024);<br>■ Combine empirical data, expert input, and system dynamics to simulate and optimize HEPs, like *Hybrid Slim* (Zhou and Lei, 2020), *SD* (Emroozi et al., 2024b) |
| | **Error Labels** | ■ **First-generation HRA**: Skill-based, Rule-based, Knowledge-based (Rasmussen, 1990); Slips, Lapses, Mistakes, Violations (Reason, 1990); Omission, Commission (Boring, 2012);<br>■ **Second-generation HRA**: Observation, Interpretation, Planning, Execution (Hollnagel, 1998); Phenotypes, Genotypes (Morais et al., 2022)<br>■ **Third-generation HRA**: "Organization-Individual-Task" chain error classification (Diaconeasa and Mosleh, 2018); Real-time deviations in human-machine interaction (Ding et al., 2025a); |
| **Human Factor Accident Data Scarcity** | **Accident Labels** | ■ incidence, accident with material loss, serious accidents with injures, serious accidents with fatalities and injures (Bird et al., 1996) |
| | **Data Quantity** | **Based on Expert Knowledge**<br>■ *Accident sequences analysis* (Olivares et al., 2018); *Fuzzy method* (Kumar et al., 2020)<br>**Based on Data Augmentation**<br>■ Basic sampling: *BL-SMOTE* (Han et al., 2005); *MWMOTE* (Barua et al., 2012);<br>■ Advanced sampling: *Clustering* (Nekooeimehr and Lai-Yuen, 2016); *Ensemble learning* (Bader-El-Den et al., 2018);<br>*Text-Based Knowledge Extraction* (Tian et al., 2022), *AccidentGPT* (Wang et al., 2024); |

Extrinsic explainability, primarily represented by post-hoc methods such as SHAP and LIME, currently prevails in the literature as a diagnostic layer for complex black-box models. Although these techniques effectively provide feature importance rankings, they frequently remain confined to an associative framework that lacks deep mechanistic insights. While they can identify statistically significant variables—such as operator fatigue or ambient lighting conditions—they fail to elucidate the underlying causal interactions or the temporal evolution of these factors within an accident sequence. The fundamental limitation of these extrinsic approaches is their lack of counterfactual stability; in complex human-machine systems, minor contextual shifts can render such probabilistic explanations invalid or inconsistent.

(2) Intrinsic Explainability: Toward Causal Fidelity and Cognitive Alignment

Beyond post-hoc interpretations, the post-2020 surge in the application of Bayesian Networks and Causal Graphs underscores a growing consensus that intrinsic transparency is the definitive pathway for safety-critical AI systems. This transition toward intrinsic explainability is essentially a pursuit of causal fidelity, achieved by embedding structured domain knowledge, such as the Human Factors Analysis and Classification System (HFACS), directly into the model through probabilistic graphical models. Such integration seeks to bridge the semantic gap between machine outputs and human cognition. However, a significant complexity ceiling remains a major obstacle; as causal networks become more realistic by incorporating the dynamic and non-linear nature of human behavior, computational demands and modeling difficulties escalate exponentially. This challenge frequently necessitates a compromise, leading researchers to revert to oversimplified linear assumptions that may fail to capture the intricate nuances of high-risk operational environments.

(3) Synthesis: The Explainability-Actionability Gap

The fundamental depth of the Q2 challenge resides not in a deficit of algorithmic solutions, but in a persistent misalignment between AI-generated explanations and the requirements of managerial actionability. While current explainable AI (XAI) technologies successfully address the mathematical transparency of models, the DD-HEARC paradigm necessitates a higher level of managerial transparency. This cognitive mismatch stems from a disparity in operational languages: while AI interprets its internal logic through artificial neurons and feature weights, safety management operates within a framework of responsibility and proactive prevention. Such a gap exerts a significant propagation effect that reinforces the identified vicious cycle. When explainability fails to provide actionable insights, practitioners often lose the incentive for high-quality data collection (Q1), causing analytical results to remain confined to the theoretical phase rather than bridging the transformation gap toward practical risk control (Q3).

### 4.3. Bridging the transformation gap (Q3)

Q3 represents the final and most critical stage of the DD-HEARC value chain: the transition from Scientific Cognition to Engineering Action. The review of the selected studies indicates that while risk control concepts have evolved from static standardization to dynamic personalization, a fundamental decoupling crisis persists between analytical insights and operational interventions.

(1) Human-Centered Control: Limitations of Static Paradigms in Dynamic Environments

Traditional human-centered control logic predominantly emphasizes personnel selection and specialized training to mitigate errors at their source. While the data-driven era is facilitating a paradigm shift toward situational matching, the practical effectiveness of this approach is often constrained by the high economic costs of personalized training and the rigid capability boundaries of operators during extreme scenarios. However, our synthesis suggests that in the absence of real-time physiological feedback—a critical gap identified in the Q1 dimension—human-centered control remains largely reactive and probabilistic. Such static models fail to account for the transient fluctuations in human reliability, such as micro-fatigue or sudden cognitive surges, rendering long-term, generalized training insufficient for preventing low-frequency, high-consequence "black swan" human error events.

(2) System-Centered Control: The Ironies of Automation in Rigid Designs

System-centered control emphasizes error-proofing and fault-tolerance, aiming to adapt the operational environment to human

**Table 3**
Studies addressing Q2.

| Q2 | Method | Core Idea | References |
|---|---|---|---|
| Extrinsic Explainability | Based on Statistical Association | Efficiently identify association patterns between PSFs and accident outcomes, providing quantitative basis for risk factor ranking. | ■ *Logistic Regression* (Maternová et al., 2023)<br>■ *LIME* (Visani et al., 2022), *SHAP* (Zhong et al., 2021)<br>■ *Counterfactual explanation* (Linardatos et al., 2020) |
| | Based on Artificial Intelligence | Powerful nonlinear modeling capability, able to handle complex interaction relationships and unstructured text data. | ■ *XGBoost* (Koc and Gurgun, 2022);<br>■ *Neural Network based* (Li et al., 2026; Maynard and Harris, 2022; Zhou and Guo, 2024);<br>■ *Attention mechanism* (Jiang et al., 2025; Ras et al., 2022)<br>■ *Agents* (Yu et al., 2025) |
| Intrinsic Explainability | Based on Rule Reasoning | Generate intuitive and understandable "if-then" decision rules, convenient for safety management personnel to understand and apply. | ■ *Rule base* (Cheng et al., 2022)<br>■ *Decision tree* (Abellán et al., 2013), |
| | Based on Probabilistic Graphical Models | Graph structure intuitively displays variable dependency relationships, providing quantitative conditional probability descriptions. | ■ *Structural equation model* (Hu et al., 2019);<br>■ *HFACS-Bayesian Networks* (Kumar et al., 2020; Morais et al., 2020), *Fuzzy-BN* (de Maya et al., 2020; Ma et al., 2022), *DBN* (Pan et al., 2024),<br>■ *FTA&ETA* (Weber et al., 2012)<br>■ *SD* (Hulme et al., 2019) |
| | Other Methods | Methods based on systems theory, information theory, and graph theory. | ■ *Activity Theory-Based* (Yoon et al., 2016, 2017);<br>■ *Safety Information Cognition Models* (Chen et al., 2021);<br>■ *Causal Association Diagrams* (Wu et al., 2024) |

**Table 4**
Studies addressing Q3.

| Q3 | Risk Control | References |
|---|---|---|
| Human-Centered Control | Selecting suitable people to perform suitable operations | ■ *Selection and training* (Lyssakov and Lyssakova, 2019; Salas et al., 2012)<br>■ *Job competency model* (Piliuhina et al., 2024)<br>■ *Personnel-task matching mechanism* (Wei et al., 2020)<br>■ *Organizational management* (Lund and Aarø, 2004; Molan and Molan, 2020) |
| System-Centered Control | Making systems adapt to human errors | ■ *Error-proof design* (physical/ logical constraints) (Manigandan, 2024; Shingo, 1986)<br>■ *Fault-tolerant mechanisms* (Stetter et al., 2020)<br>■ *Real-time warning* (Li et al., 2025; Zhang et al., 2026) |
| Human-Machine Collaboration-Centered Control | Humans understand machines, machines understand humans | ■ *Human-Machine Interface Design* (Chen et al., 2021)<br>■ *Cognitive state modeling* (Chen et al., 2018)<br>■ *Intelligent task allocation* (Yang et al., 2021)<br>■ *Human-machine trust calibration* (Ding et al., 2025b)<br>■ *Human-machine emotional interaction* (Bernardo and Seva, 2023) |

fallibility. While these mechanisms provide deterministic reliability, they frequently precipitate the ironies of automation. Over-reliance on rigid, error-proof designs can inadvertently lead to skill degradation and a diminished sense of system awareness among operators. Our findings indicate that most contemporary data-driven systems remain tethered to hard-coded rules, lacking the cybernetic adaptability essential for navigating complex, dynamic environments. When automated safety mechanisms operate without sufficient transparency—a systemic issue directly tied to the Q2 explainability challenge—they can trigger automation surprises. In such instances, operators are rendered unable to intervene effectively because the underlying logic of the system has become decoupled from their internal mental models.

(3) Human-Machine Collaboration: Trust Calibration within the Industry 5.0 Frontier

Human-machine collaboration (HMC) represents the state-of-the-art paradigm in risk control, emphasizing bidirectional learning and mutual adaptation between human operators and artificial intelligence agents. Recent studies published highlight that such synergistic interaction is fundamental to achieving collaborative gains in complex systems. However, a profound challenge within the HMC framework lies in trust calibration. If risk predictions are excessively sensitive, leading to frequent false alarms, or remain opaque due to the explainability deficiencies identified in the Q2 dimension, operators are likely to resort to system disuse or misuse. The research reveals that contemporary DD-HEARC research lacks a robust dynamic trust model—a mechanism capable of adjusting automation levels based on the real-time reliability of both the human and machine components.

### 4.4. The systemic constraints of Q1, Q2, and Q3

Through in-depth analysis, this study found that complex mutual constraint relationships exist among the three challenges Q1, Q2, and Q3, forming systematic obstacles to DD-HEARC practical application, as shown in Fig. 6. To systematically quantify the fragmentation observed across the DD-HEARC literature, we synthesized the coverage patterns of all reviewed studies with respect to the three core challenges (Q1–Q3). The results are presented in Table 5, which categorizes each study based on its explicit focus (single vs. multiple challenges), dominant methodological approaches, and representative references.

The data reveal a stark imbalance: the overwhelming majority of studies focus exclusively on one of the three challenges, while only few studies address two or more challenges in an integrated manner. Furthermore, the methodological strategies employed within each category show limited cross-pollination: Q1 studies predominantly rely on data augmentation or transfer learning without causal safeguards; Q2 studies favor post-hoc explainability tools or static Bayesian networks; and Q3 studies remain largely conceptual or domain-specific without systematic links to data or explanation layers. This empirical evidence strongly supports our central claim that the DD-HEARC field is characterized by research silos, which not only limit theoretical advancement but also reinforce the vicious cycle depicted in Fig. 6.

The three challenges form a vicious cycle of constraint relationships where the output of one phase fundamentally limits the input of the next.

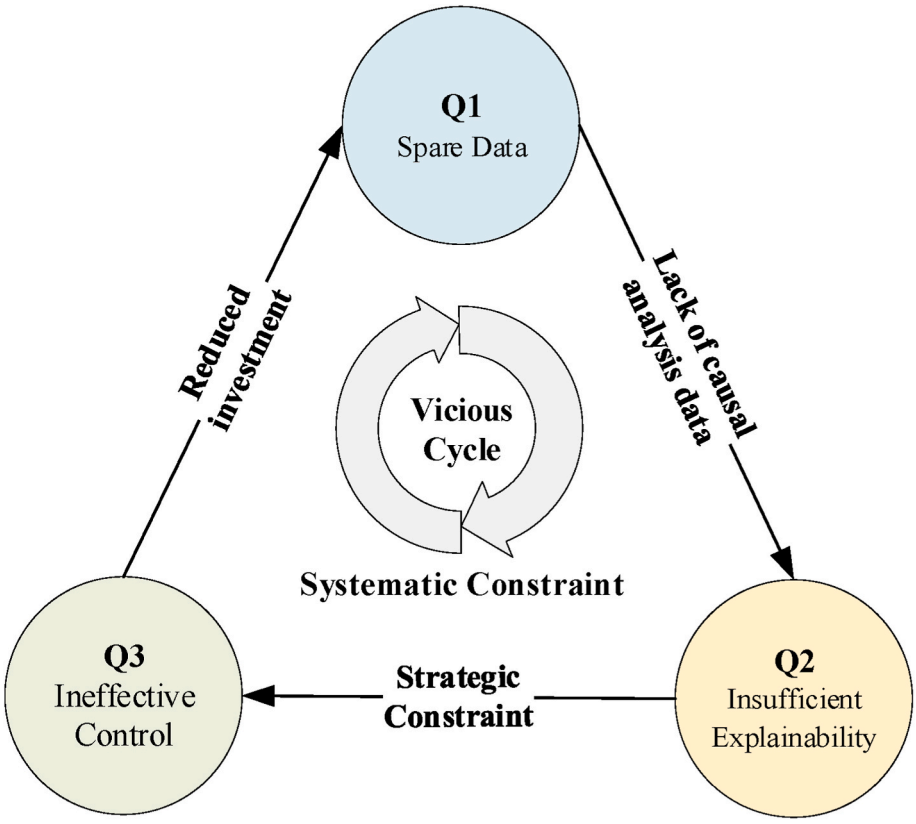


**Fig. 6.** The vicious cycle of three core challenges in DD-HEARC

**Table 5**
Cross-dimensional coverage and methodological characteristics of studies.

| Coverage Type | Number (%) | Method Category | Representative References (≥3 per cell) |
|---|---|---|---|
| Q1 only | 36.8% | Data augmentation, PSFs questionnaires, Transfer learning | Han et al. (2005); Groth and Mosleh (2009); Franciosi et al. (2019); Parsa et al. (2019); Chen et al. (2022) |
| Q2 only | 32.2% | SHAP/LIME, Logistic regression, Bayesian Networks | Zhong et al. (2021); Maternová et al. (2023); Morais et al. (2020); Hu et al. (2019); Abellán et al. (2013) |
| Q3 only | 16.1% | HMI design, Training models, Error-proofing | Chen et al. (2021); Manigandan (2024); Lyssakov and Lyssakova (2019); Salas et al. (2012); Stetter et al. (2020) |
| Q1+Q2 | 9.2% | Causal GANs + SHAP, BN-based data fusion | Wang et al. (2024); Ding et al. (2025a); Emroozi et al. (2024a); Liu et al. (2022); Pan et al. (2024) |
| Q2+Q3 | 3.4% | BN-based intervention, Rule-based control | Kumar et al. (2020); Cheng et al. (2022); Yang et al. (2021) |
| Q1+Q3 | 1.2% | Data-augmented control rules | Zhou and Lei (2020) |
| Q1+Q2+Q3 | 1.1% | Integrated simulation framework | Emroozi et al. (2024b) |

1) Q1→Q2 (Fundamental Constraint): The data foundation determines the analysis quality. Sparse or low-quality data (Q1) forces models to rely on statistical correlations, which leads to "causal blindness" or wrong causation in the explainability (Q2).
2) Q2 → Q3 (Strategic Constraint): The quality of understanding determines the effectiveness of control. When causal explanations are inaccurate or lack transparency, managers cannot formulate targeted intervention strategies, leading to ineffective or even counterproductive risk control (Q3).
3) Q3 → Q1 (Feedback Constraint): This is the crucial link that closes the vicious cycle. Poor implementation of risk control (Q3) reduces industry confidence and willingness to invest in HEA management, which in turn diminishes the resources available for high-quality data collection, resulting in even sparser data for future research (Q1).

This reveals that DD-HEARC faces not isolated technical challenges but a systematic vicious cycle: Q1's data sparsity limits Q2's causal inference accuracy, Q2's insufficient explainability hinders Q3's effective control strategy formulation, and Q3's transformation difficulties further weaken data collection investment. Therefore, DD-HEARC requires systems thinking and integrated methods. By establishing unified theoretical frameworks, developing integrated methodological systems, and improving overall evaluation standards can DD-HEARC achieve the leap from theoretical exploration to practical application.

## 5. Research gaps and future directions

### 5.1. Systematic identification of key research limitations

#### 5.1.1. Deficiencies in sparse data processing

Current sparse data processing research exhibits three fundamental deficiencies. Causal blindness deficiency: existing data augmentation methods focus on preserving statistical characteristics while ignoring causal structure integrity, producing more severe consequences than data sparsity in safety domains. Contextual simplification deficiency: most methods assume independent PSFs actions with linear

superposition modeling, but real industrial systems involve complex nonlinear interactions, temporal accumulation effects, and contextual dependencies (Miranda et al., 2013). Validation system absence deficiency: research lacks specialized quality evaluation systems for HEA domains, continuing to use general machine learning metrics while safety-critical applications require causal consistency, contextual adaptability, and ethical compliance dimensions.

#### 5.1.2. Disconnection between explainability theory and safety practice

The core challenge manifests at cognitive, technical, and application levels. Cognitive paradigm conflicts reflect fundamental differences between scientific cognition (pursuing objectivity, universality, reproducibility) and management decision-making (emphasizing contextuality, timeliness, operability), leading to scientifically correct explanations lacking management operability (Nyrup and Robinson, 2022). Technical capability boundaries appear in current explainable AI limitations: LIME and SHAP produce misleading results in complex nonlinear systems (Lundberg and Lee, 2017), while causal inference faces bottlenecks with high-dimensional variables and time-varying structures (Islam et al., 2022). Application adaptability gaps reflect mismatches between general explanation technologies and safety domain needs requiring stable, traceable, auditable explanations.

#### 5.1.3. Systematic barriers in transformation application

Transformation from causal analysis to risk control faces systematic barriers at methodological and theoretical levels. Theoretical bridge absence constitutes the fundamental cause, where existing research lacks systematic connections between causal analysis addressing "what/why" questions and risk control practice needing "how/effects" answers (Anjum and Rocca, 2019). Personalization-standardization contradictions emerge as DD-HEARC provides differentiated risk control while industrial safety management emphasizes standardization (Battles et al., 2006). Dynamic adaptation insufficiency manifests in mismatches between static analysis methods and dynamic control requirements, lacking real-time monitoring technology, dynamic updating mechanisms, and adaptive control algorithms (Seiti et al., 2022).

#### 5.1.4. Methodological gaps in cross-dimensional integration

Cross-dimensional integration research scarcity reflects fundamental methodological gaps constraining DD-HEARC implementation. Missing systems thinking appears in localized research perspectives lacking systematic DD-HEARC understanding, leading to uncoordinated technical development. Integration theory absence means research adopts empirical "trial and error" methods without scientific integration strategies. Disunified evaluation standards create difficulties in ensuring research quality, where existing studies use respective evaluation systems lacking unified benchmarks, constraining knowledge accumulation and method improvement.

### 5.2. Strategic directions for future research

#### 5.2.1. Causal-aware data intelligence

The primary direction for future research is developing causal-aware data intelligence technology to fundamentally solve data sparsity problems. This includes causal-constrained data augmentation embedding causal structure constraints in data generation to ensure synthetic data maintains consistency in both statistical characteristics and causal mechanisms. Multi-modal sparse data fusion integrates text reports, numerical monitoring, images, and sensor data to construct panoramic HEA data views. Crucially, in the context of embodied intelligence, this sensing layer must extend to dyadic interaction data, capturing the synchronized states of both the human supervisor and the autonomous agent to identify latent conflicts in collaborative tasks (Prabhakar and Murphey, 2022). Adaptive data quality assessment mechanisms dynamically monitor causal distortion risks, while causal invariance theory identifies stable relationships across HEA domains to achieve cross-domain knowledge transfer.

#### 5.2.2. Context-adaptive causal explanation

Developing context-adaptive causal explanation technology establishes explainable AI systems truly meeting safety management needs. Multi-level causal explanation frameworks provide differentiated services according to stakeholders' cognitive characteristics and decision-making needs. For systems utilizing embodied intelligence, explainability must move beyond mathematical transparency toward intent-based transparency, ensuring that a robot's physical trajectory and planned actions are predictable and aligned with the operator's mental model (Luo et al., 2025). Dynamic causal inference systems achieve transformation from static analysis to dynamic monitoring, while counterfactual reasoning answers what effects would occur if certain interventions were taken through high-fidelity simulation and multi-objective optimization.

#### 5.2.3. Intelligent risk control systems for collaborative environments

Constructing intelligent risk control systems achieves transformation from passive response to proactive prevention. A critical frontier resides in addressing the shifting error pathways within embodied intelligence, where risks transition from manual execution failures to supervisory lapses and automation biases in ostensibly unmanned processes. Individual-specific risk profiling establishes dynamic profiles for each operator, predicting error risks based on cognitive traits and experience. Adaptive intervention strategies must therefore evolve to manage human-robot trust calibration, dynamically adjusting autonomy levels or providing haptic/visual cues to maintain the human-machine team within a safe performance envelope (Guo and Yang, 2020). Effect evaluation establishes continuous optimization mechanisms through real-time monitoring and long-term causal effect tracking.

#### 5.2.4. Cross-disciplinary methodological innovation

Promoting cross-disciplinary methodological innovation establishes theoretical systems and technical standards for DD-HEARC. Multidisciplinary theoretical integration merges causal inference, machine learning, and safety science with robotics and human-computer interaction (HCI) to address the systemic complexities of Industry 5.0 (Gholamizadeh et al., 2025). Standardized evaluation systems establish unified benchmarks and multidimensional indicators for data quality, model performance, and application value. Open collaborative ecosystems construct industry-academia-research integration through cross-institutional data sharing and open-source tools, lowering the barriers to implementing resilient risk control in modern socio-technical systems.

## 6. Data-driven HEA three-dimensional analysis framework

Based on the systematic review findings, this study reveals that the DD-HEARC field suffers from systematic fragmentation and vicious cycle constraint relationships among three core challenges. To systematically break this cycle, this chapter proposes a three-dimensional analysis framework as the first systematic method integrating fragmented research elements into a comprehensive problem analysis space, as shown in Fig. 7. The framework's breakthrough value lies in: providing unified problem understanding by integrating scattered elements into temporal × logical × collaborative dimensions; establishing systematic cross-dimensional collaboration to break the Q1→Q2→Q3→Q1 vicious cycle; achieving paradigm shift from experience-driven to science-driven approaches for fundamental field transformation.

### 6.1. Temporal dimension: PSF causal evolution process

The temporal dimension reflects dynamic evolution from potential risks to actual consequences, embodying "PSFs → Human Error → HEA" temporal characteristics with distinct stage requirements. PSFs

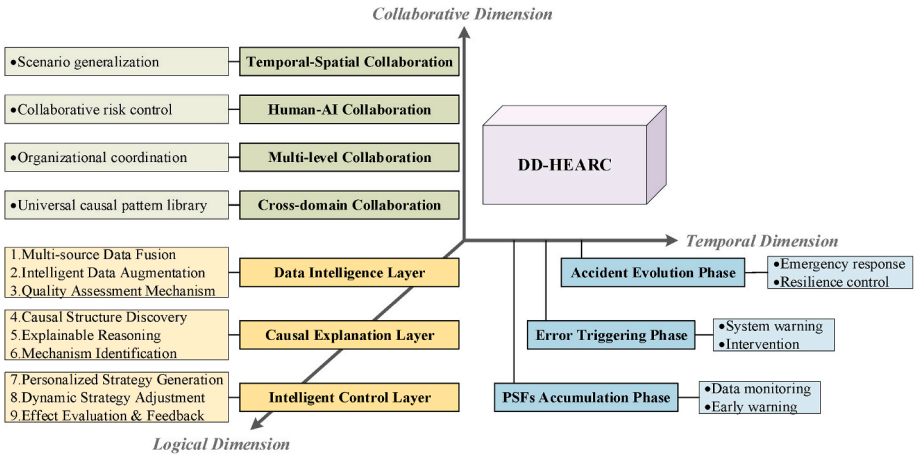


**Fig. 7.** DD-HEARC three-dimensional analysis framework.

Accumulation Phase corresponds to data scarcity problems in Section 4.2, where individual fatigue, team communication barriers, organizational defects, and system design issues manifest and reinforce. This stage requires multi-source data fusion technology, objective PSFs measurement methods, and domain knowledge-based quality control. Control strategies prioritize prevention through continuous monitoring and early warning due to relatively low costs and good intervention effects; Error Triggering Phase addresses human error data scarcity as accumulated PSFs interact with triggering factors, exceeding cognitive boundaries and causing decision errors or violations. This stage needs real-time error detection technology, dynamic causal inference methods, and rapid response mechanisms, requiring immediate intervention despite increased costs; Accident Evolution Phase tackles accident data scarcity when errors propagate through system coupling toward adverse consequences. This stage focuses on data augmentation maintaining causal authenticity, dynamic accident modeling, and emergency response systems, emphasizing loss control despite highest costs.

### 6.2. Logical dimension: problem-solving logic levels

The logical dimension embodies hierarchical DD-HEARC problem-solving from data foundation to intelligent control, forming complete cognitive-action chains through three interconnected layers. Data Intelligence Layer builds analysis foundation addressing Section 4.2 sparsity challenges through: Multi-source Data Fusion establishing cross-domain integration mechanisms for accident reports, operation records, and monitoring data; Intelligent Data Augmentation developing causal-aware technologies using GANs and VAEs with causal constraints; Quality Assessment Mechanisms ensuring data completeness, consistency, and reliability through multi-source verification; Causal Explanation Layer achieves mechanism understanding addressing Section 4.3 explainability challenges through: Causal Structure Discovery learning "PSFs → Human Error → HEA" networks using advanced algorithms; Explainable Reasoning providing transparent causal path explanations for different users; Risk Mechanism Identification revealing internal causation logic through mediation, moderation, and heterogeneity analyses.; Intelligent Control Layer implements precise intervention addressing Section 4.4 transformation challenges through: Personalized Strategy Generation creating individual-specific risk profiles and targeted interventions; Dynamic Strategy Adjustment adapting to real-time conditions through self-learning algorithms; Effect Evaluation and Feedback establishing continuous optimization through real-time monitoring and causal effect tracking.

### 6.3. Collaborative dimension: system collaborative capability

The collaborative dimension addresses Section 4.5 cross-dimensional integration gaps through four systematic collaboration types ensuring framework effectiveness. Cross-Domain Collaboration breaks traditional barriers achieving knowledge sharing among aviation, manufacturing, medical, and transportation fields through unified PSF classification systems and domain-adaptive transfer learning. Multi-Level Collaboration coordinates individual, team, organizational, and system causal relationships through inter-level transmission models and coordinated intervention measures. Human-AI Collaboration combines experiential wisdom with computational capability through expert-guided augmentation, verified discovery, and collaborative strategy formulation. Temporal-Spatial Collaboration coordinates temporal evolution and spatial distribution through dynamic causal modeling capturing time-lag effects and regional differences.

### 6.4. Application guidance value

The framework's core value lies in integrating fragmented elements into systematic cognitive tools providing strategic planning methods. Through three-dimensional cross-positioning, it identifies breakthrough opportunities like "Error Triggering + Causal Explanation + Human-Machine Collaboration" for real-time causal inference needs. For different users, it provides differentiated guidance: theoretical researchers identify weak links and innovation opportunities; method developers clarify systematic positioning and integration pathways; practitioners use it as risk diagnosis and strategy selection tools. The framework establishes common language for cross-disciplinary collaboration, clarifying integration pathways for causal inference, machine learning, and safety engineering. Through systematic quality assurance and continuous improvement mechanisms, it promotes fundamental DD-HEARC transformation from experience-driven to science-driven, fragmented to systematic research, ultimately improving modern complex system safety management levels.

To illustrate how empirical studies map onto this framework, consider Emroozi et al. (2024b), which develops a system dynamics model for human error in industrial maintenance. This study naturally occupies the Error Triggering Phase in the Temporal Dimension, as it models how task complexity and operator fatigue converge to trigger cognitive errors. In the Logical Dimension, it spans the Data Intelligence Layer (using simulation-generated time-series data) and the Causal Explanation Layer (via feedback loop analysis in system dynamics that reveal non-linear causality). In the Collaborative Dimension, it embodies Human-AI Collaboration through an expert-in-the-loop calibration process where domain knowledge refines model parameters.

Similarly, Ding et al. (2025a)'s EEG-based human reliability assessment fits into the PSFs Accumulation Phase, as it continuously monitors fatigue and mood as latent risk factors. Its Logical Dimension includes Data Intelligence (multi-modal fusion of EEG and task data) and Causal Explanation (a Bayesian network linking physiological states to situation awareness errors). The Collaborative Dimension is reflected in Multi-level Collaboration, as the model outputs feed into team-level risk profiling rather than staying at the individual level, thereby connecting cognitive science with operational safety management.

## 7. Conclusions

The primary finding of this study is the first systematic diagnosis of the structural bottlenecks in the DD-HEARC field. We identified that the field faces not isolated technical hurdles, but a vicious cycle constraint relationship (Q1→Q2→Q3→Q1). This systemic fragmentation explains why a decade of localized technical advances has failed to drive a fundamental leap in safety management.

(1) Theoretical Contributions

The theoretical significance of this research is manifested through several pioneering contributions that redefine the conceptual boundaries of the DD-HEARC domain. Primarily, this study provides the first systematic diagnosis of the structural pathologies within the field by elucidating the self-reinforcing vicious cycle between data sparsity (Q1), causal opacity (Q2), and transformation failure (Q3). By identifying this inherent pathology, the research establishes a robust scientific foundation for understanding why contemporary data-driven models frequently fail to meet the rigorous demands of safety-critical contexts. Furthermore, this work fills a significant methodological gap by integrating cybernetics, Safety-II, and cognitive engineering into a unified theoretical system, effectively transitioning the field from a reliance on fragmented algorithms toward a cohesive and integrated knowledge architecture. This synthesis is complemented by the innovation of a three-dimensional analysis framework—comprising logical, temporal, and collaborative dimensions—which serves as a systematic cognitive tool for positioning research within the broader human error accident landscape and facilitating meaningful cross-disciplinary synthesis.

(2) Practical Implications

The practical value of this research lies in its potential to catalyze a fundamental paradigm shift in safety governance by translating theoretical insights into actionable organizational strategies. For safety managers, this study serves as a critical diagnostic instrument to identify the systemic origins of intervention failures, advocating for a transition from rigid, rule-based protocols toward adaptive, causal-aware strategies that are harmonized with the operator's real-time situational awareness. Simultaneously, the work provides method developers with a strategic roadmap and the technical benchmarks required to engineer the next generation of human error accident analytical tools. By prioritizing the development of causal-aware data intelligence and context-adaptive explanations, developers can ensure that artificial intelligence outputs are both ethically trustworthy and operationally actionable within high-stakes environments. At the macro level, this research offers significant guidance for policy makers by elucidating the critical feedback loop between application failure and data investment. By highlighting how operational setbacks discourage resource allocation, the study underscores the necessity of fostering open collaborative ecosystems that lower the institutional barriers to cross-industry data sharing and collective learning, thereby ensuring the long-term resilience of modern socio-technical systems.

## CRediT authorship contribution statement

**Chongfeng Li:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Data curation, Conceptualization. **Shenghan Zhou:** Supervision, Resources, Project administration, Investigation. **Xing Pan:** Supervision, Methodology, Funding acquisition, Conceptualization. **Song Ding:** Visualization, Software, Resources. **Ziyao Li:** Writing – original draft, Validation, Investigation.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. All authors have confirmed that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China [grant number 72071011].

## Data availability

Data will be made available on request.

## References

Abellán, J., López, G., De OñA, J., 2013. Analysis of traffic accident severity using decision rules via decision trees. Expert Syst. Appl. 40, 6047–6054.
Aliabadi, M., Mohammadfam, I., Soltanian, A., Najafi, K., 2022. Human error probability determination in blasting process of ore mine using a hybrid of HEART and best-worst methods. Saf. Health Work 13, 326–335. https://doi.org/10.1016/j.shaw.2022.03.010.
Anjum, R.L., Rocca, E., 2019. From ideal to real risk: philosophy of causation meets risk analysis. Risk Anal. 39, 729–740.
Arun, P., Tauseef, S., Uniyal, U., 2022. Comparison of accident databases and analysis of past industrial accidents in the chemical process industry. Eng. Technol. Appl. Sci. Res. 12, 8922–8927.
Asgher, U., Khalil, K., Khan, M.J., Ahmad, R., Butt, S.I., Ayaz, Y., Naseer, N., Nazir, S., 2020. Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain–computer interface. Front. Neurosci. 14. https://doi.org/10.3389/fnins.2020.00584.
Ashraf, M.W., Hassan, A., Shah, I.A., 2024. V-CAS: a realtime vehicle anti collision system using vision transformer on multi-camera streams. In: 2024 International Conference on Machine Learning and Applications (ICMLA). IEEE, pp. 939–944.
Bader-El-Den, M., Teitei, E., Perry, T., 2018. Biased random forest for dealing with the class imbalance problem. IEEE Transact. Neural Networks Learn. Syst. 30, 2163–2172.
Barbierato, E., Gatti, A., 2024. The challenges of machine learning: a critical review. Electronics 13, 416.
Barua, S., Islam, M.M., Yao, X., Murase, K., 2012. MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans. Knowl. Data Eng. 26, 405–425.
Basra, G., Kirwan, B., 1998. Collection of offshore human error probability data. Reliab. Eng. Syst. Saf., Offshore Safety 61, 77–93. https://doi.org/10.1016/S0951-8320(97)00064-1.
Battles, J.B., Dixon, N.M., Borotkanics, R.J., Rabin-Fastmen, B., Kaplan, H.S., 2006. Sensemaking of patient safety risks and hazards. Health Serv. Res. 41, 1555–1575.
Bauder, M., Lecheler, K., Wech, L., Böhm, K., Paula, D., Schweiger, H.-G., 2022. Determination of accident scenarios via freely available accident databases. Open Eng. 12, 453–467.
Belkadi, A., Bachiri, M., 2025. Towards an industry 5.0 enhanced by AI: a theoretical framework. Eng. Proc. 112, 2.
Bernardo, E., Seva, R., 2023. Affective design analysis of explainable artificial intelligence (xai): a user-centric perspective. In: Informatics. MDPI, p. 32.
Bird, F., Germain, G.L., Clark, D., 1996. Practical Loss Control Leadership, Det Norske Veritas. Inc Revis. Ed, USA.
Boring, R., Blackman, H., 2007. The origins of the SPAR-H method's performance shaping factor multipliers. 2007 IEEE 8th Hum. Factors Power Plants HPRCT 13th Annu. Meet 177–184. https://doi.org/10.1109/HFPP.2007.4413202.
Boring, R.L., 2012. Fifty years of THERP and human reliability analysis. Reliab. Eng. Syst. Saf. 108, 140–148.
Burns, K.J., Bonaceto, C., 2020. An empirically benchmarked human reliability analysis of general aviation. Reliab. Eng. Syst. Saf. 194, 106227. https://doi.org/10.1016/J.RESS.2018.07.028.

Cao, C., Ma, L., Xu, Y., 2012. Adaptive Control Theory and Applications.
Chang, Y.H.J., Kim, Y., Park, J., Criscione, L., 2022. SACADA and HuREX: part 1. the use of SACADA and HuREX systems to collect human reliability data. Nucl. Eng. Technol. 54, 1686–1697. https://doi.org/10.1016/j.net.2021.10.037.
Chang, Y.J., Mosleh, A., 2007. Cognitive modeling and dynamic probabilistic simulation of operating crew response to complex system accidents. Part 2: IDAC performance influencing factors model. Reliab. Eng. Syst. Saf. 92, 1014–1040.
Chen, J.Y., Lakhmani, S.G., Stowers, K., Selkowitz, A.R., Wright, J.L., Barnes, M., 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theor. Issues Ergon. Sci. 19, 259–282.
Chen, S., Zeng, N., Li, F., Yue, H., Wang, Q., Li, Q., 2025. Proactive safety risk control system for deep foundation pit construction: situational tailoring of integrated cybernetics and dual-system theory. J. Construct. Eng. Manag. 151, 4025018. https://doi.org/10.1061/JCEMD4.COENG-15518.
Chen, Y., Feng, W., Jiang, Z., Duan, L., Cheng, S., 2021. An accident causation model based on safety information cognition and its application. Reliab. Eng. Syst. Saf. 207, 107363. https://doi.org/10.1016/j.ress.2020.107363.
Chen, Z., Huang, K., Wu, L., Zhong, Z., Jiao, Z., 2022. Relational graph convolutional network for text-mining-based accident causal classification. Appl. Sci. 12, 2482.
Cheng, C.-H., Yang, J.-H., Liu, P.-C., 2022. Rule-based classifier based on accident frequency and three-stage dimensionality reduction for exploring the factors of road accident injuries. PLoS One 17, e0272956.
Cho, J., Kim, Y., Kim, J., Park, J., Kim, D.-S., 2020. Realistic estimation of human error probability through monte carlo thermal-hydraulic simulation. Reliab. Eng. Syst. Saf. 193, 106673. https://doi.org/10.1016/j.ress.2019.106673.
Christian, M.S., Bradley, J.C., Wallace, J.C., Burke, M.J., 2009. Workplace safety: a meta-analysis of the roles of person and situation factors. J. Appl. Psychol. 94, 1103.
de Maya, B.N., Babaleye, A.O., Kurt, R.E., 2020. Marine accident learning with fuzzy cognitive maps (MALFCMs) and bayesian networks. Saf. Extreme Environ. 2, 69–78.
Diaconeasa, M.A., Mosleh, A., 2018. Performing an accident sequence precursor analysis with the ADS-IDAC dynamic PSA software platform. In: Probabilistic Safety Assessment and Management Conference. Presented at the Probabilistic Safety Assessment and Management Conference.
Ding, S., Hu, L., Pan, X., Zuo, D., Sun, L., 2025a. Assessing human situation awareness reliability considering fatigue and mood using EEG data: a bayesian neural network-bayesian network approach. Reliab. Eng. Syst. Saf. 260, 110962. https://doi.org/10.1016/j.ress.2025.110962.
Ding, S., Pan, X., Hu, L., Liu, L., 2025b. A new model for calculating human trust behavior during human-AI collaboration in multiple decision-making tasks: a bayesian approach. Comput. Ind. Eng. 200, 110872. https://doi.org/10.1016/j.cie.2025.110872.
Donaldson, M.S., Corrigan, J.M., Kohn, L.T., 2000. To Err is Human: Building a Safer Health System.
Emroozi, V.B., Kazemi, M., Pooya, A., Doostparast, M., 2024a. Evaluating human error probability in maintenance task: an integrated system dynamics and machine learning approach. Hum. Factors Ergon. Manuf. Serv. Ind. 35. https://doi.org/10.1002/hfm.21057.
Emroozi, V.B., Kazemi, M., Pooya, A., Doostparast, M., 2024b. Dynamic modeling of human error in industrial maintenance through structural analysis and system dynamics. Risk Anal. Off. Publ. Soc. Risk Anal. https://doi.org/10.1111/risa.17652.
Fathy, Y., Jaber, M., Brintrup, A., 2020. Learning with imbalanced data in smart manufacturing: a comparative analysis. IEEE Access Pract. Innov. Open Solut. 9, 2734–2757.
Forester, J., Bley, D., Cooper, S., Lois, E., Siu, N., Kolaczkowski, A., Wreathall, J., 2004. Expert elicitation approach for performing ATHEANA quantification. Reliab. Eng. Syst. Saf. 83, 207–220. https://doi.org/10.1016/j.ress.2003.09.011.
Franciosi, C., Di Pasquale, V., Iannone, R., Miranda, S., 2019. A taxonomy of performance shaping factors for human reliability analysis in industrial maintenance. J. Ind. Eng. Manag. 12, 115. https://doi.org/10.3926/jiem.2702.
Garrett, J.W., Teizer, J., 2009. Human factors analysis classification system relating to human error awareness taxonomy in construction safety. J. Constr. Eng. Manag.-Asce 135, 754–763. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000034.
Gharoun, H., Momenifar, F., Chen, F., Gandomi, A.H., 2023. Meta-learning approaches for few-shot learning: a survey of recent advances. ACM Comput. Surv. 56, 1–41. https://doi.org/10.1145/3659943.
Gholamizadeh, K., Zarei, E., Gualtieri, L., Marchi, M.D., 2025. Advancing occupational and system safety in industry 5.0: effective HAZID, risk analysis frameworks, and human-AI interaction management. Saf. Sci. https://doi.org/10.1016/j.ssci.2024.106770.
Griffith, C.D., Mahadevan, S., 2015. Human reliability under sleep deprivation: derivation of performance shaping factor multipliers from empirical data. Reliab. Eng. Syst. Saf. 144, 23–34. https://doi.org/10.1016/j.ress.2015.05.004.
Groth, K.M., Mosleh, A., 2009. A data-informed model of performance shaping factors and their interdependencies for use in human reliability analysis. In: Reliability, Risk, and Safety, Three Volume Set. CRC Press, pp. 265–272.
Guo, Y., Yang, X.J., 2020. Modeling and predicting trust dynamics in human–robot teaming: a bayesian inference approach. Int. J. Soc. Robot. 13, 1899–1909. https://doi.org/10.1007/s12369-020-00703-3.
Han, H., Wang, W.-Y., Mao, B.-H., 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing. Springer, pp. 878–887.
He, Y., Kuai, N.-S., Deng, L.-M., He, X.-Y., 2021. A method for assessing human error probability through physiological and psychological factors tests based on CREAM and its applications. Reliab. Eng. Syst. Saf. 215, 107884. https://doi.org/10.1016/j.ress.2021.107884.
Hollnagel, E., 2014. Safety-I and safety-II: the past and Future of Safety Management. Ashgate Publishing, Farnham.
Hollnagel, E., 1998. Cognitive Reliability and Error Analysis Method (CREAM). Elsevier, Oxford.
Hong, S.R., Hullman, J., Bertini, E., 2020. Human factors in model interpretability: industry practices, challenges, and needs. Proc. ACM Hum.-Comput. Interact. 4, 1–26.
Hu, S., Li, Z., Xi, Y., Gu, X., Zhang, X., 2019. Path analysis of causal factors influencing marine traffic accident via structural equation numerical modeling. J. Mar. Sci. Eng. 7, 96. https://doi.org/10.3390/jmse7040096.
Hulme, A., Stanton, N.A., Walker, G.H., Waterson, P., Salmon, P.M., 2019. What do applications of systems thinking accident analysis methods tell us about accident causation? A systematic review of applications between 1990 and 2018. Saf. Sci. 117, 164–183. https://doi.org/10.1016/j.ssci.2019.04.016.
Islam, M.R., Ahmed, M.U., Barua, S., Begum, S., 2022. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Appl. Sci. 12, 1353.
Jha, P.R., Pasquale, V.D., Saleem, J.J., Wang, X., 2024. Taxonomy of performance shaping factors in manufacturing: a systematic literature review. Hum. Factors Ergon. Manuf. Serv. Ind. 34, 367–385. https://doi.org/10.1002/hfm.21036.
Jiang, Y., Qu, X., Zhang, W., Guo, W., Xu, J., Yu, W., Chen, Y., 2025. Analyzing crash severity: human injury severity prediction method based on transformer model. Vehicles. https://doi.org/10.3390/vehicles7010005.
Jung, W., Park, J., Kim, Y., Choi, S.Y., Kim, S., 2020. HuREX – a framework of HRA data collection from simulators in nuclear power plants. Reliab. Eng. Syst. Saf. 194, 106235. https://doi.org/10.1016/j.ress.2018.07.036.
Kim, J.W., Jung, W., 2003. A taxonomy of performance influencing factors for human reliability analysis of emergency tasks. J. Loss Prev. Process. Ind. 16, 479–495. https://doi.org/10.1016/S0950-4230(03)00075-5.
Kim, Y., Park, J., 2019. Incorporating prior knowledge with simulation data to estimate PSF multipliers using Bayesian logistic regression. Reliab. Eng. Syst. Saf. 189, 210–217. https://doi.org/10.1016/J.RESS.2019.04.022.
Kim, Y., Park, J., Jung, W., Jang, I., Seong, P., 2015. A statistical approach to estimating effects of performance shaping factors on human error probabilities of soft controls. Reliab. Eng. Syst. Saf. 142, 378–387. https://doi.org/10.1016/j.ress.2015.06.004.
Koc, K., Gurgun, A., 2022. Scenario-based automated data preprocessing to predict severity of construction accidents. Autom. ConStruct. https://doi.org/10.1016/j.autcon.2022.104351.
Kumar, P., Gupta, S., Gunda, Y.R., 2020. Estimation of human error rate in underground coal mines through retrospective analysis of mining accident reports and some error reduction strategies. Saf. Sci. 123, 104555. https://doi.org/10.1016/j.ssci.2019.104555.
Kyriakidis, M., de Winter, J.C., Stanton, N., Bellet, T., van Arem, B., Brookhuis, K., Martens, M.H., Bengler, K., Andersson, J., Merat, N., Others, 2019. A human factors perspective on automated driving. Theor. Issues Ergon. Sci. 20, 223–249.
La Fata, C., Adelfio, L., Micale, R., La Scalia, G., 2023. Human error contribution to accidents in the manufacturing sector: a structured approach to evaluate the interdependence among performance shaping factors. Saf. Sci. 161, 106067.
Larsson, P., Tingvall, C., 2013. The safe system approach–a road safety strategy based on human factors principles. In: International Conference on Engineering Psychology and Cognitive Ergonomics. Springer, pp. 19–28.
Leveson, N.G., 2011. Engineering a Safer World: Systems Thinking Applied to Safety. MIT Press, Cambridge.
Li, C., Pan, X., Yang, L., Wang, J., Ma, H., 2025. Human control mode enables accurate real-time risk warning in human–machine systems. Comput. Ind. Eng. 204, 111110.
Li, C., Sun, R., Pan, X., 2023. Takeoff runway overrun risk assessment in aviation safety based on human pilot behavioral characteristics from real flight data. Saf. Sci. 158, 105992. https://doi.org/10.1016/j.ssci.2022.105992.
Li, J., Wang, S., 2025. UAV accident forensics via HFACS-LLM reasoning: Low-altitude safety insights. Drones 9, 704.
Li, Y., Wang, Y., Zhou, X., Zhang, W., Zheng, J., 2026. Detection of perceived risk during partially automated driving on real road. Int. J. Ind. Ergon. 111, 103842. https://doi.org/10.1016/j.ergon.2025.103842.
Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2020. Explainable ai: a review of machine learning interpretability methods. Entropy Int. Interdiscip. J. Entropy Inf. Stud. 23, 18.
Liu, J., Zou, Y., Wang, W., Zio, E., Yuan, C., Wang, T., Jiang, J., 2022. A bayesian belief network framework for nuclear power plant human reliability analysis accounting for dependencies among performance shaping factors. Reliab. Eng. Syst. Saf. 228, 108766. https://doi.org/10.1016/j.ress.2022.108766.
Lu, K.-Q., Dahlman, A.S., Karlsson, J., Candefjord, S., 2022. Detecting driver fatigue using heart rate variability: a systematic review. Accid. Anal. Prev. 178, 106830. https://doi.org/10.1016/j.aap.2022.106830.
Lund, J., Aarø, L., 2004. Accident prevention. Presentation of a model placing emphasis on human, structural and cultural factors. Saf. Sci. 42, 271–324. https://doi.org/10.1016/S0925-7535(03)00045-6.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.
Luo, J., Zhang, C., Si, W., Jiang, Y., Yang, C., Zeng, C., 2025. A physical human–robot interaction framework for trajectory adaptation based on human motion prediction and adaptive impedance control. IEEE Trans. Autom. Sci. Eng. 22, 5072–5083. https://doi.org/10.1109/tase.2024.3415650.
Lyssakov, N., Lyssakova, E., 2019. Human factor as a cause of aircraft accidents. Proc. II Int. Sci.-Pract. Conf. Psychol. Extreme Prof. ISPCPEP. https://doi.org/10.2991/ISPCPEP-19.2019.31, 2019.

Ma, L., Ma, X., Lan, H., Liu, Y., Deng, W., 2022. A data-driven method for modeling human factors in maritime accidents by integrating DEMATEL and FCM based on HFACS: a case of ship collisions. Ocean Eng. https://doi.org/10.1016/j.oceaneng.2022.112699.
Manigandan, DrS.K., 2024. Accident prevention system. Interantional J. Sci. Res. Eng. Manag. https://doi.org/10.55041/ijsrem36861.
Maternová, A., Materna, M., Dávid, A., Török, Á., Svabova, L., 2023. Human error analysis and fatality prediction in maritime accidents. J. Mar. Sci. Eng. 11. https://doi.org/10.3390/jmse11122287.
Maynard, E., Harris, D., 2022. Using neural networks to predict high-risk flight environments from accident and incident data. Int. J. Occup. Saf. Ergon. 28, 1204–1212.
Mehra, A., 2024. Hybrid AI models: integrating symbolic reasoning with deep learning for complex decision-making. J. Emerg. Technol. Innov. Res. 11, f693–f695.
Menezes, M.L.A., Haddad, A., Nascimento, M.L.F., 2021. Functional resonance analysis method and human performance factors identifying critical functions in chemical process safety. IEEE Access Pract. Innov. Open Solut. https://doi.org/10.1109/ACCESS.2021.3135747, 1–1.
Miranda, S., Riemma, S., Iannone, R., Di Pasquale, V., 2013. Rijeka. In: An Overview of Human Reliability Analysis Techniques in Manufacturing Operations, 221. Rij. Croac. Intechopen.
Molan, G., Molan, M., 2020. Theoretical model for accident prevention based on root cause analysis with graph theory. Saf. Health Work 12, 42–50. https://doi.org/10.1016/j.shaw.2020.09.004.
Morais, C., Moura, R., Beer, M., Patelli, E., 2020. Analysis and estimation of human errors from major accident investigation reports. ASCE-ASME J. Risk Uncertain. Eng. Syst. Part B Mech. Eng. 6, 11014. https://doi.org/10.1115/1.4044796.
Morais, C., Yung, K.L., Johnson, K., Moura, R., Beer, M., Patelli, E., 2022. Identification of human errors and influencing factors: a machine learning approach. Saf. Sci. 146, 105528. https://doi.org/10.1016/j.ssci.2021.105528.
Moura, R., Beer, M., Patelli, E., Lewis, J., Knoll, F., 2016. Learning from major accidents to improve system design. Saf. Sci. 84, 37–45.
Mujalli, R., López, G., Garach, L., 2016. Bayes classifiers for imbalanced traffic accidents datasets. Accid. Anal. Prev. 88, 37–51. https://doi.org/10.1016/j.aap.2015.12.003.
Nekooeimehr, I., Lai-Yuen, S.K., 2016. Adaptive semi-unsupervised weighted oversampling (a-SUWO) for imbalanced datasets. Expert Syst. Appl. 46, 405–416.
Nobles, C., 2018. Botching human factors in cybersecurity in business organizations. HOLISTICA – J. Bus. Public Adm 9, 71–88. https://doi.org/10.2478/hjbpa-2018-0024.
Nyrup, R., Robinson, D., 2022. Explanatory pragmatism: a context-sensitive framework for explainable medical AI. Ethics Inf. Technol. 24, 13.
Olivares, R.D.C., Rivera, S.S., Leod, J., 2018. A novel qualitative prospective methodology to assess human error during accident sequences. Saf. Sci. 103, 137–152. https://doi.org/10.1016/J.SSCI.2017.10.023.
Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Others, 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Br. Med. J. 372.
Pan, X., Du, H., Yu, H., 2024. A method for safety analysis of human-machine systems based on dynamic bayesian simulation. Reliab. Eng. Syst. Saf. 248, 110152. https://doi.org/10.1016/j.ress.2024.110152.
Park, J., 2024. A framework to determine the holistic multiplier of performance shaping factors in human reliability analysis – an explanatory study. Reliab. Eng. Syst. Saf. 242, 109727. https://doi.org/10.1016/j.ress.2023.109727.
Park, K.-S., in Lee, J., 2008. A new method for estimating human error probabilities: AHP-slim. Reliab. Eng. Syst. Saf. 93, 578–587. https://doi.org/10.1016/j.ress.2007.02.003.
Parsa, A., Taghipour, H., Derrible, S., Mohammadian, A., 2019. Real-time accident detection: coping with imbalanced data. Accid. Anal. Prev. 129, 202–210. https://doi.org/10.1016/j.aap.2019.05.014.
Pearl, J., 2019. Causal inference in statistics: an overview. Stat. Surv. 3, 96–146.
Pearl, J., 2010. An introduction to causal inference. Int. J. Biostat. 6. https://doi.org/10.2202/1557-4679.1203.
Piliuhina, K., Bushuyev, S., Cirillo, R., Pavel, G.L., Ricotti, M., Janssens, W., 2024. Development of the nuclear competences based on global trends in the nuclear industry. Nucl. Eng. Des. 421, 113046.
Prabhakar, A., Murphey, T., 2022. Mechanical intelligence for learning embodied sensor-object relationships. Nat. Commun. 13. https://doi.org/10.1038/s41467-022-31795-2.
Rahman, F.A., Arifin, K., Abas, A., Mahfudz, M., Basir Cyio, M., Khairil, M., Ali, M.N., Lampe, I., Samad, M.A., 2022. Sustainable safety management: a safety competencies systematic literature review. Sustainability 14, 6885. https://doi.org/10.3390/su14116885.
Ramezani, A., Nazari, T., Rabiee, A., Hadad, K., Faridafshin, M., 2020. Human error probability quantification for NPP post-accident analysis using cognitive-based THERP method. Prog. Nucl. Energy 123, 103281. https://doi.org/10.1016/j.pnucene.2020.103281.
Ras, G., Xie, N., Van Gerven, M., Doran, D., 2022. Explainable deep learning: a field guide for the uninitiated. J. Artif. Intell. Res. 73, 329–396.
Rasmussen, J., 1990. Human error and the problem of causality in analysis of accidents. Philos. Trans. R. Soc. Lond. B Biol. Sci. 327, 449–462. https://doi.org/10.1098/rstb.1990.0088.
Reason, J., 1990. Human error. Saf. Sci. 12, 3–14.
Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144.
Salas, E., Tannenbaum, S.I., Kraiger, K., Smith-Jentsch, K.A., 2012. The science of training and development in organizations: what matters in practice. Psychol. Sci. Public Interest 13, 74–101.
Salmon, P.M., Naughton, M., Hulme, A., McLean, S., 2022. Bicycle crash contributory factors: a systematic review. Saf. Sci. 145, 105511.
Sanneman, L., Shah, J.A., 2022. The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. Int. J. Human–Computer Interact 38, 1772–1788. https://doi.org/10.1080/10447318.2022.2081282.
Seiti, H., Makui, A., Hafezalkotob, A., Khalaj, M., Hameed, I.A., 2022. R.graph: a new risk-based causal reasoning and its application to COVID-19 risk analysis. Process Saf. Environ. Prot. 159, 585–604. https://doi.org/10.1016/j.psep.2022.01.010.
Sen, G., Wenjun, H., Hanyu, W., Qingbin, W., 2026. Assessing pilot cognitive overload risk with a random forest framework: a non-contact approach based on a novel cardiopulmonary feature. Int. J. Ind. Ergon. 111, 103865.
Seong, S., Park, J., Kim, J.H., 2025. A new measurement for workload assessment in agricultural tasks: EDA-based real-time model. Int. J. Ind. Ergon. 108, 103771. https://doi.org/10.1016/j.ergon.2025.103771.
Sharma, K., Niforatos, E., Giannakos, M., Kostakos, V., 2020. Assessing cognitive performance using physiological and facial features. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 4, 1–41. https://doi.org/10.1145/3411811.
Shin, D., Shin, E.Y., 2023. Data's impact on algorithmic bias. Computer 56, 90–94. https://doi.org/10.1109/MC.2023.3262909.
Shingo, S., 1986. Zero Quality Control: Source Inspection and the poka-yoke System. Productivity Press.
Siddaway, A.P., Wood, A.M., Hedges, L.V., 2019. How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. Annu. Rev. Psychol. 70, 747–770.
Spirtes, P., Zhang, K., 2016. Causal discovery and inference: concepts and recent methodological advances. In: Applied Informatics. Springer, p. 3.
Stetter, R., Göser, R., Gresser, S., Till, M., Witczak, M., 2020. Fault-tolerant design for increasing the reliability of an autonomous driving gear shifting system. Eksploat. Niezawodn. 22, 482–492.
Sun, Z., Li, Z., Gong, E., Xie, H., 2012. Estimating human error probability using a modified CREAM. Reliab. Eng. Syst. Saf. 100, 28–32. https://doi.org/10.1016/j.ress.2011.12.017.
Takeda, S., Kitada, T., 2024. Bayesian inference based on monte carlo technique for multiplier of performance shaping factor. ASCE-ASME J. Risk Uncertain. Eng. Syst. Part B Mech. Eng. https://doi.org/10.1115/1.4065531.
Tang, X., Qiu, J., Chen, R., Zhang, W., Iosifidis, V., Liu, Z., Meng, W., Zhang, M., Zhang, J., 2020. A data-driven human responsibility management system. In: 2020 IEEE International Conference on Big Data (Big Data). Presented at the 2020 IEEE International Conference on Big Data (Big Data). IEEE, Atlanta, GA, USA, pp. 5834–5838. https://doi.org/10.1109/BigData50022.2020.9378484.
Tanguy, L., Tulechki, N., Urieli, A., Hermann, E., Raynal, C., 2016. Natural language processing for aviation safety reports: from classification to interactive analysis. Comput. Ind. 78, 80–95.
Tian, D., Liu, H., Chen, S., Li, M., Liu, C., 2022. Human error analysis for hydraulic engineering: comprehensive system to reveal accident evolution process with text knowledge. J. Construct. Eng. Manag. https://doi.org/10.1061/(asce)co.1943-7862.0002366.
Triplett, W.J., 2022. Addressing human factors in cybersecurity leadership. J. Cybersecur. Priv. 2, 573–586. https://doi.org/10.3390/jcp2030029.
Vinodkumar, M.N., Bhasi, M., 2010. Safety management practices and safety behaviour: assessing the mediating role of safety knowledge and motivation. Accid. Anal. Prev. 42, 2082–2093. https://doi.org/10.1016/j.aap.2010.06.021.
Visani, G., Bagli, E., Chesani, F., Poluzzi, A., Capuzzo, D., 2022. Statistical stability indices for LIME: obtaining reliable explanations for machine learning models. J. Oper. Res. Soc. 73, 91–101.
Wang, L., Ren, Y., Jiang, H., Cai, P., Fu, D., Wang, T., Cui, Z., Yu, H., Wang, X., Zhou, H., Others, 2024. Accidentgpt: a v2x environmental perception multi-modal large model for accident analysis and prevention. In: 2024 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp. 472–477.
Weber, P., Medina-Oliva, G., Simon, C., Iung, B., 2012. Overview on bayesian networks applications for dependability, risk analysis and maintenance areas. Eng. Appl. Artif. Intell., Special Section: Dependable System Modelling and Analysis 25, 671–682. https://doi.org/10.1016/j.engappai.2010.06.002.
Wei, M., Tian, X., Geng, J., Zhang, M., 2020. A Balanced task-personnel Matching Method for Aircraft Assembly Coordination Planning. Xibei Gongye Daxue XuebaoJournal Northwest, 38. Polytech. Univ., pp. 130–138
Wood, S.D., Kieras, D.E., 2002. Modeling human error for experimentation, training, and error-tolerant design. In: Proceedings of the Interservice/Industry Training, Simulation, and Education Conference, pp. 1075–1085.
Wu, B., Wang, J., Cai, H., Shen, Y., Qu, B., Fu, Y., 2024. Research on the analysis method of production safety accidents based on accident event causal association diagram. Min. Metall. Explor. https://doi.org/10.1007/s42461-024-01088-x.
Wu, B., Yan, X., Wang, Y., Soares, C.G., 2017. An evidential reasoning-based CREAM to human reliability analysis in maritime accident process. Risk Anal. 37, 1936–1957.
Wu, B., Yip, T., Yan, X., Soares, C., 2021. Review of techniques and challenges of human and organizational factors analysis in maritime transportation. Reliab. Eng. Syst. Saf. 219, 108249. https://doi.org/10.1016/j.ress.2021.108249.
Xu, Z., Shang, S., Su, X., Qian, H., Pan, X., 2023. Handling dependencies among performance shaping factors in SPAR-H through DEMATEL method. Nucl. Eng. Technol. https://doi.org/10.1016/j.net.2023.04.017.
Xue, J., Fang, J., Wu, J., Pang, S., Zheng, N., 2024. Collaborative multiple autonomous systems. Strateg. Study Chin. Acad. Eng. 26, 101–116.

Yang, C., Zhu, Y., Chen, Y., 2021. A review of human–machine cooperation in the robotics domain. IEEE Trans. Hum.-Mach. Syst. 52, 12–25.

Yoon, Y.S., Ham, D.-H., Yoon, W., 2017. A new approach to analysing human-related accidents by combined use of HFACS and activity theory-based method. Cognit. Technol. Work 19, 759–783. https://doi.org/10.1007/s10111-017-0433-3.

Yoon, Y.S., Ham, D.-H., Yoon, W., 2016. Application of activity theory to analysis of human-related accidents: method and case studies. Reliab. Eng. Syst. Saf. 150, 22–34. https://doi.org/10.1016/j.ress.2016.01.013.

Yu, R., Xu, X., Peng, S., 2025. The impact of individual AI proficiency on human–agent collaboration: higher sensitivity to discern the comprehension ability of intelligent agents for users with higher AI proficiency levels. Int. J. Ind. Ergon. 107, 103745. https://doi.org/10.1016/j.ergon.2025.103745.

Zhang, J., Jia, Q., Li, S., Zhang, S., Chen, G., 2026. Intelligent prediction of ergonomics evaluation metrics in human-AI collaboration based on machine learning. Int. J. Ind. Ergon. 111, 103851. https://doi.org/10.1016/j.ergon.2025.103851.

Zhang, L., Liu, J., Zou, Y., 2020. Framework of performance shaping factors for human reliability analysis of digitized nuclear power plants. In: International Conference on man-machine-environment System Engineering. Springer, pp. 1013–1020.

Zhong, S., Zhang, K., Wang, D., Zhang, H., 2021. Shedding light on "Black Box" machine learning models for predicting the reactivity of HO radicals toward organic compounds. Chem. Eng. J. 405, 126627.

Zhou, J., Lei, Y., 2020. A slim integrated with empirical study and network analysis for human error assessment in the railway driving process. Reliab. Eng. Syst. Saf. 204, 107148. https://doi.org/10.1016/j.ress.2020.107148.

Zhou, J., Tu, R., Xiao, H., 2022. Large-scale group decision-making to facilitate inter-rater reliability of human-factors analysis for the railway system. Reliab. Eng. Syst. Saf. 228, 108806. https://doi.org/10.1016/j.ress.2022.108806.

Zhou, J.-L., Guo, Z.-M., 2024. A hybrid SNN-STLSTM method for human error assessment in the high-speed railway system. Adv. Eng. Inform. 60, 102408. https://doi.org/10.1016/j.aei.2024.102408.