INL/JOU-24-76319-Revision-0

The role of AI in detecting and mitigating human errors in safety-critical industries: A review

November 2024

Ezgi Gursel, Mahboubeh Madadi, Jamie B. Coble, Vaibhav Yadav, Ronald Laurids Boring PhD, Anahita Khojandi, Vivek Agarwal

Changing the World's Energy Future

INL is a U.S. Department of Energy National Laboratory operated by Battelle Energy Alliance, LLC

INL/JOU-24-76319-Revision-0

The role of AI in detecting and mitigating human errors in safety-critical industries: A review

Ezgi Gursel, Mahboubeh Madadi, Jamie B. Coble, Vaibhav Yadav, Ronald Laurids Boring PhD, Anahita Khojandi, Vivek Agarwal

November 2024

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Highlights

The Role of AI in Detecting and Mitigating Human Errors in Safety-Critical Industries: A Review

Ezgi Gursel, Mahboubeh Madadi, Jamie Baalis Coble, Vivek Agarwal, Vaibhav Yadav, Ronald L. Boring, Anahita Khojandi

- This review considers AI/ML for human errors in safety-critical industries.
- Studies are categorized into descriptive, predictive, prescriptive, generative modeling types.
- Findings suggest AI/ML can be utilized to help with human error in safety-critical industries.

The Role of AI in Detecting and Mitigating Human Errors in Safety-Critical Industries: A Review

Ezgi Gursel[a], Mahboubeh Madadi[b], Jamie Baalis Coble[c], Vivek Agarwal[d], Vaibhav Yadav[d], Ronald L. Boring[d], Anahita Khojandi[a,1]

[a]*Department of Industrial and Systems Engineering, University of Tennessee, Knoxville, 37996, TN, USA*
[b]*Department of Marketing and Business Analytics, University of Tennessee, San Jose, 95192, CA, USA*
[c]*Department of Nuclear Engineering, University of Tennessee, Knoxville, 37996, TN, USA* [d]*Idaho National Laboratory, PO Box 1625, Idaho Falls, 83415, ID, USA*

Abstract

For safety-critical industries, human error (HE) presents continual risks to system productivity, reliability and safety. Artificial intelligence (AI) and machine learning (ML) methods have emerged as promising approaches to understand, categorize and mitigate the risk of HE in safety-critical industries. This review offers an examination of the current landscape regarding the utilization of AI/ML with regards to HE in safety-critical industries, categorizing literature into descriptive modeling, predictive modeling, prescriptive modeling, and generative modeling techniques. Additionally, the review aims to provide insights regarding themes in literature, challenges, and future research directions. Findings of the review suggest that AI/ML methods can prove useful in addressing the HE problem across safety-critical industries.

Keywords: safety-critical industries, human error, artificial intelligence, machine learning

1. Introduction and Background

Human error (HE) is a broadly defined term that refers to human performance errors which may adversely impact system safety, performance, and human health. HE plays a critical role in virtually every industry, particularly safety-critical industries. Safety-critical industries are those that involve a high degree of human dependency with a paramount importance on safety, in which the failure of such systems can threaten the safety of humans, or result in consequential environmental or property damage [1]. By this definition, many

[1] Corresponding author
Email address: khojandi@utk.edu (Anahita Khojandi)

important industries and systems central to an efficient society fall within the category of 'safety-critical,' including, but not limited to, nuclear energy, aviation (all aerospace fields), transportation, maritime, healthcare, oil and gas, and chemical process industries. Due to its significance and complexity, the study of HE spans multiple disciplines and research areas, including psychology [2], cognitive science [3], industrial engineering [4] and management [5], safety engineering, ergonomics, and human-computer interaction. While there is extensive research on HE, literature is often either domain-specific - considering the HE factors and solutions within those specific contexts - or too broad, taking a more general approach to HE. This study aims to bridge the gap between domain-specific and broad literature on HE to achieve both breadth and depth and to provide a cross-cutting perspective on how data-driven computational approaches can be used to detect, reduce or mitigate HE across different safety-critical industries. Specifically, in our review, we analyze how artificial intelligence (AI) techniques, including machine learning (ML) and reinforcement learning (RL), are used to detect and reduce the risk of HE in safety-critical industries. While other studies, such as [6, 7, 8] have also done reviews on HE or the use of analytics and modeling [9, 10], most existing studies tend to be either field specific (such as focusing on railway, maritime, or healthcare specifically) or do not specifically consider studies with a focus on HE. [11] presents a thorough review on the use of AI in safety-critical systems; [12] discusses the applications of ML as how to relates to reliability engineering and safety applications, and [13] present a bibliometric review on the use of ML for occupational accident analysis. The emphasis on our review is on AI as how it relates to HE in safety-critical systems. To the best of our knowledge, this review is the first review on the use of AI/ML for HE in safety-critical industries, particularly focusing on studies from the last 10 years. The research questions this review aims to address are as follows:

- *R1: How are AI/ML models currently being applied to detect and mitigate human error in safety-critical industries?*
- *R2: What are the differences in application of modeling types among descriptive, predictive, prescriptive and generative modeling in addressing human error across safety-critical industries?*
- *R3: What are the limitations and challenges for the use of AI/ML models in safety-critical industries for human error?*
- *R4: What key insights can be identified in recent literature with respect to the application of AI/ML models for safety-critical industries and the modeling types utilized?*

The remainder of this paper is structured as follows. The rest of Section 1 provides a background and discusses definitions, importance, and classifications of HE. Section 2 introduces the types of AI modeling techniques, namely descriptive modeling, predictive modeling, prescriptive modeling, and generative modeling. Section 3 presents the methodology and a review of relevant literature on the applications of AI to detect or reduce

HE in various safety-critical industries by grouping them into descriptive, predictive, prescriptive modeling, and generative modeling. Section 4 discusses opportunities for advancing AI models, through digital twins and human-in-the-loop. Section 5 provides an overview of challenges that can be encountered when using AI in HE and provides approaches that can help address these challenges. Finally, Section 6 and 7 offers insights regarding the state of the literature and concludes the review.

1.1. Human Error Definition

HE is a broad concept that has many definitions that can vary across studies, with some scholars such as [14] preferring to use alternative terms such as 'erroneous action' instead to specifically emphasize the *action* rather than the *cause* of undesirable events. In a broader sense, HE is defined as a discrepancy or deviance between the action taken by the human and the intended action [15]. [8] defines it as any failure to accomplish a particular task, which can disrupt the normal course of operations or result in property damage or equipment failure. HE can take on many forms, including slips and lapses (errors from unplanned actions), mistakes (errors of judgement), and deliberate violations of established rules and procedures [16]. Both unintentional acts and intentional acts that do not achieve the intended outcome can be considered HE [17, 18]. However, the classification of an act as HE may vary depending on the nature of the error and the original intention, such as malevolence. A variety of factors can contribute to HE, including inadequate training or skill of personnel, poor equipment or job design, insufficient layout or ergonomics, personal stressors, use of improper tools, poorly written maintenance and operation procedures, and improper environmental conditions, such as poor lighting of the work area or high noise levels, to name a few [18, 19].

1.2. Importance of HE

Despite the increased automation in many safety-critical industries, humans, and thus HE, continue to play a significant role in system reliability and performance. The impact of HEs has been widely studied across many industries, such as healthcare [20]. Although many HEs have negligible impact on system cost and operations, the impacts of HE can sometimes be far-ranging and consequential, including significant financial losses and damage to the national and global economy, loss of life, irreversible environmental degradation, and widespread public health damage, as detailed in studies like [21]. For example, detailed retrospective analyses of major system failures and accidents like the Texas City Disaster, Piper Alpha, Bhopal Gas Leak, Texas City Refinery, Deepwater Horizon, Chevron, Chernobyl, and Fukushima have all found evidence of HE contributing to the magnitude of the incident [22].

Furthermore, 20-30% of system failures across all industries can be attributed to human actions [23]. For complex systems such as aircraft, ships, and nuclear power plants (NPPs), it is estimated that 70-90% of all accidents occur either as a direct or indirect consequence of HE [24]. Many studies have investigated the quantitative impact of HE in accidents across safety-critical industries. Although the estimated percentage differs across various industries, studies and depending on how HE is defined or calculated, the statistics remain

nonetheless significant [25]. For example, in the aviation industry, HE has been found responsible for 75% of all accidents. The statistics are similarly significant across maritime operations (75-96% of causalities) [26], process industries (80%) [22], oil and gas (70%) [27], and NPPs (90%), where HE was found to be an aggravator of the top 10 nuclear accidents in recent history
[28].

1.3. HE Classification

As discussed in Section 1.1, the term HE encompasses a large range of human-induced errors that can adversely impact the performance, safety, or reliability of a system. Given the broad definition of HE, classification methods have been developed to systematically identify and address HEs. Although there are many different classification taxonomies present in the literature, in this section, we expand on the definitions provided in Section 1.1 by exploring three common groups of HE classification taxonomies, namely behavior-oriented, task-oriented, and system-oriented schemes [24]. It is important to note that these classifications are not mutually exclusive and HE is often a result of a combination of many factors, and can be expressed in various taxonomies. While these classifications provide a conceptual framework for HE, they are beyond the main focus of our review, and are instead included to contextualize the discussion on HE and its impact on safety-critical systems.

Behavior-Oriented: Behavior-oriented schemes categorize HE based on human behavior. Although many schemes in literature can fall under this category, [29], [30], and [18] in particular are considered to be among the more popular and fundamental works [31]. According to [29], HE can be generally categorized into slips and mistakes, where slips are errors resulting from unintentional actions whereas mistakes result from intentional actions. For example, when a worker knows what to do yet still makes an error, it is categorized as a *slip*, whereas it is considered a *mistake* when the worker intentionally chooses the wrong procedure to complete the task. [18] expands the taxonomy from [29] and introduces the category of lapses. While slips and lapses are both unintentional errors which occur due to unplanned actions, mistakes are intentional errors that occur due to errors in judgement.

Task-Oriented: Task-oriented schemes define and categorize HE based on specific tasks and across specific domains. Examples of studies utilizing task-oriented schemes address the information transfer problem [32], distraction [33] and categorization [34]. The goal of behavior-oriented schemes is to classify behavior independent of the task [24]. However, behavior-oriented schemes may also be considered task-oriented, if behavior is particularly influenced by the nature of the task and its requirements.

System-Oriented: System-oriented schemes are more comprehensive schemes, in that they extend their scope to a wider range of tasks within a particular domain [24]. System-oriented schemes can be considered a combination of behavior-oriented and task-oriented schemes, in that they consider the interaction between human actions and behavior, the environment, and tasks. For example, in [35], errors in an aircraft simulation study were categorized into categories such as related to navigation, communication problems, systems errors, among others. Some issues can be considered as arising from a combination of the

task-related and behavior-related elements, such as the pilot failing to deploy the automatic pilot when it could have been helpful - a possible slip or lapse.

In addition to these classifications, HE can be broadly categorized based on visibility, into active and latent errors [36]. Active failures are the errors that directly and immediately contribute to an adverse event, often committed by front-line operators or individuals otherwise directly involved with a task. Latent conditions, on the other hand, are those errors that can lie dormant within the system for an extended period of time, such as long-term problems in communication, organizational structure, policies, or management.

1.4. Industry-Specific HE Classification

In addition to general HE classification schemes discussed above, many industries have recognized the unique complexities inherent in their specific domains and have adapted standards and classification systems to help identify and mitigate HE. Although HE classification can vary across industries, classification often shares similarities with regards to human behavior and task requirements, and many classification systems are based of the granularity of task, as discussed in Section 1.3. For example, the Human Factors Analysis and Classification System (HFACS) [37] was based on [36] to analyze HE and the underlying factors of accidents in the aviation industry. Per the HFACS framework, HE is categorized into four levels: unsafe acts of operators, preconditions of unsafe behavior, unsafe supervision, and organizational influences. HFACS is a popular and versatile HE framework and has been adapted across many safety-critical industries, such as maritime [38], healthcare [39], nuclear [40], and oil and gas [41]. Another similar framework based on HFACS, titled Human Error Awareness Training (HEAT), was developed in [42] with application for the construction industry. Under the HEAT model, HEs are categorized as organizational influences, supervisory influences, preconditions, and acts/events [42]. Another industry-specific model, AGAPE-ET, considers human error analysis methodology for emergency tasks in a NPP [43].

2. Modeling Types

AI models are algorithms or computer programs that find patterns in data. In the context of AI, four categories of modeling techniques can be discussed: Descriptive modeling (DM), predictive modeling (PM), prescriptive modeling (PSM), and increasingly, generative modeling (GM) [44]. These categories can be generally considered to be progressive [44], based on increasing levels of modeling complexity and added decision-making value [45, 46]. While many reviews often group AI/ML techniques into more conventional classifications of classification, regression, and clustering, our approach emphasizes the levels of decision-making support these models provide. This focus allows us to explore the purpose and outcomes of these models, rather than their technical methodologies specifically. By using this specific classification scheme, we aim to highlight the progressive complexity of these models and their respective contributions to decision-making in the context of reducing HE in safety-critical contexts.

DM focuses on analyzing historical data to understand patterns and relationships, in order to provide insights as to what has happened in the past or is currently happening

regarding HE. In the context of HE analysis, DM helps to summarize and analyze existing data in a meaningful way, allowing for error identification and classification and providing industry personnel with a comprehensive understanding of the nature and characteristics of HE incidents. PM goes beyond describing past or current events and instead aims to make predictions on future events based on past events. It leverages historical data, often obtained from DMs, to identify patterns and build models that can forecast future trends or outcomes. PM can assist in proactive decision-making, risk assessment, and development of HE mitigation strategies. PSM expands the predictive analysis and aims to provide reasonable actions or recommendations for managing and mitigating HE. PSM considers different constraints, objectives, and available courses of actions to determine best possible decisions in preventing, identifying and mitigating HE incidents. PSMs can extend the insights obtained in DMs and PMs by providing actionable recommendations, and even decision automation through implementation. Thus, PSMs complement descriptive and predictive modeling by bridging the gap between data analysis and decision implementation.

It is possible for these modeling types to overlap. For example, a model that identifies an error (DM) can also provide the probability of its occurrence (PM). Similarly, a model may build upon DM and PM for a PSM that aids in operators' decision-making abilities. Additionally, the lines between the modeling types can be blurry in application, so it is possible that a study may be used in different modeling applications than generally considered.

In addition, a fourth category, generative modeling (GM), has gained significant interest and relevance, especially with the rise of generative AI models and large language models (LLMs). Unlike the other modeling types which focus on extracting and applying knowledge from data, GMs create new data by learning and sampling from the probability distribution of the existing datasets [46].

2.1. Descriptive Modeling (DM)

DM is an approach within AI that makes inferences about the past by using historical records [47]. DM establishes a link between past events and present outcomes. The analysis of historical data with DM can provide insights into the factors contributing to HE in an industry or organization. The primary objective of DM with regards to HE is to answer the question of "why did the error occur?" [48]. This allows for a better understanding of HE prevention and mitigation measures.

The concept of DM has been studied extensively and various DM algorithms have been applied in HE detection and mitigation studies. For example, clustering methods, such as *k*-means clustering and associations [49] can help group similar data points and explore relationships that may not be immediately apparent. Additionally, regression analysis can fit under the DM umbrella - if the goal is exploratory analysis between dependent and independent variables, rather than prediction [50] - as well as ML dimensionality reduction techniques such as principal component analysis (PCA). Decision trees can also be considered DM algorithms depending on the nature of the task. For example, decision trees can be employed to provide insights into historical data.

2.2. Predictive Modeling (PM)

PM is the process of using historical data to create, process, and validate a model that can be used to forecast future outcomes. Unlike DMs, PMs aim to offer insights about the future from past events. Although the main objective of PM is not to identify errors, error identification often serves as an initial step in the assessment process. Prior to assessing probabilities, different DM techniques can be employed to identify and classify HE. In some cases, the same model can be considered both a DM and PM - used for both error identification and probability assessment.

Many classification and regression algorithms can fit under the umbrella of PM [46] if the goal is to produce predictions of future events [50]. For example, different regression analysis algorithms, such as linear regression, logistic regression, neural network (NN) regression, can be used as PMs to assess error probability by establishing a correlation between a dependent variable, such as the occurrence or the likelihood of HEs, and independent variables, such as task characteristics like task complexity [51] or various human factors, such as fatigue [52]. Regression algorithms can also be combined with various traditional human reliability analysis (HRA) methods such as Technique for Human Error Rate Prediction (THERP) and Human Cognitive Reliability Correlation (HCR) to assess the probability of an error [53]. For instance, the Human Reliability data EXtraction (HuREX) framework, developed to collect HRA data from NPP simulators and events, utilizes various data analysis techniques, including logistic regression, to provide HE probability estimates [54]. Additional commonly used PMs include pattern recognition models such as artificial neural networks (ANNs), Bayesian Networks (BNs) and support vector machines (SVMs), classification models such as decision trees, and nearest neighbor search [9, 49].

2.3. Prescriptive Modeling (PSM)

PSM is an approach that focuses on providing adaptable and real-time sequences of suggested actions based on situational conditions [55]. It aims to build upon DM and PM to answer the questions about “what should be done?” and “why should this be done?” [55]. In comparison with DM and PM, which are both well-established in literature, PSM is a relatively new area of research that has been gaining added interest, particularly in the context of the Internet of Things Revolution. PSM is a real-time sequential decision-making model [56], with the goal of intervening at the *right time* in order to take the *right decision*. The goal of PSM is to deliver actionable recommendations and suggest the optimal decision. Various modeling techniques can fall under the umbrella of PSM, as noted in [48], and can include Markov decision processes (MDP), genetic and greedy algorithms, and optimization models, such as linear programming, among others. RL is another commonly utilized ML technique that can be used in PSM [57]. RL is concerned with how agents should behave in an environment in order to maximize the cumulative reward received over time [58]. RL’s ability to adapt to changing circumstances makes it particularly well-suited as a PSM in dynamic and evolving environments.

2.4. Generative Modeling (GM)

GM is an modeling type that has gained increased popularity, especially with the rise of large language models (LLMs). Generative models include techniques such as generative adversarial networks (GANs), autoregressive models, diffusion models, and variational autoencoders (VAEs) [59, 60]. Depending on the problem statement, GMs can be utilized for clustering, regression, or classification purposes [61]. The primary objective of these models is to model the underlying distribution of the dataset and generate new points representative of that distribution [46]. In the context of HE, GM offers several valuable applications. Namely, generative models can be used for synthetic data creation for training purposes, which can be valuable in settings where simulating anomalous scenarios or obtaining data representative of HE can be difficult. The process of curating and labeling real datasets required for AI models can itself be subject to HE, which can negatively impact the accuracy and reliability of the resulting models, so synthetic datasets created by GMs can also address this predicament [62]. Additionally, GMs can be employed for the purpose of anomaly detection. In adversarial scenarios in safety-critical settings, anomaly detection models can help human operators with incident diagnosis and mitigation tasks [63]. LLMs have also found applications in safety-critical industries, where they are being utilized to analyze large sets of textual reports, such as accident reports, and generate insights into the data [64]. LLMs' ability in generating human-like text has been considered for developing safety management and HE prevention procedures [64].

3. Methodology and AI/ML Applications for HE

As discussed in Sections 1.3 and 1.4, classification of HE varies across or even within industries. In this section, we break down selected literature as DM, PM, PSM, and GM. It is important to note that schemes and algorithms are not mutually exclusive and a study may also be considered under different or multiple categories simultaneously, even though it may not be explicitly listed.

While our review is not an exhaustive list of all relevant literature, our main aim in our selection is to explore the diverse array of studies conducted across various safety-critical industries and systems to explore similarities in methods and applications. In particular, we consider studies within the last ten years of publication (2013-2024, to date). We choose this range specifically to study to gain insights into the state-of-the-art and the recent trends in the field of HE detection and mitigation across safety-critical industries. Although HE is an active field of research historically, the use of AI/ML models have only recently began to gain interest. Our methodology involved a systematic review across Google Scholar, a scholarly search engine, and multiple academic and scientific databases like IEEE Xplore and Scopus. We used a combination of keywords and Boolean operators to refine our search. Examples of keywords considered include a combination of terms such as "human error/factor," industry name (e.g. aviation, maritime, nuclear, medical), "safety-critical," "algorithm," "machine learning," "descriptive," "predictive," "prescriptive," "generative." Following the initial keyword search, we applied a manual screening process to further refine our selection. This

involved reviewing the abstracts, and when necessary, full texts of documents to exclude those studies that did not fit within the scope of this review. For each selected paper, we reviewed the method and the modeling types utilized. We also examined that citations of the selected literature to check for further relevant studies that may not be captured through keyword searches alone, as well as any related derivative works. For example, we notice that many relevant studies do not include the term "safety-critical" and may not necessary include the terms "human error" but rather related terms like operator or personnel error, or not even mention human error, but still fit within our scope. In total, 58 studies are included.

Figure 1 shows a stacked bar chart of the literature in this study, breaking them down into industry and modeling type. Figure 2 shows the distribution of the selected studies into modeling type. Tables 1, 2, 3, and 4 categorize the reviewed studies into modeling type, industry, model(s) utilized, major objective of the study, the theme of the paper as it relates to HE, and detection occurrence. In 'detection occurrence,' pre- refers to pre-occurrence, which refers to error detection that happens before the error has taken place, based on what-if scenarios. Opt refers to operational detection, with the goal of working in real-time to catch or alert at an probability of error, based on operator actions. Finally, post- refers to detection post-occurrence, after the error has already occurred.

3.1. Network Visualization

Figure 3 shows a network visualization of the co-occurrence of keywords in the studies considered in this review. The VOSViewer software is used to extract and visualize the cooccurrence networks of key terms from studies included in the review [65]. This visualization serves as a tool to identify key themes and relationships present in the reviewed literature, providing insights into the frequencies and relationships between key terms. The distance between the keywords represents the strength of the relationship of the terms. For example, 'clustering' is closely related to 'human factors,' which highlights the focus on clustering methods for human factor analysis. 'Natural language processing' is also similarly related to 'aviation safety'. The size of the keywords represents the frequency of the keywords. Specifically in this analysis, 'machine learning,' 'reinforcement learning,' and 'human factors' are among the most frequent keywords. The size of these terms suggests the growing importance of AI/ML models in addressing HE.

closely

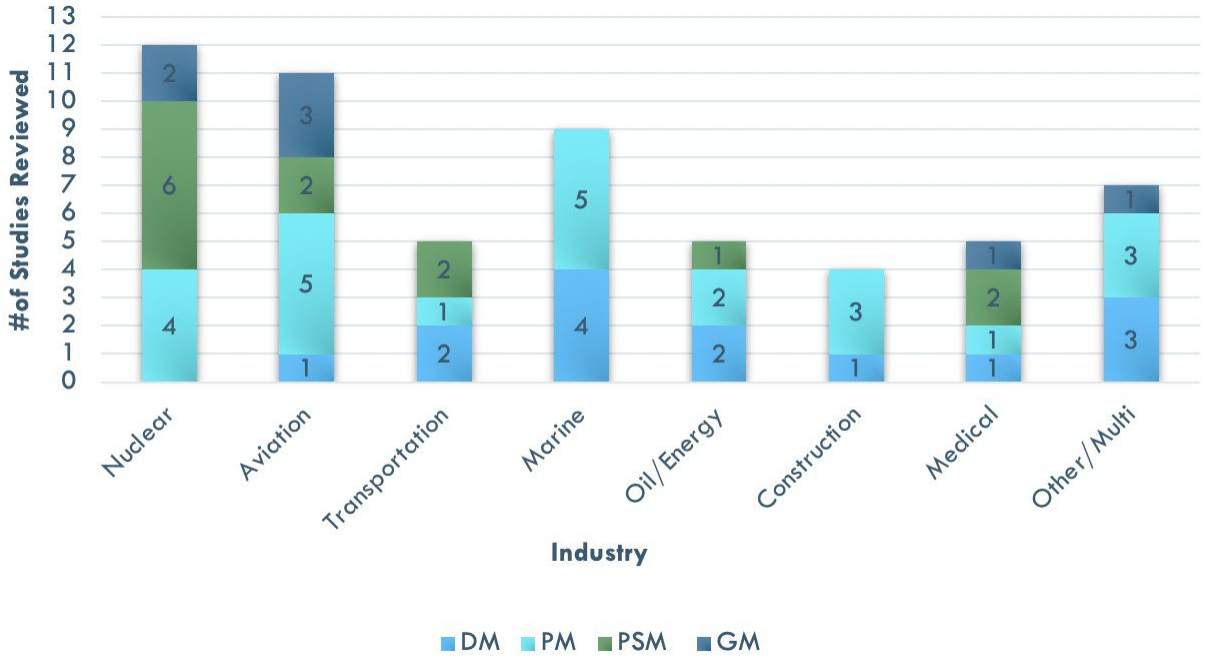

Figure 1: Breakdown of Included Literature into Industry and Type of Modeling (DM: Descriptive Modeling, PM: Predictive Modeling, PSM: Prescriptive Modeling, GM: Generative Modeling), *n* = 58.

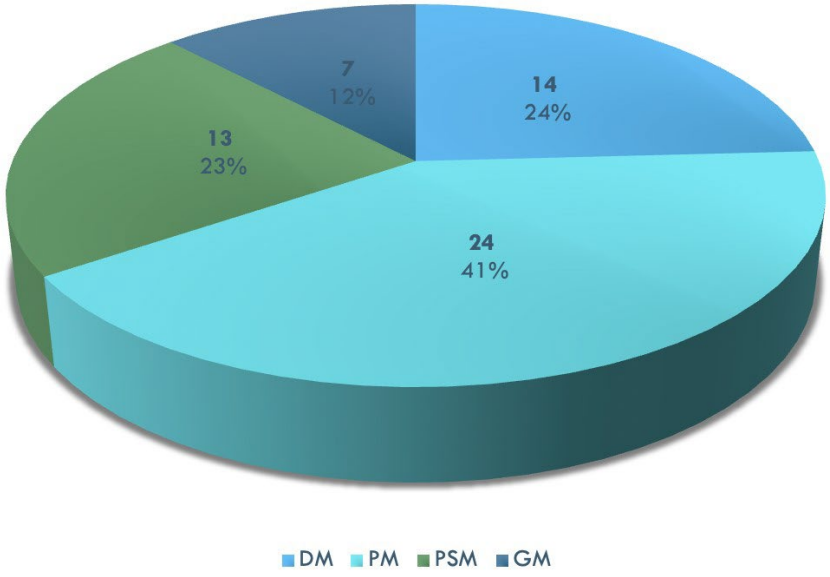

Figure 2: Distribution of Studies according to Type of Modeling (DM: Descriptive Modeling, PM: Predictive Modeling, PSM: Prescriptive Modeling,GM: Generative Modeling), *n* = 58.

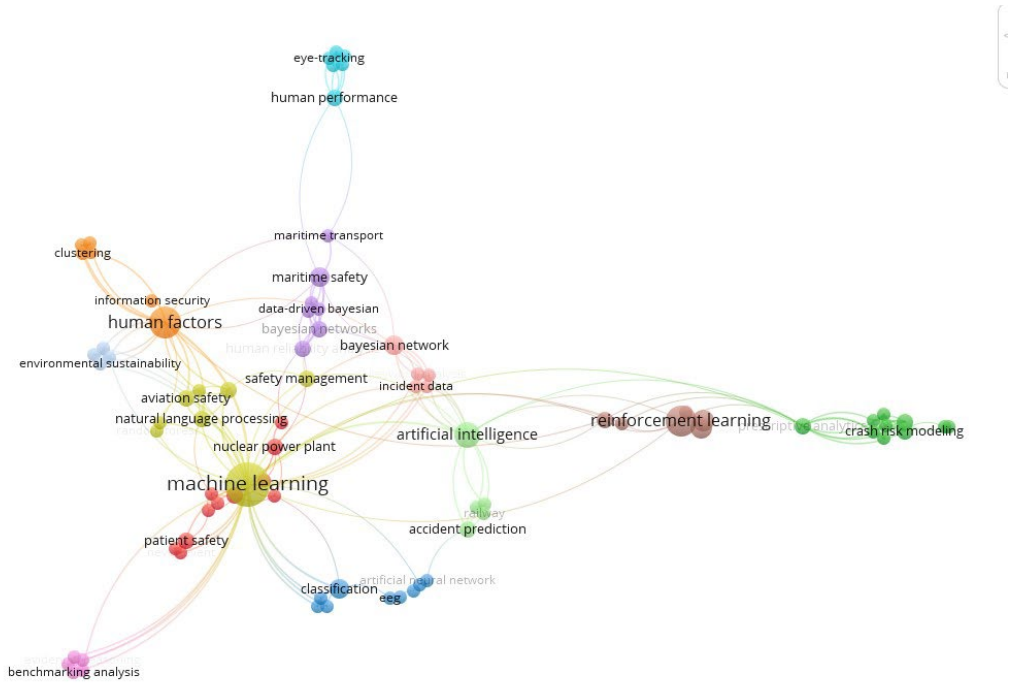

Figure 3: Network Visualization of Studies Reviewed using VOSViewer Software

3.2. Descriptive Modeling

DM studies primarily focus on retrospectively exploring the underlying human factors that contribute to an HE incident, identify causal factors of accidents, or classifying accidents. As descriptive analytics is often the first step in HE analyses, DM techniques often serve as a foundational step in many HE studies. As such, we consider those studies in particular that focus entirely or have a significant element of DM in their methodology. Table 1 presents an overview of the DM studies considered.

Most DM studies identified in this paper fall under the theme of incident/accident analysis or classification [66, 67, 68, 69, 70, 71, 72, 73], in which the goal is to analyze historical incident/accident reports and understand the underlying causes or human factors that contributed to an event. For instance, in [68], construction site incidents from a construction company between 2014 and 2020 are analyzed and patterns are extracted between features and type of incident using a decision tree classifier and the Apriori algorithm used in association rule mining. Similarly, in [72], contributing factors to HEs in surgery are analyzed using random forests (RFs) from historical event reports. Classification methods, particularly decision trees and ensemble tree models like RFs used to prevent overfitting, are widely used as DM and PM techniques. These models are favored in HE analysis given their interpretable structure and ability to consider the relationships between many input variables (e.g. human factors or characteristics of incidents) to a target output [12, 67].

DM studies are particularly prevalent across the transportation industries, with a majority of the studies considered in this area falling under maritime, railway, and aviation,

due, in part, to the availability of relatively larger public incident databases. For example, the goal of both [66] and [67] is to incident analysis and classification, with the underlying goal of understanding factors or patterns of railway accidents. Similarly, [74, 75, 76] all consider historical incident databases from the maritime industry to identify human factors or classify maritime incidents.

While DM studies are not inherently PSMs, they often do include a 'prescriptive' element by recommending specific measures for safety enhancement. Many DM studies provide additional insights for improving safety within the industry. For example, based on the results from the DM on maritime accidents, [76] suggests that improved training and enhanced communication protocols could reduce the occurrence of maritime incidents. The emphasis on communication is also mentioned in [72], among others. [72] suggests that improved communication between surgery teams and adopting a specialized safety approach for the characteristics of each surgery could lessen the likelihood of preventable medical error in surgeries.

3.3. Predictive Modeling

As discussed in Section 2.2, studies that address HE probability calculations, in which the likelihood of HE is calculated using available datasets or past incident/accident data, can fall under the umbrella of PM. Generally, PM in the context of HE focuses on three main categories: (1) predicting HE or accident probability, (2) predicting the outcome of an incident, and (3) predicting accident severity. Prediction of HE risk, severity, or category can be done retroactively, i.e. after an incident or HE has occurred, before an incident based on what-if scenarios, or in fewer cases, operative, based on the actions of the human operator. Table 2 presents an overview. Studies such as [80, 81, 82, 83, 84] can fall under the umbrella of HE risk prediction. Bayesian networks are popular methods for this task. Bayesian networks are particularly used in the field of HE analysis, given their ability to handle complex interdependencies in human factors data points and ability to inherently capture uncertainty [85]. For example, [81], a Bayesian network model based on an accident database from safety-critical and other high-technology industries is proposed to calculate probability estimates of errors of cognition and execution, or human failure probabilities. The use of text mining and natural language processing is also utilized in PM and HE analysis. In [86], the authors develop a PM to predict human factors from aviation incident reports. The HFACS-ML model is developed with the goal of extracting HE factors for ML applications. Related to HE risk analysis, in [87], the authors develop a maritime vigilance assessment model based on Shallow NN. Vigilance assessment is especially important for operators in safety-critical roles such as traffic controllers, whose tasks required continuously elevated attention levels. The goal of the study is to predict the onlookers' reaction time base on their gaze patterns. Results of the Shallow NN model are compared against other algorithms such as bagged trees, decision tree and SVM. While the Shallow ANN provides R2 near 0.8, the decision tree and bagged trees provide the poorest performance of 0.053 and 0.1, respectively, likely due to overfit.

Studies such as [88, 89, 90, 91] can be included under the accident risk prediction category, in which the goal is to quantify the risk of an accident based on human factor traits

or environmental conditions. In [92], the authors develop an accident prediction model for repair and maintenance accidents in oil refineries. In [91], a PM is build based on RFs and multi-class SVMs using Boolean kernel to predict the type of maritime accident (e.g.

			Descriptive Modeling				
					Detection		**Occurrence**
Study	**Industry**	**Model**	**Objective**	**Theme**	**Pre-**	**Opt**	**Post**
Cheng et al. (2013) [77]	Petrochemical	classification & regression tree	Analyze causes of major petrochemical accidents	identify and categorize causal themes in accident data			X
Youn et al. (2018) [78]	Maritime	decision tree, SVM, KNN	Use optical sensors to simulate lookout activities and develop a classification model to monitor navigators' lookout actions	classification			X
Chen & Yu (2018) [79]	Aviation	PCA, regression analysis	Investigate the relationship between intervention strategy and unsafe acts	risk mitigation			X
Hua et al. (2019) [66]	Railway	text mining, NLP	Extract accident risk factors from text reports	incident/accident analysis/classification			X
Alawad et al. (2019)[67]	Railway	decision tree	Classify accidents and patterns Predict characteristics of accidents	incident/accident analysis/classification			X
Fan et al. (2020) [74]	Maritime	BN	Risk assessment for human factors contributing to maritime accidents	human reliability analysis			X
Chen et al. (2020) [75]	Maritime	association rules	Finding human factor cause chain of ship accidents	human factor classification			X
Ugur et al. (2021) [68]	Construction	decision tree, association rule mining	Analyze construction site incident reports to understand incident causes & relationships to other factors (incident type, severity).	incident/accident analysis/classification			X
Sattari et al. (2021) [69]	Oil/Gas	BN	Label and classify incident reports	incident/accident analysis/classification			X
Paolo et al. (2021) [76]	Maritime	semisupervised hierarchical methods	Classify sea accidents from database using clustering method to analyze contributing factors and contribution of HEs to accident data	identify and categorize causal themes in accident data			X
Morais et al. (2022) [70]	Multi	SVM	Develop a `virtual human factors classifier' to automatically read and classify accident reports based on individual, organizational, or technological factors	incident/accident analysis/classification			X

Ouache et al. (2022) [71]	Fire Safety	ANN, classification	Identify HE factors contributing to fires	incident/accident analysis/classification	X
Arad et al. (2023) [72]	Medical	RF	Analyze observations and RCAs to identify contributing factors to the occurrence of preventable errors in surgery	incident/accident analysis/classification	X
Nallathambi et al. (2023) [73]	Fireworks	rule-based	Investigate the impact of human factors contributing to HE	incident/accident analysis /classification	X

Table 1: Overview of DM studies included in the review (organized by publication year).

collision, hull fire, etc.) based on the presence of various human factor features. Then, feature ranking is applied to rank the influence of the human factor features. In [93], an SVM and an ANN are optimized using genetic algorithm (GA) and particle swarm optimization (PSO) to predict occupational accidents using a steel mill case study. Accident data is categorized as injury, near miss and property damage. The best performing classification model, PSO-SVM, is used to generate rules on the contributing factors of accidents. Similar accident outcome prediction studies include [94], in which RFs and stochastic gradient tree boosting is utilized to predict injury type, energy type and body part affected from construction injury reports. In practice, the user inputs the work package and observations about the job site, and the model predicts injury type, energy type (e.g. biological, chemical, gravity, mechanical, etc.), and body part that would be most likely to be affected should an accident occur.

Accident severity studies, such as [95, 96, 97, 98] aim to quantify the severity of an accident based on its characteristics. In [99], association rule mining is used to find patterns in accident dataset and the severity of the incident. Various time series forecasting models, including linear regression (LR), Gaussian processes (GP), MultiLayer perceptron (MP), and Sequential Minimum Optimization Regression (SMOreg) are utilized to predict one year of aviation accident trends. The findings from these prediction studies contributes to a better understanding of the impact of HE and human factors and provide insights into mitigation and prevention strategies.

In addition to these categories, in [100], the authors develop a model based on LSTM that can predict the trend of key plant parameters following operator actions in emergency situations. The model can be considered an operator support system that can be utilized to allow the operator to check intentions with the predictions of the model to prevent HE. Although the study is formulated as a prediction model, the practical development of this model into a diagnostic and error mitigation tool can also classify it as PSM.

3.4. Prescriptive Modeling

Since PSMs are often build upon DMs and PMs, PSM studies are rarely standalone PSM. Additionally, certain studies not considered to be PSM although not PSM at the current stage, can have PSM elements. PSM studies are prevalent in the energy and transportation and aviation industries. Healthcare is also well-suited for PSMs due to its complex, time-critical and personalized nature, and the large scope of possible diagnoses and treatments [105]. Table 3 presents an overview of the studies.

In [106], the authors consider how a combination of PM and PSM methods can be utilized in crash risk minimization in the transportation industry. PMs such as logistic regression, Poisson regression, NNs, and XGBoost are utilized to predict the crash probability in a simulated example. *K*-shortest path is used to find the shortest routes in an example route network and the results from the PMs are used to rank the suggested routes according to crash risk. In [107], an RL model that uses historical multi-modal railway data is proposed to predict causes of railway accidents, and uses to develop accident prevention strategies.

PSM research largely revolves around developing models that actively provide recommendations and suggested actions, thus the vast majority of PSM studies fall under the operator or decision support system category. The nuclear industry makes extensive use of PSMs with the goal of aiding operators in their tasks, such as NPP heat-up operations [108] and preventing HE. For example, [109] propose an operator support system based on Petri

		Predictive Modeling					
					Detection	**Occurrence**	
Study	**Industry**	**Model**	**Objective**	**Theme**	**Pre-**	**Opt**	**Post**
Cai et al.(2013) [80]	Oil & Gas	BN	Quantify human reliability of offshore blowouts	HE risk analysis			X
Tixier et al.(2016) [94]	Construction	RF, stochastic gradient tree boosting	Develop a PM that can predict the injury type, energy type and body part affected	accident prediction	X		
Burnett & Si (2017) [88]	Aviation	decision tree, *k*-nearest neighbors, SVM, ANN	Predict conditions that can lead to accidents	accident/risk prediction			X
Madeira et al. (2017) [86]	Aviation	NLP, label spreading, SVM	Classify human factor categories from incident reports	human factor classification			X
Morais et al.(2018) [81]	Multi	BN	Predict HE probability in various sectors	HE prediction	X		
Liao et al. (2018) [82]	Construction	BN	Quantify the impact of improper work environment on HE probability	HE probability prediction			X
Shao-Yu et al. (2018) [89]	Aviation	fuzzy clustering, back propagation NN	Predict the affect of pilot age on flight accident risk	accident/risk prediction			X
Zaranezhad et al. (2019) [92]	Oil	ANN, fuzzy, GA, Ant Colony Optimization	Create accident prediction model for repair and maintenance accidents	accident prediction	X		
Sarkar et al.(2019) [93]	Steel	SVM, ANN optimized with GA, Particle Swarm	Predict occupational accident outcomes from accident data (e.g. injury/near miss/property damage)	accident prediction	X		
Wang et al. (2019) [90]	Transportation	RF, Adaboost with decision tree, GBDT, and XGboost	Predict driving risk from crash and violation records	accident/risk prediction	X		
Abesamis et al.(2020) [99]	Aviation	association rule mining time series forecasting	Predict conditions that can lead to accidents	accident prediction	X		
Coraddu et al. (2020) [91]	Maritime	RF, multi-class SVM	Predict accident type based on presence of human factors; rank the influence of different human factor features	accident/risk prediction			X
Suh & Yim (2020) [84]	Nuclear	SVM, bio-signals	Identify worker's fitness for duty status based on bio-signals	HE risk analysis	X		

Zhang et al. (2020) [101]	Nuclear	SVM with bootstrap	Predict operator performance in NPPs using multiple sources of task and physiological information	operator performance prediction			X
Bae et al. (2021) [100]	Nuclear	LSTM	Predict parameter trends after operator actions in NPP emergency situations	operator support system		X	
Zhu et al. (2021) [96]	Construction	logistic regression, decision tree, SVM, Naive Bayes, *k*-nearest neighbor, RF, Multi-Layer Perceptron and AutoML	Predict the severity of construction accidents	accident severity prediction			X
Li et al. (2022) [87]	Maritime	shallow ANN	Assess vigilance levels predict reaction time based on gaze patterns	vigilance assessment		X	
Tamascelli et al.(2022) [95]	Chemical	linear, Deep NN, hybrid wide & deep	Correlate the features of an accident to predict accident severity	accident severity prediction			X
Yan et al.(2022) [83]	Nuclear	back propagation NN	Predict HE probability from eye response, situational awareness, workload	HE probability prediction	X		
Nogueira et al.(2023) [97]	Aviation	RF, ANN, active learning	Predict fatalities based on accident data	accident severity prediction			X
Ganguly et al. (2023) [102]	Medical	NLP, linear SVM, MLP, CNN	Categorize medical error reports	error type prediction			X
Lan et al. (2023) [98]	Maritime	association rule mining, complex network, RF	Predict the severity of ship accidents by understanding correlation of risk factors	accident severity prediction	X		
Fan & Yang (2023) [103]	Maritime	ANN	Conduct human performance measurement and predict operators' experience level based on physiological data	operator experience prediction			X
Fan & Yang (2024) [104]	Maritime	Lasso, BN	Identify critical fatigue-related factors; build a fatigue prediction model	fatigue prediction			X

Table 2: Overview of PM studies included in the review (organized by publication year).

Nets and deep NNs that detect operator error during emergency situations, such as a coolant loss accident, and notifies the operator through a warning system of an deviancy from the procedure. Similarly, [110] and [111] also discuss operator support systems with the goal of alerting operators if there is a presence of HE. [112] proposes an XAI operator diagnostic assistant that can diagnose scenarios and provide rationale behind the diagnosis, with the goal of supporting NPP operators in their diagnosis tasks during abnormal operating conditions.

In healthcare, decision support systems can come in the form of clinical decision support systems. These systems are intended to provide clinicians with suggestions regarding possible diagnoses and treatment models, which, in turn, can reduce medical error. [105] propose a clinical decision system PSM that is built on a PSO PM. [113] propose an real-time operator assistance system based on POMPD that can aid stroke patients during their daily tasks and provide feedback when they err in their actions.

AI-based PSMs are also utilized in aviation. For example, [114] propose the use of a PSM based on MDP to make decisions as to whether to give the control of unmanned aircraft system from the pilot to the computer system to avoid pilot loss of control. Similarly, [115] propose the use of human-in-the-loop RL model that utilizes computer vision to understand

the pilot's psychological reaction and the simulated flight path to learn flying skills, with the goal that, once sufficiently trained, the RL can fly the aircraft and support the pilot in flight.

3.5. Generative Modeling

Generative models are closely related to the other modeling types, in that they can incorporate elements of data analysis (DM), PM, and suggesting recommendations (PSM). Based on the reviewed literature, GMs are considered across a range of applications for HE in safety-critical industries, including incident analysis, fault detection and diagnosis, and anomaly detection.

The potential of generative language models, such as Bidirectional Encoder Representations from Transformers (BERT) and generative pre-trained transformers (GPT) used in LLMs such as ChatGPT, have been considered in safety-critical literature, especially in relation to incident and human factors analysis [116]. In healthcare and medical fields, for example, GPT models are considered by radiologists and clinicians for diagnosis support. [117] consider the performance of ChatGPT in evaluating and correctly categorizing diagnostic errors in healthcare, a time-intensive task often performed manually by physicians. Results of the study suggest that the ChatGPT model was able to successfully detect 95% of the diagnostic errors and identify the contributing factors of the errors. GPT models are also considered in the aviation and aerospace fields where they are utilized in the accident data analysis. For example, [118] uses a GPT model to classify aviation "decision errors" based on the HFACS framework. While the performance of the GPT model did not meet the expectations of the subject matter expert, especially without prompt engineering, the authors highlight the potential of these models for accident analysis, in particular when used in conjunction with the domain knowledge of experts. In another study [116], authors consider the use of ChatGPT for generating incident summaries based on narratives and later identifying human factor causes based on aviation incident summaries, the results of which are validated against real safety analysts. Relatedly, LLMs are also being considered for the purpose of fault diagnosis. Fault diagnosis in complex systems such as NPPs and chemical processing plants can involve time-sensitive tasks that are prone to HE. [119] consider fine-tuning pre-trained LLM for the purpose of fault diagnosis in complex systems.

Anomaly detection is another important application area of generative models. Similar to fault diagnosis, the manual process of detecting anomalies in safety-critical industries can itself lead to HE through time pressures and cognitive overload. Thus, anomaly detection techniques can either point to outliers that may be resulting from HE or aid human personnel mitigate problems in a timely manner. GANs and GAN-based techniques in particular have emerged as promising anomaly detection tools. The basic GAN architecture consists of two subnetworks that work in adversarial fashion - a generator, which attempts to generate synthetic data representative of the input data, and a discriminator, which attempts to categorize between real and generated data. [120] propose the use of GAN-based anomaly detection model to identify mismatches between automatically collected sensor and manually collected surveillance data in an NPP, suggesting the potential presence of HE. Autoencoders and VAEs can also be employed for the purpose of anomaly detection. [121] employ an autoencoder model to identify pilot error in multivariate time series data. [63]

propose a LSTM-VAE model to detect system and component anomalies in an NPP, under the idea that anomalous accident data will not be reconstructed properly in comparison with the normal input data.

4. Opportunities for Advancing AI Models

Going beyond DM, PM, PSM, and GM, digital twins (DT) and the concept of human-in-the-loop (HITL) present opportunities for advancing AI models through the opportunity for real-time simulation, enriched human-machine collaboration, and improved decision-making. In this section, we discuss the role of digital twins and human-in-the-loop for HE.

4.1. Digital Twins

Digital twins (DTs) can effectively act as a combination of DM, PM, PSM, and GM depending on the task and the complexity of the model simulated [124]. The use of DT in various industries, including for aviation [125, 126], is an emerging, but promising concept. A DT serves as a digital copy of a dynamic system which simulates the real-world outcomes in real time. In DTs, historical data and data collected from the system can be used as a DM to display the real-time status of the system. PMs can be utilized to anticipate critical parameters or forecast trends related to system or component health or potential failure risk. This proactive approach can enable effective decision-making, facilitating timely maintenance activities and preemptively addressing issues before they escalate in failures or unplanned downtime [127]. PSMs can be used to detect critical situations and provide recommendations on corrective measures. DT can model future outcomes and can recommend the best course of actions for any pre-specified outcome. Additionally, a backward simulation can be created based on the historical incident data to find the root cause of an error [128] and reduce the probability of late detected errors. In automated systems especially, DTs can reduce HE by minimizing human input and thus increasing the integration between the virtual and physical system [129], among one of the promises of Industry 4.0 [130]. DTs require a working environment or an actual asset to model, which makes their practical

			Prescriptive Modeling				
					Detection Occurrence		
Study	**Industry**	**Model**	**Objective**	**Theme**	**Pre-**	**Opt**	**Post**
JeanBaptise et al. (2015) [113]	Medical	Partially Observ-able MPD	Monitor patients during ‘activities of daily living’ tasks Provide feedback when HE is detected	operator assistance		X	
Kruse et al. (2019) [114]	Aviation	MDP	Assess risk of pilot loss of control in UAS	operator support system		X	

Vemuru et al. (2019) [115]	Aviation	computer vision, RL	Learn to fly from pilot's reaction and flight path	operator support system		X
Park et al. (2020) [122]	Nuclear	RL	Learn diagnosis of safety functions	operator support system		X
Hu et al. (2020) [106]	Transportation	PMs (logistic regression, Poisson regression, NN, XGBoost), *k*-shortest path	Combine PMs to rank shortest routes from the crash risk	crash risk prediction	X	
Ahn et al.(2020, 2021, 2022) [109, 110, 111]	Nuclear	Colored Petri Nets, NNs	Detect operator error in emergency situations, Alert operator of deviancy from procedure	operator support system		X
Park et al. (2022) [112]	Nuclear	GRU-AE, Light-GBM, SHAP	Diagnose abnormal operating scenarios	operator diagnostic assistant		X
Park et al. (2022) [108]	Nuclear	Deep RL	Perform automatic control during NPP heat-up stage	operator support system		X
Yan et al. (2023) [107]	Railway	RL	Predict causes of railways accidents	accident prediction	X	
Hoyos et al. (2023) [105]	Medical	Fuzzy cognitive maps PSO, GA	Suggest treatments, follow-up, prevention	decision support system	X	
Lui et al. (2024) [123]	Oil & Gas	functional resonance analysis method, RL	Update and optimize emergency procedures in dynamic conditions	decision support system		X

Table 3: Overview of PSM studies included in the review (organized by publication year).

implementation in safety-critical systems particularly challenging, given the infrastructure, deployment costs and issues with data quality and big-data analysis [131]. Regulatory implications, in the nuclear industry in particular, are a large hindrance in the application of DT technologies [132]. Furthermore, a lack of an unified DT architecture across industries exacerbates the gap between theoretical DT concepts and practical applications [133]. Although full-scale DT implementations are limited and largely conceptual in some industries

Generative Modeling							
					Detection Occurrence		
Study	**Industry**	**Model**	**Objective**	**Theme**	**Pre-**	**Opt**	**Post**
Park et al. (2021) [63]	Nuclear	LSTMVAE	Detect system & component anomalies in NPP	anomaly detection		X	
Tikayat et al. (2023) [116]	Aviation	GPT	Generate incident summaries based on aviation safety analyses reports	safety analysis			X
Gursel et al. (2023) [120]	Nuclear	GAN	Correlate NPP data; suggest the presence of HE anomalies	anomaly detection			X
Harada et al. (2024) [117]	Medical	GPT	Detect diagnostic errors; identify contributing factors to errors	error analysis			X
Mural et al. (2024) [121]	Aviation	Autoencoder	Quantify pilot error in multivariate data	anomaly detection			X
Saunders et al. (2024) [118]	Aviation	GPT	Analyzing aviation accidents based on HFACS	accident analyses			X
Zheng et al. (2024) [119]	Multi	LLaMa, GPT	LLM model for fault diagnosis	fault diagnosis		X	

Table 4: Overview of GM studies included in the review (organized by publication year).

due to these aforementioned challenges, various DTs are already used in different capacities in safety-critical industries. The construction industry, for example, actively uses DTs. Further, national labs and research institutions are actively working to bridge the gap between concept and implementation. [132] discuss the potential applications of DTs in the nuclear industry and provide a roadmap for practical implementation given regulatory requirements. Additionally, there are several large-scale research projects focused on DTs in nuclear, such as Advanced Research Projects Agency-Energy (ARPA-E) GEMINA, developing a predictive maintenance DT for the General Electric advanced reactor, the BWRX-300, and Idaho National Lab MAGNET, which attempts to build a DT of an experimental testbed using sensor data [134].

The concept of human digital twins (HDT) has also been explored in recent literature, a review of which is provided in [135]. Similar to DTs, which are digital replicas of physical system, an HDT is a digital replica of a human [135]. Although the application of HDTs is still in its infancy and largely conceptual, an HDT may be used to analyze and understand HEs and the conditions under which errors occur [136]. Additionally, HDTs may be used to evaluate the readiness of personnel [136]. HDTs can be particularly useful in domains such as

healthcare by aiding in personalized diagnoses and real-time patient monitoring. For example, [137] suggest a perioperative (during the period of surgery) HDT that can assess the situation of a patient in real-time and test every possible diagnosis to come to an optimal course of action. Another recent example of a preliminary development of a HDT for a safety-critical industry is the Human Unimodel for Nuclear Technology to Enhance Reliability (HUNTER) [138]. HUNTER, when integrated with NPP simulation models, acts as a 'virtual operator' that can do dynamic and computational HRA, such as calculating HE probabilities and time spent on task. In [139], the authors developed a HDT for process industries that mimics operators' behavior during abnormal operating conditions. In addition to existing challenges with DTs as discussed above, additional concerns with privacy and data security, complexity of modeling human cognition, and regulation with regards to ethics and user data make the development of HDTs particularly challenging [135].

4.2. Human-in-the-Loop and HE

In recent years, safety-critical systems have become increasingly reliant on automation with the intention of improving system safety and efficiency. While automation can offer numerous advantages in improving system performance and reducing HE by automating certain tasks that may be more prone to HE (such as data entry, etc.) and lighten the load of personnel when used at an appropriate level, automation is not a solution to HE [140]. In fact, the increased reliance on automation can aggravate the human factor component of safety [28]. Highly automated systems risk significantly increasing the monitoring workload and lowering freedom for operators to override the system's decisions, given the lack of transparency in AI models [141, 142]. One significant unintended consequence of automation is the out-of-the-loop (OOTL) performance problem, which arises when humans become complacent and lose vigilance [140]. In this case, humans lose situational awareness, impairing their ability to correctly observe or interpret the system in cases where the automation system fails to do so, thereby increasing the likelihood of HE. The OOTL problem can lead to operator delays or failures in system intervention or inadequate mitigation of critical conditions [140, 143].

Recognizing the indispensability of humans in safety-critical systems, the solution to the OOTL problem does not lie in building fully automated systems and removing humans from the system altogether [144]. Increased automation in complex systems necessitates more human oversight since humans are needed to diagnose the system in cases of emergencies or abnormal operating conditions [145, 146] In safety-critical systems, the expertise of humans and trained operators remains unparalleled, and in cases where safety is at stake, automation should work to support, rather than replace humans in the decision-making pipeline [147]. Proper human-machine interaction and a meaningful level of human control can address the problems caused by OOTL. The development of human-in-the-loop (HITL) systems, in particular, can help solve the OOTL problem. A HITL system can leverage a priori human knowledge into the ML model. This is especially valuable in certain safety-critical industries, such as healthcare, where training data can be sparse and a reliable and timely diagnosis is necessary [148, 149]. By incorporating human expertise and decision-making abilities, a HITL system can enhance the performance and reliability of AI models. This

collaborative approach enables ML models to benefit from the strengths of both human personnel and machine intelligence, leading to more effective and safer outcomes in safety-critical systems. HITL systems mitigate the risks associated with over-reliance on automation and ensure that human expertise remains utilized in the system by actively employing humans in the decision-making process. For example, according to [150], even the most advanced autonomous shipping systems necessitate some form of human control and support, highlighting the importance of designing resilient HITL systems.

HITL systems prevent errors in safety-critical industries through a combination of real-time monitoring, human expertise, and the ability of rapid intervention. Two main issues with highly automated systems include lack of transparency and controllability, thus the goal of HITL systems is to address these two main concerns [151]. HITL systems can take on many forms depending on the context, and the level of autonomy of the system. In the context of ML, human expertise can be utilized in data annotation, such as for active learning, and for model validation and fine-tuning the predictions of a model [149, 152]. For example, in healthcare settings, as discussed in Section 3.4, HITL can take the form of 'clinician-in-the-loop,' in which clinicians set the initial truth labels for training samples (e.g. diagnosis labels), and iteratively provide feedback to the AI model regarding the validity or quality of the results [153]. Ultimately, the suggestions of the AI are combined with the expertise and decision-making capability of the human; it is the decision of the clinician to follow the suggestions of the model [154]. In systems like NPPs or aircrafts, HITL takes a similar form of 'operator-in-the-loop.'

HITL is often achieved through the use of effectively designed user interfaces (UIs). The integration of DM, PM, PSM, and GMs can lead to the development of intelligent (AI-based) UIs. AI-based UIs can enable effective human-machine interaction (HMI), and facilitate HITL interaction. Well-designed UIs incorporate human behavior, allowing for an efficient collaboration between humans and machines. For example, the main control room (MCR) in an NPP is a typical example of HMI in a safety-critical system. A MCR can be equipped with various AI technologies, such as computer vision, natural language processing, and RL to improve HCI [155]. Effective HMI in the MCR can reduce mental and physical pressure on personnel, the complexity of data to be monitored, and the steps required for proper operation [155]. While AI-based UIs can bring many advantages, the development of these interfaces for complex systems can be particularly challenging due to high stakes and the various factors involved in effective design. The design and implementation of AI-based UIs necessitate careful consideration of factors such as system reliability, human factors and safety. User-centered design is necessary to ensure optimal usability, trust, and effectiveness of these interfaces.

5. Challenges in using AI/ML Models for HE

Despite the benefits and the many applications of AI and ML in safety-critical industries, the use of AI, especially in safety-critical industries, is hampered by several challenges. The most common challenges to using AI to detect or mitigate HE frequently brought up in safety-critical literature include biases, erroneous data collection techniques or lack of appropriate

training data, generalizability, trustworthiness, explainability, uncertainty, and security/privacy concerns. Many of these challenges are widely noted across the studies reviewed, as well as in other reviews with similar focuses on AI/ML in safety and reliability contexts, such as [12, 156]. In this section, we explore these challenges in greater detail and discuss possible techniques for addressing these issues.

Biases: Biases in the AI model may be present throughout the model development pipeline, from the data collection process to algorithmic bias. Additionally, there is also a risk of automation bias, which is the humans' tendency to accept the output of the AI/ML model as accurate even in cases where the output may not be true. This is a particular problem for decision-support systems, where operators may rely heavily on AI-generated warnings or suggestions for error mitigation, presenting additional opportunities for HE [157]. As discussed in Section 4.2, this necessitates careful design.

Data Collection and Lack of Training Data: The trustworthiness of a model can be compromised by incorrect data collection methods, the lack adequate training data or lack of enough training data representative of the population. For example, the lack of sufficient labeled fault or accident data from real NPPs proves to be a challenge in building AI models that provide satisfactory results [158]. Collecting data is a time and cost-intensive task, which makes building extensive datasets all the more challenging, especially for rarely documented incidents/accidents or fault conditions, which may be considered anomalies. The dataset challenge is similarly encountered in the HE context, where the overall scarcity of HE data complicates the quantification of HE and HRA across many industries [159]. To address the dataset predicament inherent in AI/ML model development, it may be necessary to perform data augmentation by generating synthetic data through the use of simulators or generative models, such as GANs, as discussed in Section 2.4, using regularization techniques for unbalanced datasets, or using transfer learning to share domain knowledge [158, 160]. However, the use of simulators to generate training data may introduce its own challenges. In nuclear energy for example, generating synthetic data for probabilistic safety assessment often requires a large number of thermal-hydraulic (TH) code runs, which demands significant resources in terms of time, cost, and manpower [161]. The role of TH code is to simulate the trends of key plant parameters following an initiating event. Thus, an immense number of TH code runs is required to catalog the possible undesirable scenarios in an NPP [161]. Furthermore, the computational cost of TH code runs is exacerbated by uncertainty quantification for best estimate runs, making cataloging of all anomalous event progressions computationally infeasible [162] (curse of TH code runs). To this end, research efforts ([163, 162]) have focused on addressing the curse of TH code runs. For example, [162] has explored the use of reduced-order models (ROMs) to approximate TH code runs with less computational overhead. Additionally, existing datasets may also be messy or incomplete, particularly exacerbated by a lack of guidelines on the data collection process, necessitating data cleaning techniques to enhance the quality and usability of these datasets.

Generalizability: Generalizability is the extent to which the results of a study are applicable to a wide range of populations or situations, outside of the datasets from which the model was trained from. Generalizability can be particularly challenging for AI models. It

is difficult to build generalizable models within an industry, much less across industries. In healthcare, for example, the lack of generalizable AI models is one hindrance towards the practical implementation of AI-based clinical decision support systems, which can reduce medical errors by helping clinicians come to a correct diagnosis [164]. Generalizability can also be impacted by the lack of quality or representative training data. In the context of generative modeling for example, incomplete datasets or datasets not representative of the entire patient population can perpetuate biases and ultimately impact the fairness of models [156]. Confidentiality concerns, combined with lack of sufficient time to collect data and/or lack of funding, makes building generalizable datasets particularly challenging. Additionally, public access to many human factors and safety databases are largely limited; the integration of ML models with safety databases across industries can facilitate comparisons [12] and improve model generalizability. Generalizability is also a concern for XAI models, since explanatory needs will change among users [165]. When possible, open access and collaborative datasets can help in building more generalizable models.

Trustworthiness: The trustworthiness of AI models may be undermined by many factors, such as erroneous methods of data collection, biases and lack of fairness, ethical and privacy concerns, and lack of model interpretability [166, 167]. Another issue regarding trustworthiness is possible lack of trust in the AI's decision-making capabilities. In [168], among the top concern of interviewed pilots regarding the use of AI-driven cockpit assistant systems was the belief that the operational complexity would exceed the capabilities of the AI model.

Explainability: Trustworthiness and explainability in AI are often interconnected, in that increasing explainability in AI can also lead to increased trust [169]. In many aspects, most AI models are still 'black boxes,' which is an especially significant problem for high stakes environments like safety-critical industries. The lack of rationalization into the model algorithm and outputs may make users feel like the decisions made are arbitrary and therefore intensify trust issues, which is a problem especially for real-time decision systems. In NPPs, for example, a lack of a convincing explanation for a diagnosis made by AI, can make the operator feel accountable for any adverse effects and thus increase reluctancy to use the AI support system [112]. Thus, explainable AI, or XAI, is a rising area of interest within AI research that aims to offer insights about the model's predictions [170]. To this end, XAI frameworks such as SHapley Additive ExPlanations (SHAP) have been used by researchers to improve model explainability in safety-critical contexts [112] to give insights to the human operators on the most significant features contributing to the model prediction. For example, [171] utilize XGBoost along with SHAP models to understand the contribution of features to build a maritime accident prediction model. However, XAI is also hampered by several challenges, and does not always lead to increased trust [172]. Among the challenges of XAI include the lack of expertise of the AI model to be assessed, the dynamic and context dependent nature of decision-making, and inherent bias in developing the algorithms [172]. On the other hand, as discussed in Section 4, too much trust through increased reliance on automation and AI can also hinder human decision-making capabilities and thus increase HE.

To this end, effective AI design, as discussed in Section 4.2, plays an important role in ensuring that human expertise and AI capabilities work off each other.

Uncertainty: Given the stochastic nature of AI models and the fact that AI models only provide an approximation of relationship between model input and outcome, uncertainty in ML is inherent. Uncertainty can be defined as the likelihood that the model outcome is erroneous or outside the accepted range [173]. Uncertainties in AI models can occur as a consequence of three factors: data quality, model fit, and adherence to scope [173]. Managing uncertainty is important when actions of AI/ML systems can have a direct impact on humans and environment, as in the case of safety-critical industries. Uncertainty quantification (UQ) methods are useful in reducing the impact of uncertainties in model evaluation and can increase the reliability of results. UQ can also increase trustworthiness in the model by providing increased transparency regarding data quality or the model architecture [174].

Security: Security, especially in the context of safety-critical industries, presents a significant risk for the effective utilization of AI technologies. The integration of AI systems introduces added security vulnerabilities and risks, as all ML models are vulnerable to attacks compromising the integrity, availability, and the privacy of data and models [175]. Generative AI models and LLMs add another level of concern, given issues regarding 'hallucinations' of current models and the lack of privacy in training and input data. Erroneous information caused by model hallucinations can have direct negative implications on the end user, especially when used in the context of healthcare and medical fields, such as for medical diagnosis [176]. This drawback underscores the need for HITL systems and human experts to validate domain knowledge [118], as discussed in Section 4.2. Concerns about safeguarding data privacy also limit the practical use of publicly available LLMs in safety-critical industries, outside of academic contexts [177, 176]. Reviews of AI security risks are provided in [175, 178]. To this end, [179] propose a software architecture for deep learning with a focus on safety and security for safety-critical industries.

In addition to these overarching challenges inherent to AI, effective implementation of AI models in safety-critical systems is hindered by lack of staff competency and industry support [180]. All of these challenges ultimately culminate in the 'certification' process; certification is required for the practical deployment of ML models in safety-critical industries [181]. In addition, industries may experience knowledge gaps between how to operate an AI model and this too can amplify hesitations regarding trust and explainability. Addressing these requires partnerships between all stakeholders to ensure proper training and communication regarding AI/ML practices in industry. Especially, as noted in [181], strengthening collaboration between academia and industry partners can allow for more relevant datasets to be utilized and the development of ML models that better match the task and reliability requirements and constraints imposed in safety-critical industries.

6. Discussion

In this section, we provide a discussion of key insights drawn from the literature reviewed and outline possible future research directions. Based on the studies reviewed in this paper, we can broadly categorize literature into two categories: those that directly consider the

impact of HE and aim to categorize or mitigate it and those that aim to benefit human operators with their tasks, which, in turn, help to reduce HE. The former includes studies that aim to quantify or categorize HE, with the goal of understanding how HE occurs in the respective industry. For the latter, we can consider studies that focus on improving tasks that human operators do, with the goal of enhancing the reliability and efficiency of tasks, and ultimately, creating an environment that is less prone to HE.

Our review reveals that the majority fall under the umbrella of PM, with fewer concentrating on DM, followed by PSM and GM. As discussed in Section 2, it is often the case that many studies do not fit neatly into a single modeling type; they often incorporate multiple dimensions, such as building a PM from DM or integrating a PM with PSM. In DM, studies are generally similar in purpose, with the goal of identifying causal factors, categorizing HE or drawing human reliability insights, from historical incident datasets. Commonly employed techniques include classification models, such as decision trees, RFs, and SVMs. In the PM category, generally, studies often concern HE or accident prediction, accident outcome, or accident severity prediction. We find that in our reviewed studies, HE risk analysis remains heavily considered, suggesting the ongoing importance of understanding HE in safety-critical domains. Many PMs are based on NNs or SVMs, rather than regression analysis models, which is consistent with [93]. Sometimes the quantity and complexity of relationship between human factors or data features can increase the risk of overfit for other 'simpler' models, such as DTs. PSM studies often include some DM or historical data analysis and may build on PM with accident/HE prediction. Operator support systems are common forms of PSMs. We find that in almost every study reviewed, more than one ML method is utilized, often combining algorithms or using multiple models are implemented to find the best performing model.

Another key insight relates to the distribution of modeling types in the literature. Although DMs are popular in the grand scheme of HE literature, we observe that more DM studies involved a combination of DM and PM or PSM, rather than purely DM using AI/ML. Statistical models or the use of HRA/classification schemes, as discussed in Section 1.4, remain popular methods for descriptive analytics in safety-critical industries, likely due to being well-established and accepted methods in practice. This trend toward other modeling methods can be driven by the need of safety-critical industries to adapt to new technologies and provide operators with additional methods of support. While DM still remains invaluable, these complex industries often require more advanced models that go beyond describing past events but anticipate future events or provide recommended actions. This observation is also consistent with a review by [10], and [182], in which the authors state that some industries, such as healthcare, may value prediction more than explanation, due to the potential severity of consequences resulting from delayed or inappropriate actions in these fields. Consequently, PM or PSMs can be more valuable in providing timely insights that enhance decision-making capabilities. Interestingly however, we also find that although PSMs have gained added interest in recent years, fewer studies focus specifically on PSMs compared to DM and PM. Many PM studies, for example, express the potential for future research to develop PSMs. The current disparity in literature between DM and PMs and PSMs

suggests an opportunity to transition more towards developing actionable models that provide specific recommendations and guidance. This shift from describing and predicting to *prescribing* can provide a more holistic view to HE as it relates to safety-critical industries. Similarly, we find that despite the rising popularity of GMs across a wide range of applications, GMs remain largely underrepresented in safety-critical contexts. However, we anticipate the number of GM studies to increase, particularly in light of increased prevalence and application range of GMs, such as LLMs.

Although we find that there are plenty of research opportunities in this field, especially with regards to PSM and GMs, certain limitations need to be acknowledged. Namely, as also discussed in Section 5, challenges with data quality and/or availability are widely noted in the studies reviewed. Additionally, the general lack of regulation or standards on building industry-acceptable HE AI/ML models can make model comparisons and robustness testing more challenging.

7. Conclusion

This study focuses on literature contributions related to AI and ML in addressing HE in safety-critical industries, with a focus on studies published between 2013 and 2024, to date. Our objective in this review is to offer a comprehensive and multi-industry overview of literature and to provide a holistic understanding of how AI/ML models are being applied to detect and mitigate HE in safety-critical industries. Specifically, we explore (1) how different modeling types (DM, PM, PSM, and GM) are utilized to detect and mitigate HE in safety-critical industries, (2) the differences in applications among the modeling types, (3) the limitations and challenges of AI/ML, and (4) key insights from the literature. As explored in Section 3, AI/ML models are applied by researchers in various ways across a wide range of applications and industries to identify patterns and relationships, predict the risk and severity of HE, and assist human operators in their tasks to reduce the likelihood of HE, suggesting the potential of these technologies in enhancing safety and reducing opportunities for HE. We also identify several key challenges in Section 5, especially with regards to data quality and availability and safety-critical specific challenges that hinder practical deployment of ML models. Lastly, we explore key insights identified in Section 6. Our review highlights under-represented modeling types that hold significant opportunities for future research to enhance practical applications in safety-critical contexts.

This review has certain limitations. Namely, as mentioned in Section 3, the search is not exhaustive, as we are limited by the specific keywords utilized and space limitations. As is consistent with findings from other reviews on safety-critical industries [11, 183], studies in this area are fragmented, from a variety of outlets, making an exhaustive review especially challenging. Notably, the chosen keywords may not be included in every relevant work, further complicating the search process. Additionally, 'safety-critical' systems covers a wide range of industries and applications, and focus for HE may be different across industries. A future paper can consider additional studies and can expand to include manufacturing. Manufacturing, although not considered safety-critical by our definition, can also provide additional insights for safety-critical industries. Similar to safety-critical industries,

manufacturing is held to high reliability and safety standards and HE is a significant concern that can affect the efficiency of the system. Thus, there is value in exploring manufacturing studies. Additionally, given the increasing popularity of autonomous driving systems, this topic could warrant a separate, focused review.

CReDiT

Ezgi Gursel: Investigation, Formal Analysis, Visualization, Writing - Original Draft, Writing - Review & Editing. **Mahboubeh Madadi:** Writing - Review & Editing, Funding Acquisition. **Jamie Baalis Coble:** Writing - Review & Editing, Funding Acquisition. **Vivek Agarwal:** Writing - Review & Editing, Funding Acquisition. **Vaibhav Yadav:** Writing - Review & Editing, Funding Acquisition. **Ronald L. Boring:** Writing - Review & Editing, Funding Acquisition. **Anahita Khojandi:** Writing - Review & Editing, Supervision, Conceptualization, Methodology, Funding Acquisition.

Acknowledgments

This work of authorship was prepared as an account of work sponsored by the U.S. Department of Energy, an agency of the U.S. Government, under Grant No. DE-NE0008978 to University of Tennessee-Knoxville. The authors would like to thank Nesar Titu for his valuable contributions to the literature review. Neither the U.S. Government, nor any agency thereof, nor any of their employees makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.

References

[1] John C Knight. Safety critical systems: challenges and directions. In *Proceedings of the 24th international conference on software engineering*, pages 547–550, 2002.

[2] Jeffrey Pfeffer. Building sustainable organizations: The human factor. *Academy of Management Perspectives*, 24(1):34–45, 2010.

[3] David Woods, Leila J Johannesen, Richard I Cook, and Nadine B Sarter. Behind human error: Cognitive systems, computers, and hindsight. 1994.

[4] Harmen Kragt. Enhancing industrial performance: Experiences of integrating the human factor. *Ergonomics*, 38(8):1674–1685, 1995.

[5] Richard E Iliffe, Paul WH Chung, Trevor A Kletz, and Malcolm Preston. The application of active databases to the problems of human error in industry. *Journal of Loss Prevention in the Process Industries*, 13(1):19–26, 2000.

[6] Carine Dominguez-Péry, Lakshmi Narasimha Raju Vuddaraju, Isabelle CorbettEtchevers, and Rana Tassabehji. Reducing maritime accidents in ships by

tackling human error: A bibliometric review and research agenda. *Journal of Shipping and Trade*, 6:1–32, 2021.

[7] Velibor Božić. Application of artificial intelligence in reducing risks caused by the human factor.

[8] BS Dhillon and Y Liu. Human error in maintenance: A review. *Journal of quality in maintenance engineering*, 2006.

[9] Faeze Ghofrani, Qing He, Rob MP Goverde, and Xiang Liu. Recent applications of big data analytics in railway transportation systems: A survey. *Transportation Research Part C: Emerging Technologies*, 90:226–246, 2018.

[10] Panagiota Galetsi and Korina Katsaliaki. A review of the literature on big data analytics in healthcare. *Journal of the Operational Research Society*, 71(10):1511–1529, 2020.

[11] Yue Wang and Sai Ho Chung. Artificial intelligence in safety-critical systems: A systematic review. *Industrial Management & Data Systems*, 122(2):442–470, 2022.

[12] Zhaoyi Xu and Joseph Homer Saleh. Machine learning for reliability engineering and safety applications: Review of current status and future opportunities. *Reliability Engineering & System Safety*, 211:107530, 2021.

[13] Sobhan Sarkar and J Maiti. Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis. *Safety Science*, 131:104900, 2020.

[14] Erik Hollnagel. The phenotype of erroneous actions. *International Journal of Man-Machine Studies*, 39(1):1–32, 1993.

[15] ISO 14224:2016. Petroleum, petrochemical and natural gas industries: Collection and exchange of reliability and maintenance data for equipment. Standard, 2016.

[16] Steve Mason. Improving maintenance-reducing human error. In *Railway Safety Papers from the Railway Technology Conference held at Railtex 2000*, Birmingham, UK, November 2000.

[17] Frederick D Hansen. Human error: A concept analysis. *Journal of Air Transportation*, 11(3), 2007.

[18] James Reason. *Human Error*. Cambridge University Press, 1990.

[19] Linda T Kohn, Janet M Corrigan, Molla S Donaldson, et al. Why do errors happen? In *To Err is Human: Building a Safer Health System*. National Academies Press (US), 2000.

[20] Pascale Carayon and Kenneth E Wood. Patient safety. *Information Knowledge Systems Management*, 8(1-4):23–46, 2009.

[21] Mfundo Nkosi, Kapil Gupta, and Madindwa Mashinini. Causes and impact of human error in maintenance of mechanical systems. In *MATEC Web of Conferences*, volume 312, page 05001. EDP Sciences, 2020.

[22] Esmaeil Zarei, Faisal Khan, and Rouzbeh Abbassi. Importance of human reliability in process operation: A critical analysis. *Reliability Engineering & System Safety*, 211:107607, 2021.

[23] David Meister. *The History of Human Factors and Ergonomics*. CRC Press, 2018.

[24] William B Rouse and Sandra H Rouse. Analysis and classification of human error. *IEEE Transactions on Systems, Man, and Cybernetics*, (4):539–549, 1983.

[25] Marilyn Sue Bogner. *Human Error in Medicine*. CRC Press, 2018.

[26] Bing Wu, Tsz Leung Yip, Xinping Yan, and C Guedes Soares. Review of techniques and challenges of human and organizational factors analysis in maritime transportation. *Reliability Engineering & System Safety*, 219:108249, 2022.

[27] Punitkumar Bhavsar, Babji Srinivasan, and Rajagopalan Srinivasan. Pupillometry based real-time monitoring of operator's cognitive workload to prevent human error during abnormal situations. *Industrial & Engineering Chemistry Research*, 55(12):3372–3382, 2016.

[28] Irina-Maria Dragan and Alexandru Isaic-Maniu. The reliability of the human factor. *Procedia Economics and Finance*, 15:1486–1494, 2014.

[29] Donald A Norman. Categorization of action slips. *Psychological Review*, 88(1):1, 1981.

[30] Jens Rasmussen. Human errors. a taxonomy for describing human malfunction in industrial installations. *Journal of Occupational Accidents*, 4(2-4):311–333, 1982.

[31] Neville A Stanton and Paul M Salmon. Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems. *Safety Science*, 47(2):227–237, 2009.

[32] Charles E Billings and William D Reynard. Dimensions of the information transfer problem. *Information transfer problems in the aviation system*, pages 9–14, 1981.

[33] William P Monan. Distraction-a human factor in air carrier hazard events. *NASA Technical Memorandum*, 78608:2–23, 1979.

[34] Leon H Nawrocki, Michael H Strub, and Ross M Cecil. Error categorization and analysis in man-computer communication systems. *IEEE Transactions on Reliability*, 22(3):135–140, 1973.

[35] HP Ruffell Smith. A simulator study of the interaction of pilot workload with errors, vigilance, and decisions. Technical Report NASA-TM-78482, 1979.

[36] James Reason. Human error: models and management. *BMJ*, 320(7237):768–770, 2000.

[37] Scott A Shappell and Douglas A Wiegmann. The human factors analysis and classification system–HFACS. 2000.

[38] Mehmet Kaptan, Songu¨l Sarıaliog˘lu, Ozkan Ug˘urlu, and Jin Wang. The evolution¨ of the HFACS method used in analysis of marine accidents: A review. *International Journal of Industrial Ergonomics*, 86:103225, 2021.

[39] Thomas Diller, George Helmrich, Sharon Dunning, Stephanie Cox, April Buchanan, and Scott Shappell. The human factors analysis classification system (HFACS) applied to health care. *American Journal of Medical Quality*, 29(3):181–190, 2014.

[40] Sa Kil Kim, Yong Hee Lee, Tong Il Jang, Yeon Ju Oh, and Kwang Hyeon Shin. An investigation on unintended reactor trip events in terms of human error hazards of korean nuclear power plants. *Annals of Nuclear Energy*, 65:223–231, 2014.

[41] Stephen C Theophilus, Victor N Esenowo, Andrew O Arewa, Augustine O Ifelebuegu, Ernest O Nnadi, and Fredrick U Mbanaso. Human factors analysis and classification system for the oil and gas industry (HFACS-OGI). *Reliability Engineering & System Safety*, 167:168–176, 2017.

[42] JW Garrett and Jochen Teizer. Human factors analysis classification system relating to human error awareness taxonomy in construction safety. *Journal of Construction Engineering and Management*, 135(8):754–763, 2009.

[43] Jae W Kim, Wondea Jung, and Jaejoo Ha. AGAPE-ET: A methodology for human error analysis of emergency tasks. *Risk Analysis: An International Journal*, 24(5):1261–1277, 2004.

[44] Jiadong Gong. Harnessing the power of ai in materials digital transformation: a synergistic hybrid approach. *The Bridge*, pages 30–37, 2024.

[45] Aurelien Teguede Keleko, Bernard Kamsu-Foguem, Raymond Houe Ngouna, and Amèvi Tongne. Artificial intelligence and real-time predictive maintenance in industry 4.0: a bibliometric analysis. *AI and Ethics*, 2(4):553–577, 2022.

[46] David Bendig and Antonio Br¨aunche. The role of artificial intelligence algorithms in information systems research: a conceptual overview and avenues for research. *Management Review Quarterly*, pages 1–46, 2024.

[47] Mehmed Kantardzic. *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons, 2011.

[48] Katerina Lepenioti, Alexandros Bousdekis, Dimitris Apostolou, and Gregoris Mentzas. Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, 50:57–70, 2020.

[49] Lian Duan and Ye Xiong. Big data analytics and business analytics. *Journal of Management Analytics*, 2(1):1–21, 2015.

[50] Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

[51] Jinkyun Park, Hee Eun Kim, and Inseok Jang. Empirical estimation of human error probabilities based on the complexity of proceduralized tasks in an analog environment. *Nuclear Engineering and Technology*, 54(6):2037–2047, 2022.

[52] Bart Roets and Johan Christiaens. Shift work, fatigue, and human error: An empirical analysis of railway traffic control. *Journal of Transportation Safety & Security*, 11(2):207–224, 2019.

[53] Xing Pan, Ye Lin, and Congjiao He. A review of cognitive models in human reliability analysis. *Quality and Reliability Engineering International*, 33(7):1299–1316, 2017.

[54] Yung Hsien James Chang, Yochan Kim, Jinkyun Park, and Lawrence Criscione. SACADA and HuREX: Part 1. the use of SACADA and HuREX systems to collect human reliability data. *Nuclear Engineering and Technology*, 54(5):1686–1697, 2022.

[55] Reza Soltanpoor and Timos Sellis. Prescriptive analytics for big data. In *Australasian Database Conference*, pages 245–256. Springer, 2016.

[56] Casey C Bennett and Kris Hauser. Artificial intelligence framework for simulating clinical decision-making: A markov decision process approach. *Artificial Intelligence in Medicine*, 57(1):9–19, 2013.

[57] Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 4(1):1–103, 2010.

[58] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, 2018.

[59] Ceren Güzel Turhan and Hasan Sakir Bilge. Recent trends in deep generative models: a review. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, pages 574–579. IEEE, 2018.

[60] Prashnna Ghimire, Kyungki Kim, and Manoj Acharya. Generative ai in the construction industry: Opportunities & challenges. *arXiv preprint arXiv:2310.04427*, 2023.

[61] GM Harshvardhan, Mahendra Kumar Gourisaria, Manjusha Pandey, and Siddharth Swarup Rautaray. A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38:100285, 2020.

[62] Fahad Umer and Niha Adnan. Generative artificial intelligence: synthetic datasets in dentistry. *BDJ open*, 10(1):13, 2024.

[63] Ji Hun Park, Hye Seon Jo, and Man Gyun Na. System and component anomaly detection using lstm-vae. In *2021 5th International Conference on System Reliability and Safety (ICSRS)*, pages 131–137. IEEE, 2021.

[64] Li Fan, Lee Ching-Hung, Han Su, Feng Shanshan, Jiang Zhuoxuan, and Sun Zhu. A new era in human factors engineering: A survey of the applications and prospects of large multimodal models. *arXiv preprint arXiv:2405.13426*, 2024.

[65] Nees Jan Van Eck and Ludo Waltman. VOSviewer manual. *Leiden: Univeristeit Leiden*, 1(1):1–53, 2013.

[66] Lingling Hua, Wei Zheng, and Shigen Gao. Extraction and analysis of risk factors from chinese railway accident reports. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 869–874. IEEE, 2019.

[67] Hamad Alawad, Sakdirat Kaewunruen, and Min An. Learning from accidents: Machine learning for safety at railway stations. *IEEE Access*, 8:633–648, 2019.

[68] Ozgur Ugur, Ali Atilla Arisoy, Murat Can Ganiz, and Berkay Bolac. Descriptive and prescriptive analysis of construction site incidents using decision tree classification and association rule mining. In *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6. IEEE, 2021.

[69] Fereshteh Sattari, Renato Macciotta, Daniel Kurian, and Lianne Lefsrud. Application of bayesian network and artificial intelligence to reduce accident/incident rates in oil & gas companies. *Safety Science*, 133:104981, 2021.

[70] Caroline Morais, Ka Lai Yung, Karl Johnson, Raphael Moura, Michael Beer, and Edoardo Patelli. Identification of human errors and influencing factors: A machine learning approach. *Safety Science*, 146:105528, 2022.

[71] R Ouache, E Bakhtavar, G Hu, K Hewage, and R Sadiq. Evidential reasoning and machine learning-based framework for assessment and prediction of human error factorsinduced fire incidents. *Journal of Building Engineering*, 49:104000, 2022.

[72] Dana Arad, Ariel Rosenfeld, and Racheli Magnezi. Factors contributing to preventing operating room “never events”: A machine learning analysis. *Patient Safety in Surgery*, 17(1):1–9, 2023.

[73] Indumathi Nallathambi, Padmaja Savaram, Sudhakar Sengan, Meshal Alharbi, Samah Alshathri, Mohit Bajaj, Moustafa H Aly, and Walid El-Shafai. Impact of fireworks industry safety measures and prevention management system on human error mitigation using a machine learning approach. *Sensors*, 23(9):4365, 2023.

[74] Shiqi Fan, Eduardo Blanco-Davis, Zaili Yang, Jinfen Zhang, and Xinping Yan. Incorporation of human factors into maritime accident analysis using a data-driven bayesian network. *Reliability Engineering & System Safety*, 203:107070, 2020.

[75] Dejun Chen, Yilou Pei, and Qian Xia. Research on human factors cause chain of ship accidents based on multidimensional association rules. *Ocean Engineering*, 218:107717, 2020.

[76] Fadda Paolo, Fancello Gianfranco, Frigau Luca, Mandas Marco, Medda Andrea, Mola Francesco, Pelligra Vittorio, Porta Mattia, and Serra Patrizia. Investigating the role of the human element in maritime accidents using semi-supervised hierarchical methods. *Transportation Research Procedia*, 52:252–259, 2021.

[77] Ching-Wu Cheng, Hong-Qing Yao, and Tsung-Chih Wu. Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. *Journal of Loss Prevention in the Process Industries*, 26(6):1269–1278, 2013.

[78] Ik-Hyun Youn, Deuk-Jin Park, and Jeong-Bin Yim. Analysis of lookout activity in a simulated environment to investigate maritime accidents caused by human error. *Applied Sciences*, 9(1):4, 2018.

[79] Jeng-Chung Chen and F Yu Vincent. Relationship between human error intervention strategies and unsafe acts: The role of strategy implementability. *Journal of Air Transport Management*, 69:112–122, 2018.

[80] Baoping Cai, Yonghong Liu, Yunwei Zhang, Qian Fan, Zengkai Liu, and Xiaojie Tian. A dynamic bayesian networks modeling of human factors on offshore blowouts. *Journal of Loss Prevention in the Process Industries*, 26(4):639–649, 2013.

[81] C Morais, R Moura, M Beer, and E Patelli. Attempt to predict human error probability in different industry sectors using data from major accidents and bayesian networks. *14th Probabilistic Safety Assessment and Management, PSAM 2018*, 2018.

[82] Pin-Chao Liao, Mei Liu, Yu-Sung Su, Hui Shi, and Xintong Luo. Estimating the influence of improper workplace environment on human error: Posterior predictive analysis. *Advances in Civil Engineering*, 2018, 2018.

[83] Shengyuan Yan, Kai Yao, Fengjiao Li, Yingying Wei, and Cong Chi Tran. Constructing neural network model to evaluate and predict human error probability in nuclear power plants based on eye response, workload rating, and situation awareness. *Nuclear Technology*, 208(10):1540–1552, 2022.

[84] Young A Suh and Man-Sung Yim. A worker's fitness-for-duty status identification based on biosignals to reduce human error in nuclear power plants. *Nuclear Technology*, 206(12):1840–1860, 2020.

[85] Riccardo Patriarca, Marilia Ramos, Nicola Paltrinieri, Salvatore Massaiu, Francesco Costantino, Giulio Di Gravio, and Ronald Laurids Boring. Human reliability analysis: Exploring the intellectual structure of a research field. *Reliability Engineering & System Safety*, 203:107102, 2020.

[86] Tomás Madeira, Rui Melício, Duarte Valério, and Luis Santos. Machine learning and natural language processing for prediction of human factors in aviation incident reports. *Aerospace*, 8(2):47, 2021.

[87] Fan Li, Chun-Hsien Chen, Ching-Hung Lee, and Shanshan Feng. Artificial intelligenceenabled non-intrusive vigilance assessment approach to reducing traffic controller's human errors. *Knowledge-Based Systems*, 239:108047, 2022.

[88] R Alan Burnett and Dong Si. Prediction of injuries and fatalities in aviation accidents through machine learning. In *Proceedings of the International Conference on Compute and Data Analysis*, pages 60–68, 2017.

[89] Liu Shao-Yu, Tang Hui-Chin, and Yu-Cheng Wang. The study on the prediction models of human factor flight accidents by combining fuzzy clustering methods and neural networks. *Journal of Aeronautics, Astronautics and Aviation*, 50(2):175–185, 2018.

[90] Chen Wang, Lin Liu, Chengcheng Xu, and Weitao Lv. Predicting future driving risk of crash-involved drivers based on a systematic machine learning framework. *International journal of environmental research and public health*, 16(3):334, 2019.

[91] Andrea Coraddu, Luca Oneto, Beatriz Navas de Maya, and Rafet Kurt. Determining the most influential human factors in maritime accidents: A data-driven approach. *Ocean Engineering*, 211:107588, 2020.

[92] Abbas Zaranezhad, Hasan Asilian Mahabadi, and Mohammad Reza Dehghani. Development of prediction models for repair and maintenance-related accidents at oil refineries using artificial neural network, fuzzy system, genetic algorithm, and ant colony optimization algorithm. *Process Safety and Environmental Protection*, 131:331–348, 2019.

[93] Sobhan Sarkar, Sammangi Vinay, Rahul Raj, Jhareswar Maiti, and Pabitra Mitra. Application of optimized machine learning techniques for prediction of occupational accidents. *Computers & Operations Research*, 106:210–224, 2019.

[94] Antoine J-P Tixier, Matthew R Hallowell, Balaji Rajagopalan, and Dean Bowman. Application of machine learning to construction injury prediction. *Automation in Construction*, 69:102–114, 2016.

[95] Nicola Tamascelli, Riccardo Solini, Nicola Paltrinieri, and Valerio Cozzani. Learning from major accidents: A machine learning approach. *Computers & Chemical Engineering*, 162:107786, 2022.

[96] Rongchen Zhu, Xiaofeng Hu, Jiaqi Hou, and Xin Li. Application of machine learning techniques for predicting the consequences of construction accidents in china. *Process Safety and Environmental Protection*, 145:293–302, 2021.

[97] Rui PR Nogueira, Rui Melicio, Duarte Val´erio, and Lu´ıs FFM Santos. Learning methods and predictive modeling to identify failure by human factors in the aviation industry. *Applied Sciences*, 13(6):4069, 2023.

[98] He Lan, Xiaoxue Ma, Weiliang Qiao, and Wanyi Deng. Determining the critical risk factors for predicting the severity of ship collision accidents using a data-driven approach. *Reliability Engineering & System Safety*, 230:108934, 2023.

[99] Pierre Pauline R Abesamis, Remedios de Dios Bulos, and Michelle Ching. Improving aviation incidents using association rule mining algorithm and time series analysis. In *IOP Conference Series: Materials Science and Engineering*, volume 946, page 012005. IOP Publishing, 2020.

[100] Junyong Bae, Geunhee Kim, and Seung Jun Lee. Real-time prediction of nuclear power plant parameter trends following operator actions. *Expert Systems with Applications*, 186:115848, 2021.

[101] Xiaoge Zhang, Sankaran Mahadevan, Nathan Lau, and Matthew B Weinger. Multisource information fusion to assess control room operator performance. *Reliability Engineering & System Safety*, 194:106287, 2020.

[102] Indrila Ganguly, Graham Buhrman, Ed Kline, Seong K Mun, and Srijan Sengupta. Automated error labeling in radiation oncology via statistical natural language processing. *Diagnostics*, 13(7):1215, 2023.

[103] Shiqi Fan and Zaili Yang. Towards objective human performance measurement for maritime safety: A new psychophysiological data-driven machine learning method. *Reliability Engineering & System Safety*, 233:109103, 2023.

[104] Shiqi Fan and Zaili Yang. Accident data-driven human fatigue analysis in maritime transport using machine learning. *Reliability Engineering & System Safety*, 241:109675, 2024.

[105] William Hoyos, Jose Aguilar, Mayra Raciny, and Mauricio Toro. Case studies of clinical decision-making through prescriptive models based on machine learning. *Computer Methods and Programs in Biomedicine*, 242:107829, 2023.

[106] Qiong Hu, Miao Cai, Nasrin Mohabbati-Kalejahi, Amir Mehdizadeh, Mohammad Ali Alamdar Yazdi, Alexander Vinel, Steven E Rigdon, Karen C Davis, and Fadel M Megahed. A review of data analytic applications in road traffic safety. Part 2: Prescriptive modeling. *Sensors*, 20(4):1096, 2020.

[107] Dongyang Yan, Keping Li, Qiaozhen Zhu, and Yanyan Liu. A railway accident prevention method based on reinforcement learning–active preventive strategy by multimodal data. *Reliability Engineering & System Safety*, 234:109136, 2023.

[108] JaeKwan Park, TaekKyu Kim, SeungHwan Seong, and SeoRyong Koo. Control automation in the heat-up mode of a nuclear power plant using reinforcement learning. *Progress in Nuclear Energy*, 145:104107, 2022.

[109] Jeeyea Ahn and Seung Jun Lee. Deep learning-based procedure compliance check system for nuclear power plant emergency operation. *Nuclear Engineering and Design*, 370:110868, 2020.

[110] Jeeyea Ahn, Junyong Bae, and Seung Jun Lee. A human error detection system in nuclear power plant operations. *Nuclear Science and Engineering*, 2021.

[111] Jeeyea Ahn, Junyong Bae, Byung Joo Min, and Seung Jun Lee. Operation validation system to prevent human errors in nuclear power plants. *Nuclear Engineering and Design*, 397:111949, 2022.

[112] Ji Hun Park, Hye Seon Jo, Sang Hyun Lee, Sang Won Oh, and Man Gyun Na. A reliable intelligent diagnostic assistant for nuclear power plants using explainable artificial intelligence of GRU-AE, LightGBM and SHAP. *Nuclear Engineering and Technology*, 54(4):1271–1287, 2022.

[113] Emilie MD Jean-Baptiste, Pia Rotshtein, and Martin Russell. Pomdp based action planning and human error detection. In *Artificial Intelligence Applications and Innovations: 11th IFIP WG 12.5 International Conference, AIAI 2015*, volume 11, pages 250–265, Bayonne, France, September 2015. Springer.

[114] Liam A Kruse, Justin M Bradley, and Marilyn Wolf. A control authority switching system for avoiding multicopter loss of control using a markov decision process. In *AIAA Scitech 2019 Forum*, page 1688, 2019.

[115] Krishnamurthy V Vemuru, Steven D Harbour, and Jeffery D Clark. Reinforcement learning in aviation, either unmanned or manned, with an injection of AI. In *20th International Symposium on Aviation Psychology*, page 492, 2019.

[116] Archana Tikayat Ray, Anirudh Prabhakara Bhat, Ryan T White, Van Minh Nguyen, Olivia J Pinon Fischer, and Dimitri N Mavris. Examining the potential of generative language models for aviation safety analysis: Case study and insights using the aviation safety reporting system (asrs). *Aerospace*, 10(9):770, 2023.

[117] Yukinori Harada, Tomoharu Suzuki, Taku Harada, Tetsu Sakamoto, Kosuke Ishizuka, Taiju Miyagami, Ren Kawamura, Kotaro Kunitomo, Hiroyuki Nagano, Taro Shimizu, et al. Performance evaluation of chatgpt in detecting diagnostic errors and their contributing factors: an analysis of 545 case reports of diagnostic errors. *BMJ Open Quality*, 13(2):e002654, 2024.

[118] Declan Saunders, Kyle Hu, and Wen-Chin Li. The process of training chatgpt using hfacs to analyse aviation accident reports. 2024.

[119] Shuwen Zheng, Kai Pan, Jie Liu, and Yunxia Chen. Empirical study on fine-tuning pre-trained large language models for fault diagnosis of complex systems. *Reliability Engineering & System Safety*, page 110382, 2024.

[120] Ezgi Gursel, Bhavya Reddy, Anahita Khojandi, Mahboubeh Madadi, Jamie Baalis Coble, Vivek Agarwal, Vaibhav Yadav, and Ronald L Boring. Using artificial intelligence to detect human errors in nuclear power plants: A case in operation and maintenance. *Nuclear Engineering and Technology*, 55(2):603–622, 2023.

[121] Prashant C Mural, Rathna GN, and Virat Bhola. Autoencoder-based pilot error quantification model for aviation safety. In *AIAA SCITECH 2024 Forum*, page 2584, 2024.

[122] JaeKwan Park, TaekKyu Kim, and SeungHwan Seong. Providing support to operators for monitoring safety functions using reinforcement learning. *Progress in Nuclear Energy*, 118:103123, 2020.

[123] Xuan Liu, Huixing Meng, Xu An, and Jinduo Xing. Integration of functional resonance analysis method and reinforcement learning for updating and optimizing emergency procedures in variable environments. *Reliability Engineering & System Safety*, 241:109655, 2024.

[124] Romina Eramo, Francis Bordeleau, Benoit Combemale, Mark van Den Brand, Manuel Wimmer, and Andreas Wortmann. Conceptualizing digital twins. *IEEE Software*, 2021.

[125] Claudio Mandolla, Antonio Messeni Petruzzelli, Gianluca Percoco, and Andrea Urbinati. Building a digital twin for additive manufacturing through the exploitation of blockchain: A case analysis of the aircraft industry. *Computers in Industry*, 109:134–152, 2019.

[126] Cho Yin Yiu, Kam KH Ng, Ching-Hung Lee, Chun Ting Chow, Tsz Ching Chan, Kwok Chun Li, and Ka Yeung Wong. A digital twin-based platform towards intelligent automation with virtual counterparts of flight and air traffic control operations. *Applied Sciences*, 11(22):10923, 2021.

[127] Dong Zhong, Zhelei Xia, Yian Zhu, and Junhua Duan. Overview of predictive maintenance based on digital twin technology. *Heliyon*, 2023.

[128] Y Hirotsu, K Suzuki, M Kojima, and K Takano. Multivariate analysis of human error incidents occurring at nuclear power plants: Several occurrence patterns of observed human errors. *Cognition, Technology & Work*, 3(2):82–91, 2001.

[129] Richard Williams, John Ahmet Erkoyuncu, Tariq Masood, et al. Augmented reality assisted calibration of digital twins of mobile robots. *IFAC-PapersOnLine*, 53(3):203–208, 2020.

[130] Ján Vachálek, Lukás Bartalský, Oliver Rovný, Dana Šišmišová, Martin Morháč, andˇ Milan Lokšík. The digital twin of an industrial production line within the industry 4.0 concept. In *2017 21st International Conference on Process Control (PC)*, pages 258–262. IEEE, 2017.

[131] Diego M Botín-Sanabria, Adriana-Simona Mihaita, Rodrigo E Peimbert-García, Mauricio A Ramírez-Moreno, Ricardo A Ramírez-Mendoza, and Jorge de J Lozoya-Santos. Digital twin technology challenges and applications: A comprehensive review. *Remote Sensing*, 14(6):1335, 2022.

[132] Michael D Muhlheim, Pradeep Ramuhalli, Alex Huning, Askin Guler Yigitoglu, Richard Thomas Wood, and Abhinav Saxena. Status report on regulatory criteria applicable to the use of digital twins. 2022.

[133] Angira Sharma, Edward Kosasih, Jie Zhang, Alexandra Brintrup, and Anisoara Calinescu. Digital twins: State of the art theory and practice, challenges, and open research questions. *Journal of Industrial Information Integration*, 30:100383, 2022.

[134] Harleen Kaur Sandhu, Saran Srikanth Bodda, and Abhinav Gupta. A future with machine learning: review of condition assessment of structures and mechanical systems in nuclear facilities. *Energies*, 16(6):2628, 2023.

[135] Yujia Lin, Liming Chen, Aftab Ali, Christopher Nugent, Cleland Ian, Rongyang Li, Dazhi Gao, Hang Wang, Yajie Wang, and Huansheng Ning. Human digital twin: A survey. *arXiv preprint arXiv:2212.05937*, 2022.

[136] Michael E Miller and Emily Spatz. A unified view of a human digital twin. *Human Intelligent Systems Integration*, pages 1–11, 2022.

[137] Hannah Lonsdale, Geoffrey M Gray, Luis M Ahumada, Hannah M Yates, Anna Varughese, and Mohamed A Rehman. The perioperative human digital twin. *Anesthesia & Analgesia*, 134(4):885–892, 2022.

[138] Ronald Boring, Thomas Ulrich, Jooyoung Park, Yunyeong Heo, and Jeeyea Ahn. The HUNTER dynamic human reliability analysis tool: Overview of the enhanced framework for modeling human digital twins. In *Proceedings of the Probabilistic Safety Assessment and Management (PSAM 16) Conference*, 2022.

[139] Bharatwaajan Balaji, Mohammed Aatif Shahab, Babji Srinivasan, and Rajagopalan Srinivasan. ACT-R based human digital twin to enhance operators' performance in process industries. *Frontiers in Human Neuroscience*, 17:18, 2023.

[140] Jonas Gouraud, Arnaud Delorme, and Bruno Berberian. Autopilot, mind wandering, and the out of the loop performance problem. *Frontiers in Neuroscience*, 11:541, 2017.

[141] David B Kaber, Norbert Stoll, Kerstin Thurow, Rebecca S Green, Sang-Hwan Kim, and Prithima Mosaly. Human–automation interaction strategies and models for life science applications. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 19(6):601–621, 2009.

[142] Susanne Niehaus, Matthias Hartwig, Patricia H Rosen, and Sascha Wischniewski. An occupational safety and health perspective on human in control and AI. *Frontiers in Artificial Intelligence*, 5:868382, 2022.

[143] Mica R Endsley and Esin O Kiris. The out-of-the-loop performance problem and level of control in automation. *Human Factors*, 37(2):381–394, 1995.

[144] Shahabuddin Muhammad. Modeling operator performance in human-in-the-loop autonomous systems. *IEEE Access*, 9:102715–102731, 2021.

[145] Lisanne Bainbridge. Ironies of automation. In *Analysis, Design and Evaluation of Man–machine Systems*, pages 129–135. Elsevier, 1983.

[146] Sambit Ghosh and B Wayne Bequette. Process systems engineering and the humanin-the-loop: The smart control room. *Industrial & Engineering Chemistry Research*, 59(6):2422–2429, 2019.

[147] Ronald L Boring, Torrey J Mortenson, Thomas A Ulrich, and Roger Lew. Humans with/as big data in nuclear energy. *Human Factors in Energy: Oil, Gas, Nuclear and Electric Power*, 54:56, 2022.

[148] Xinyuan Zhang, Shiqi Wang, Jie Liu, and Cui Tao. Towards improving diagnosis of skin diseases by combining deep neural network and human knowledge. *BMC Medical Informatics and Decision Making*, 18(2):69–76, 2018.

[149] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381, 2022.

[150] J Hans van den Broek, JR Jaco Griffioen, and M Monique van der Drift. Meaningful human control in autonomous shipping: an overview. In *IOP Conference Series: Materials Science and Engineering*, volume 929, page 012008. IOP Publishing, 2020.

[151] Enes Yigitbas, Kadiray Karakaya, Ivan Jovanovikj, and Gregor Engels. Enhancing human-in-the-loop adaptive systems through digital twins and VR interfaces. In *2021 International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*, pages 30–40. IEEE, 2021.

[152] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062, 2021.

[153] Jingqing Zhang, Atri Sharma, Luis Bolanos, Tong Li, Ashwani Tanwar, Vibhor Gupta, and Yike Guo. A scalable workflow to build machine learning classifiers with clinicianin-the-loop to identify patients in specific diseases. *arXiv preprint arXiv:2205.08891*, 2022.

[154] Benjamin Smith, Anahita Khojandi, and Rama Vasudevan. Bias in reinforcement learning: A review in healthcare applications. *ACM Computing Surveys*, 56(2):1–17, 2023.

[155] Chuanzan Wang, Tao Huang, Aicheng Gong, Chao Lu, Rui Yang, and Xiu Li. Humanmachine interaction in future nuclear power plant control rooms–a review. *IFACPapersOnLine*, 53(5):851–856, 2020.

[156] Michael R Pinsky, Armando Bedoya, Azra Bihorac, Leo Celi, Matthew Churpek, Nicoleta J Economou-Zavlanos, Paul Elbers, Suchi Saria, Vincent Liu, Patrick G Lyons, et al. Use of artificial intelligence in critical care: opportunities and obstacles. *Critical Care*, 28(1):113, 2024.

[157] Mary Cummings. Automation bias in intelligent time critical decision support systems. In *AIAA 1st Intelligent Systems Technical Conference*, page 6313, 2004.

[158] Xianping Zhong and Heng Ban. Pre-trained network-based transfer learning: A smallsample machine learning approach to nuclear power plant classification problem. *Annals of Nuclear Energy*, 175:109201, 2022.

[159] Ar Ryum Kim, Jinkyun Park, Yochan Kim, Jaewhan Kim, and Poong Hyun Seong. Quantification of performance shaping factors (PSFs)'weightings for human reliability analysis (HRA) of low power and shutdown (LPSD) operations. *Annals of Nuclear Energy*, 101:375–382, 2017.

[160] Khanhvi Tran, Johan Peter Bøtker, Arash Aframian, and Kaveh Memarzadeh. Artificial intelligence for medical imaging. In Adam Bohr and Kaveh Memarzadeh, editors, *Artificial Intelligence in Healthcare*, pages 143–162. Academic Press, 2020.

[161] Jinkyun Park and Hyeonmin Kim. A case study to address the limitation of accident scenario identifications with respect to diverse manual responses. *Reliability Engineering & System Safety*, page 110406, 2024.

[162] Jinkyun Park and Hyeonmin Kim. Addressing the limitations of accident scenario identifications with respect to diverse manual responses affecting the progression of an initiating event. *Available at SSRN 4480466*.

[163] Jong Woo Park and Seung Jun Lee. Simulation optimization framework for dynamic probabilistic safety assessment. *Reliability Engineering & System Safety*, 220:108316, 2022.

[164] Sobhan Moazemi, Sahar Vahdati, Jason Li, Sebastian Kalkhoff, Luis JV Castano, Bastian Dewitz, Roman Bibo, Parisa Sabouniaghdam, Mohammad S Tootooni, Ralph A Bundschuh, et al. Artificial intelligence for clinical decision support for monitoring patients in cardiovascular icus: A systematic review. *Frontiers in Medicine*, 10:1109411, 2023.

[165] Uwe Peters and Mary Carman. Unjustified sample sizes and generalizations in explainable AI research: Principles for more inclusive user studies. *arXiv preprint arXiv:2305.09477*, 2023.

[166] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society*, 35:611–623, 2020.

[167] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. Trustworthy artificial intelligence: A review. *ACM Computing Surveys (CSUR)*, 55(2):1–38, 2022.

[168] Zelun Tony Zhang, Yuanting Liu, and Heinrich Hußmann. Pilot attitudes toward AI in the cockpit: Implications for design. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, pages 1–6. IEEE, 2021.

[169] Roberto Confalonieri, Ludovik Coba, Benedikt Wagner, and Tarek R Besold. A historical perspective of explainable artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(1):e1391, 2021.

[170] Mostafa Amini, Ali Bagheri, and Dursun Delen. Discovering injury severity risk factors in automobile crashes: A hybrid explainable ai framework for decision support. *Reliability Engineering & System Safety*, 226:108720, 2022.

[171] Cheng Zhang, Xiong Zou, and Chuan Lin. Fusing xgboost and shap models for maritime accident prediction and causality interpretability analysis. *Journal of Marine Science and Engineering*, 10(8):1154, 2022.

[172] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 39(2):101666, 2022.

[173] Michael Kläs and Anna Maria Vollmer. Uncertainty in machine learning applications: A practice-driven classification of uncertainty. In Barbara Gallina, Amund Skavhaug, Erwin Schoitsch, and Friedemann Bitsch, editors, *Computer Safety, Reliability, and Security*, pages 431–438, Cham, 2018. Springer International Publishing.

[174] Benjamin Lambert, Florence Forbes, Alan Tucholka, Senan Doyle, Harmonie Dehaene, and Michel Dojat. Trustworthy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. *arXiv preprint arXiv:2210.03736*, 2022.

[175] Ayodeji Oseni, Nour Moustafa, Helge Janicke, Peng Liu, Zahir Tari, and Athanasios Vasilakos. Security and privacy for artificial intelligence: Opportunities and challenges. *arXiv preprint arXiv:2102.04661*, 2021.

[176] Ping Yu, Hua Xu, Xia Hu, and Chao Deng. Leveraging generative ai and large language models: a comprehensive roadmap for healthcare integration. In *Healthcare*, volume 11, page 2776. MDPI, 2023.

[177] Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. Security and privacy challenges of large language models: A survey. *arXiv preprint arXiv:2402.00888*, 2024.

[178] Qiang Liu, Pan Li, Wentao Zhao, Wei Cai, Shui Yu, and Victor CM Leung. A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE Access*, 6:12103–12117, 2018.

[179] Alessandro Biondi, Federico Nesti, Giorgiomaria Cicero, Daniel Casini, and Giorgio Buttazzo. A safe, secure, and predictable software architecture for deep learning in safety-critical systems. *IEEE Embedded Systems Letters*, 12(3):78–82, 2019.

[180] Artificial Intelligence Roadmap. A human-centric approach to AI in aviation. *European Aviation Safety Agency*, 1, 2020.

[181] Florian Tambon, Gabriel Laberge, Le An, Amin Nikanjam, Paulina Stevia Nouwou Mindom, Yann Pequignot, Foutse Khomh, Giulio Antoniol, Ettore Merlo, and François Laviolette. How to certify machine learning based safety-critical systems? A systematic literature review. *Automated Software Engineering*, 29(2):38, 2022.

[182] Ritu Agarwal and Vasant Dhar. Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research*, 25(3):443–448, 2014.

[183] Jon Perez-Cerrolaza, Jaume Abella, Markus Borg, Carlo Donzella, Jesus Cerquides, Francisco J Cazorla, Cristofer Englund, Markus Tauber, George Nikolakopoulos, and Jose Luis Flores. Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey. *ACM Computing Surveys*, 56(7):1–40, 2024.