Proceedings of the ASME 2024 International Design Engineering Technical Conferences
Computers and Information in Engineering Conference
IDETC/CIE2024
August 25-28, 2024, Washington, DC, USA

IDETC/CIE2024-143591

IDENTIFYING HUMAN ERRORS AND ERROR MECHANISMS FROM ACCIDENT REPORTS USING LARGE LANGUAGE MODELS

Lukman Irshad[2,1,*], Hannah Walsh[1]

[1]Intelligent Systems Division, NASA Ames Research Center, Moffett Field, CA 94035
[2]KBR, Inc., Moffett Field, CA 94035

ABSTRACT

Emerging operational concepts for aviation hinge on novel paradigms for human machine interaction. Critical to their safe operation is early consideration of human error into the design process. Existing methods for consideration of human error require significant expert input, which is challenging both in early design and in novel systems for which there is little existing safety expertise. In this research, we propose a methodology for identifying human error, error producing factors, and mechanisms in early design from historical incident reports. Additionally, we hypothesize that cross-domain sharing of lessons learned can aid with early design human considerations in circumstances where data is not relevant or incomplete. This is addressed by identifying causes of human error in aviation and railway domains through applying state-of-the art natural language processing techniques to historical incident reports. Using this method, it is possible to extract extensive reports on human error from past incidents. Using the proposed approach, we identify nine human errors from railway reports and fourteen from aviation reports, with three errors common to both domains. There is at least one error producing conditions for each human error while a majority of the errors have more than one error mechanism. We also found that a majority of the human errors, error producing factors, and error mechanisms (even if they are not common between the domains) can be used to inform safe operations across domains as long as the errors are not domain specific and are interpreted and contextualized using engineering judgement.

Keywords: Human Errors, Hazard Analysis, LLMs, Human Error Mechanisms, BERTopic, sBERT

1. INTRODUCTION

Many emerging aviation operational concepts include new paradigms for human machine interaction, such as m:N (multi-vehicle) operations for Unmanned Aerial Systems (UAS). In the mid-term future, these paradigms form the basis for Advanced Air Mobility (AAM) for passenger use as well as new concepts for emergency response operations [1]. As these new paradigms for human involvement are introduced, it is increasingly important to consider human behavior early in the design, so safety can be built into systems. One way to ensure safety is to consider the human errors during early design hazard and safety assessments to ensure that the system is built from the ground up with an emphasis on mitigating human errors. Currently, however, consideration of human errors, particularly in early design, is limited. This has been recognized by, and is currently being considered by, standards organizations, notably through the formation of the Society of Automotive Engineers (SAE) S-18H Human Considerations for Safety Assessment Committee [2]. Existing methods for assessing human error tend to be expert-driven, require detailed expert interviews, and/or require detailed models or data about the system in order to implement [3]. This detailed information and analysis are often not available in early design stages. Limited research has addressed this challenge of human considerations in early design [4–7]; however, more work is needed, particularly for designs with high novelty.

For novel technologies, there is little safety expertise to rely on, which is critical when relying on expert-driven hazard assessment processes [8]. For instance, there is little historical precedent for autonomy in the aviation domain; however, there is more historical precedent for autonomy in other domains, such as automotive (i.e., self-driving cars) and railway (i.e., driverless trains [9]). It is known that, generally, archetypal mechanisms causing human error have cross-domain relevance. One reason for this is that a common cognitive model and understanding of human behavior can explain many domain-agnostic causes of human error [10]. Consequently, Human Reliability Assessment (HRA) methods and applications tend to be relatively consistent across domains albeit with slight modifications, which are typically based on conditions that affect human performance that tend to be different across domains (i.e., performance shaping factors and human error probabilities). For instance, the Human

*Corresponding author: lukman.irshad@nasa.gov

Error Assessment and Reduction Technique (HEART) has been applied with minor modifications to industrial settings [11], the aviation domain and air traffic management in particular [12], nuclear power [13], railways [14], and maritime applications [15]. Another method, the Cognitive Reliability and Error Analysis Method (CREAM) [16], has been extended for aviation applications to include characteristics seen in long-distance flights [17]. The factors or conditions affecting human performance that are considered in CREAM have also been adjusted for other domains such as maritime [18] and space flight [19]. In sum, with minor modifications or adaptations, human error assessment tools tend to remain largely applicable across domains, indicating the ability of human error-related knowledge to be transferred from one application area to another.

One way in which human error knowledge is documented and used to inform future operations is through incident and accident reports. Historically, it has been difficult to capture lessons in these reports due to sheer volume of information collected – it is simply not feasible for designers to manually review thousands of reports of varying degrees of relevance. However, recent advances with Large Language Models (LLMs) have made possible the rapid, and, for the most part, highly accurate, use of these large repositories of historical documents. It is now possible to discover and extract relevant information quickly and efficiently using state-of-the-art natural language processing techniques. In particular, research on Manager for Intelligent Knowledge Access (MIKA) [20] has previously shown that techniques such as LLM-enabled topic modeling and semantic search can be used within a human-in-the-loop process to extract useful information that can inform early design hazard assessment [21]. Prior research has not, however, specifically analyzed human errors, error producing factors, and error mechanisms. Moreover, it has been assumed that incident reports from a relevant domain must be used, which raises challenges for implementing these techniques for systems with a high degree of novelty. However, as has been established, human errors have substantial cross-domain applicability, and, consequently, this will likely also apply to human errors found in incident and accident reports. As such, for systems with a high degree of novelty, it may be possible to use incident reports from one domain to inform early design hazard assessment of a system in another domain, thereby reducing risk associated with the integration of the novel technology. Even later in the design process, complete enumeration of possible human error causes and mechanisms can assist with human reliability assessment.

The main contributions of this paper are (1) introducing a methodology to identify more causes and mechanisms of human error into design (and particularly early design hazard assessment) and (2) evaluate cross-domain knowledge transfer for human error. To this end, we use the proposed LLM-enabled methodology to (1) identify human errors in incident reports in the aviation domain, (2) identify human errors in incident reports for railways, and (3) assess whether identified human errors-related information have any cross-domain applicability. Human errors are identified from historical domain documents using the natural language processing toolkit Manager for Intelligent Knowledge Access (MIKA) [20]. For the aviation domain documents, aviation incident reports from the National Transportation Safety Board (NTSB) are used [22]. For the railway domain documents, accident investigation reports from the European Union Agency for Railways [23] and NTSB are used. The MIKA workflow includes two natural language processing techniques: topic modeling using BERTopic [24] and information retrieval using semantic search with sentence-BERT [25]. Causes and mechanisms of failure are extracted using this pipeline, with steps for human expert interpretation to disambiguate the natural language processing results as needed. The cross-domain use of the results are then demonstrated and discussed.

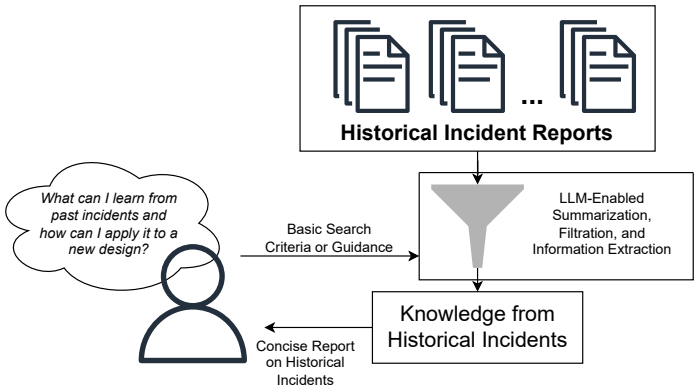

FIGURE 1: The vision for MIKA as an assistive design tool that can extract useful, succinct summaries of information from historical incident reports.

2. BACKGROUND

There is a sizable precedent for the application of Natural Language Processing (NLP) to various tasks in the engineering design process. For instance, NLP has been applied and adapted to use in extracting information from maintenance work orders [26], function knowledge [27], and design ideas [28]. Topic modeling in particular (one of the approaches used in the NLP pipeline in this paper) has been applied successfully to study themes in large sets of incident reports [29, 30]. Significant technical advances in NLP over the past several years (for example, and notably, Bidirectional Encoder Representations from Transformers, or BERT [31], and GPT-4 [32]) have led to numerous novel applications of these technologies. State-of-the-art methods that rely on Large Language Models (LLMs) have demonstrated high performance in diverse tasks, including assistive system modeling [33]. The Manager for Intelligent Knowledge Access (MIKA) has been developed to support diverse knowledge discovery tasks, including topic modeling using multiple applicable algorithms such as BERTopic and Latent Dirichlet Allocation and derivatives [34], and information retrieval tasks, specifically semantic search using sentence-BERT [35]. MIKA has been applied to engineering documents with a particular focus on extracting information related to risk [21]. The vision for MIKA is to extract a succinct report of lessons learned from historical incident reports to assist with early design failure analysis, as summarized in Fig. 1. MIKA has been applied successfully for supporting development of fishbone diagrams [36] and model-based failure modes and effects development [37] using information extracted from historical incident reports. In this study, we extend the processes

developed in this prior work to extracting human errors, error producing factors, and error mechanisms from historical incident reports and comparing the results across domains.

There have been other research efforts dedicated to mining historical safety reports to understand human operators' role in safety, notably to the human's contribution to safety [38]. Researchers have studied different NLP approaches to extracting resilient operator behaviors, including state-of-the-art approaches such as BERT, beyond the standard metadata searches possible in many safety reporting system databases [39]. These efforts differ from the focus of this paper in that they are searching for resilient operator behavior whereas the focus of this research is understanding human error. Each of these concepts represents a useful, but different, goal. In recent years, limited studies have begun to use machine-learning and NLP-enabled approaches to extract various aspects of human performance. In particular, Sawyer et al. used NLP to extract mental health indicators from aviation safety reports (e.g., related to organizational and stress factors) [40, 41]. Other approaches have used machine learning classification approaches to identify causes of human error [42]. Compared to these approaches, this paper considers a broader set of human factors considerations (as opposed to a very specific set such as mental health indicators). Other studies have evaluated the potential use of LLMs for summarization and attribution of human error in incident reports [43]. These are useful methods for improving the value of datasets, but have a different goal than this paper, which is to efficiently learn relevant information from large sets of reports. Building on this recent work, in this study, we use state-of-the-art NLP tools, namely BERT-based approaches to topic modeling and semantic search, to assess whether human error information extracted from one domain is applicable to another, and moreover to evaluate the joint knowledge discovery and information retrieval processes developed and refined in prior work [36, 37].

3. METHODOLOGY

Our primary goal in this research is to study if NLP can be used to extract information related to human error from historic incident/accident reports to inform early design hazard assessments. We further propose an approach to qualitatively assess if data from different domains have cross-domain applicability, so lessons from other domains can be applied to design in the domain of interest. To this effect, we have chosen the aviation and railways domains due to the availability of data. For aviation domain, we use a dataset from the National Transportation Safety Board (NTSB) aviation accident database, which holds investigation reports for all civil aviation accidents and selected incidents that were investigated by NTSB since 1962 [22]. All completed reports for incidents that occurred after January 1^{st}, 2020 were chosen for this study, resulting in a total of 6501 investigation reports. For railways domain, data from investigation reports from NTSB railways investigation database and European Railway Accident Information Links (ERAIL) database were selected. The NTSB railways investigation database holds investigation reports for accidents that occurred since 2010 [22], which amounts to 160 total reports. The ERAIL database, which is maintained by the European Union Agency for Railways holds all accident investigation reports that were submitted by the member nations since 2006 [23], amounting to 3707 total reports. The total number of investigation reports recorded here are a snapshot from when the data was downloaded on February 28^{th}, 2024. Note that the aviation data we use for this study is only a subset of the available data whereas we use all available data from the railways databases. The main reason for this is that we want to maintain a large enough dataset to demonstrate how NLP can be used to extract human error related information when it is not feasible manually, while also making sure that it is not too large to where it makes this demonstration convoluted.

With the datasets identified, we expand upon previous applications of MIKA [36, 37] to extract human error related information. When assessing human errors, the error producing conditions (also known as performance shaping factors or performance influencing factors) must be accounted for to fully understand the errors and the context in which they are produced (as done in a majority modern human reliability assessment methods such as CREAM [16] and ATHEANA [44]) [45]. Recent research has also identified that understanding the underlying error mechanisms is also important to understanding human error [45]. The understanding of these error producing conditions and error mechanisms can help in deriving mitigations and design recommendations that will minimize the likelihood of human errors. For example, if the human error is a pilots failure to perceive a signal, understanding underlying contributing factors such as (lack of) attention or fatigue that caused the missed perception have important implications for design. Hence, the knowledge extraction related to hazards caused by human error must account for these error causing conditions and mechanisms. Additionally, this research aims to explore the cross-domain applicability of historic incident data, so knowledge extracted from one domain can be used in another to complement safety assessments. Accounting for these needs, we first propose an approach (as summarized in Fig.2) to identify human error caused hazards, human error causing conditions, and contributing factors from historic documents. Next, we propose a qualitative assessment that can help study the cross domain applicability of the extracted knowledge.

3.1 Extracting Human Error Related Knowledge from Historic Incident Reports

As shown in Fig. 2, the process for extracting human error related knowledge from historic documents involves three high level steps; (1) pre-processing the data, (2) identifying human error related hazard themes, and (3) extracting more details about the hazard themes to generate human error related knowledge. In the following subsections, we explore each of these steps and provide details on the application of these steps to the chosen aviation and railways datasets.

3.1.1 Pre-processing the Data. The goal of the pre-processing step is to identify any natural language fields of interest in the dataset while also making sure that there are no anomalies in the data. With the NTSB datasets (both aviation and railways), the column of interest was "Probable Cause," which included a natural language description of the cause of the accident/incident. However, this column was blank for some records. We omitted the blank probable cause records because there was no natural

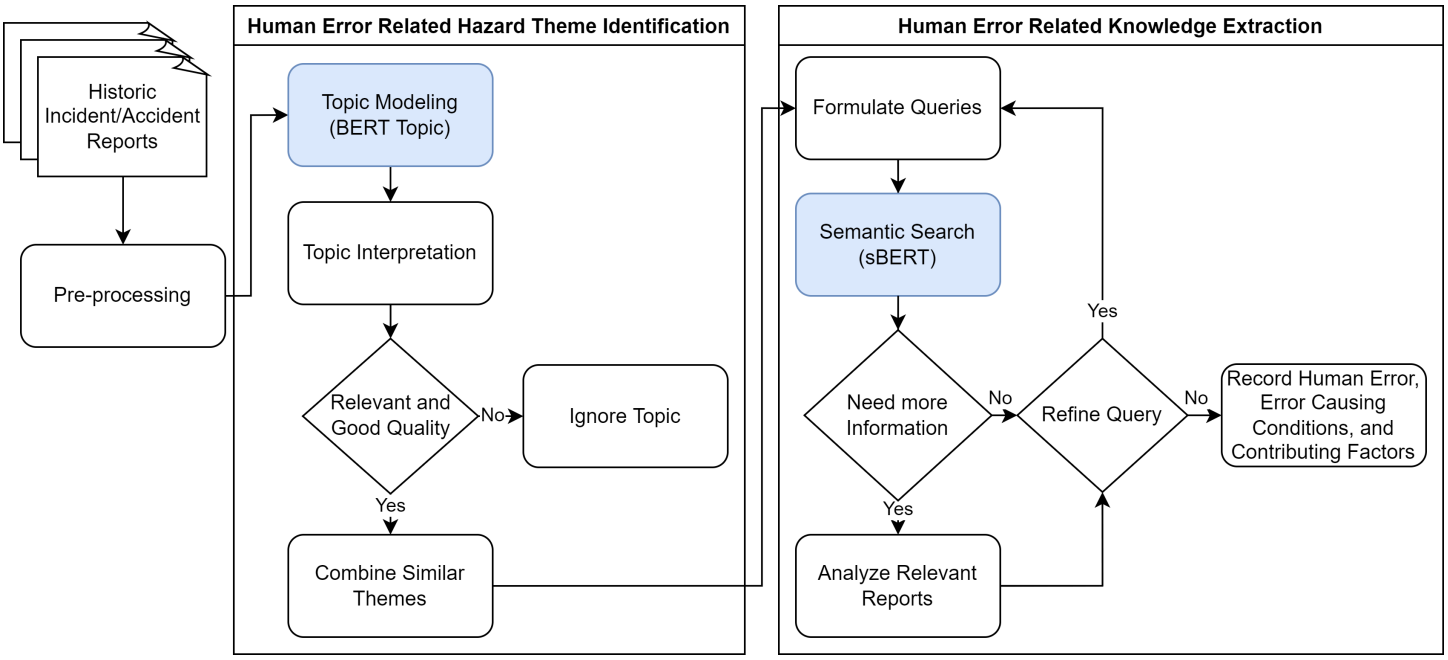

FIGURE 2: An approach to extracting human error related knowledge from historic incident/accident reports using Large Language Models (LLMs) via the MIKA toolkit, where the shaded boxes indicate steps performed through MIKA and white boxes indicate steps performed by experts.

language to process, resulting in 3,826 and 127 remaining records for aviation and railways, respectively. With the ERAIL dataset, the columns of interest were "Direct cause description" and "Underlying root causes description." As with the NTSB datasets, any records with both these columns left blank were omitted from the study. Additionally, some entries had languages other than English and some had a different language and an English translation. We omitted any entry that had only non-English descriptions and removed the non-English portion from the description for entries with English translations. After this omission, 2,852 entries remained in the dataset. Finally, we combined both the NTSB railways and ERAIL datasets (by appending the "Probable Cause" column in the NTSB dataset with "Direct cause description" column in the ERAIL dataset) since these datasets are similar in nature and the extracted hazards may be duplicated if they are analyzed separately. This is also a method to manage the occasional challenge of contributors entering very short descriptions into individual fields (columns). After the pre-processing of the data, the total entries to be processed were 3,826 and 2,979 for aviation and railways, respectively.

3.1.2 Human Error Related Hazard Theme Identification. To identify human errors, error causing conditions, and contributing factors from historic documents, we must first understand the types of human error related hazards present in the documents, so specific information related to these hazards can be extracted. We propose topic modeling to identify these human error caused hazard themes. Topic modeling returns themes which are represented through a list of words that are shared among documents that contain the theme [36]. Past research has used topic modeling to extract hazard themes [34, 36, 37]. While there are many different topic modeling approaches (e.g., Latent Dirichlet Allocation (LDA) [46] and Hierarchical Dirichlet Process (HDP) [47]), in this research, we use BERTopic [24] topic modeling via the MIKA toolkit because of its ease of use and its ability to produce high-quality results [36]. In particular, BERTopic tends to produce highly human-readable topics compared to other techniques. In this method, documents are transformed using an embedding model into a vector representation, in particular using the sentence transformer model, which embeds sentences [25]. This method leads to improved context understanding compared to methods that embed words individually. Once topics are modeled, there are several options for representing them. Past work has suggested that short phrases can represent topics better than single words [48]. We set the number of words per topic to ten to make sure that enough words are returned per theme to be able to build a coherent interpretation while avoiding over-constraining the topic. Additionally, we use n-grams with $n = 3$ as the tokens used in the algorithm rather than single words. An example of an n-gram is "Space Shuttle Program," in which these three words are treated as a single entity rather than three separate words. This is a meaningful option to use when there are key phrases that have a specific meaning apart from their individual words, as is often the case with engineering texts. Finally, we set the minimum number of documents a theme should be represented in to three to make sure that we are extracting themes that are relatively common in the accident reports. The aviation and railways datasets returned 260 and 208 total topics, respectively.

The next step in the human error related hazard identification is to interpret the topic modeling results to identify the hazard themes. This must be done based on expert judgement

TABLE 1: A sample of topic modeling results and their interpretation from the aviation data with highlighted rows showing ignored topics.

	Topic Words	Hazard Theme Interpretation
1	disorientation, spatial disorientation, spatial, meteorological conditions, meteorological, instrument, instrument meteorological conditions, instrument meteorological, control spatial, control spatial disorientation	Pilot was using instruments due to poor weather, which caused spatial disorientation.
2	failure maintain proper, maintain proper, glidepath, proper, proper glidepath, path, glide path, maintain proper glidepath, glide, approach	Pilot fails to maintain proper glide path during the approach.
3	power takeoff, throttle, engine power takeoff, power takeoff undetermined, takeoff undetermined reasons, takeoff undetermined, undetermined reasons, properly secure throttle, secure throttle, takeoff reasons determined	Pilot fails to properly secure throttle during takeoff.
4	decision abort, decision abort takeoff, delayed decision abort, abort, abort takeoff, delayed decision, pilots delayed decision, pilots delayed, delayed, abort takeoff resulted	Pilot's delayed decision to abort takeoff resulted in some incident.
5	collision bird, inflight collision bird, bird, inflight collision, bird inflight, collision bird inflight, bird inflight collision, inflight, collision, aerobatics inflight collision	Bird strike during the flight. Ignored because this is not related to human error.
6	grass, wet, turf, grass runway, wet grass runway, wet grass, turf runway, grass pilots, grass pilots decision, grass runway resulting	Ignored due to poor quality.

to ensure that the group of topics are converted to actionable hazards. A sample of topics from the BERTopic results and their interpretation are presented in Tables 1 and 2 for aviation and railway datasets, respectively. Next, an expert must identify hazard themes that are high quality and relevant to human error and ignore the remaining topics. In this study, in addition to non-human related hazard themes and poor quality themes, any hazard themes that were caused by external factors (e.g., weather with no human element, pedestrians in railway level crossings, and road vehicle causing railway accident) were ignored, resulting in 83 and 42 themes for aviation and railway datasets, respectively. The shaded rows (last two) in Tables 1 and 2 show examples of themes that were ignored in both datasets. The final step in human error related hazard theme identification is to use human factors and systems engineering judgement to merge similar themes to create a final list of unique human error related hazard themes present in the historic documents. In the case of the Aviation and Railways datasets, 25 and 20 unique human error related hazard themes were identified.

3.1.3 Human Error Related Knowledge Extraction. While the themes identified can give information on the high level human error caused hazards, they give very little information on the error producing conditions and mechanisms. To retrieve information related to error producing conditions and mechanisms, the dataset must be further analyzed. Even though a keyword search on the documents relevant to each theme can give some details on error producing conditions and mechanisms, a better approach would be to perform a semantic search, so the search is "context aware," yielding more accurate search results. We propose formulating a few queries for each identified theme to perform the semantic search, so the search is not too narrow or broad. The hazard themes identified in the previous steps keep the queries relevant to the data present in the documents, which is impossible without deep knowledge about the dataset. In other words, the human error related hazard identification step removes the need for expert knowledge about datasets, enabling the assessment to be performed by practitioners who are even new to the dataset. In the case of the railway and aviation datasets in this research, we formulated one to three queries per identified theme. When formulating the queries, we set out to extract more details on each theme by approaching the theme through a variety of subjects (i.e., causes, consequences, and error circumstances), as appropriate for each theme. For example, for the theme "Pilot exceeds angle of attack" in Table 3, the first query covers causes, the second covers consequences, and the last covers error circumstances. For some themes, one or more of these aspects may already be apparent from the theme description itself (such as the circumstances, for example), in which case fewer queries may be necessary. A sample of identified themes and formulated queries for each theme for both datasets are presented in Table 3. Note that the query formulation is a trial and error-based iterative process, where the queries can be refined based on the search results to improve the relevance of the results.

To perform the semantic search, we use MIKA's sentence-BERT based semantic search (information retrieval) capability. It is possible when replicating the approach used in this paper to substitute other search methods. However, semantic search, being context aware, shows substantial improvement over keyword search. For example, when searching for documents about cybersecurity incidents, a keyword search might return documents about physical building security due to the two concepts sharing the word "security" even in different contexts [35]. As we are already using MIKA for BERTopic, the datasets are already loaded into MIKA, meaning using MIKA's semantic search capability is not too time-consuming. The search capability is asymmetric: a short phrase or question, similar to what you might enter into a standard web search engine, is used to represent an information need and query the dataset, and a ranked list of documents is returned (in contrast, a symmetric search would require either

TABLE 2: A sample of topic modeling results and their interpretation from the railways data with highlighted rows showing ignored topics.

	Topic Words	Hazard Theme Interpretation
1	inattention driver drive, driver drive, inattention driver, inattention, drive, drive inattention driver, driver drive inattention, drive inattention, following signs, following signs driver	Driver fails to follow signs due to not paying attention.
2	traffic, traffic controllers, controllers, systems, traffic control, wrong routings, train 1b78, routings, 1b78, driver train 1b78	Traffic controller routing the trains incorrectly.
3	train dispatcher, error train dispatcher, dispatcher, error train, trains train dispatcher, 52760, train dispatcher derailment, train dispatcher deficient, train communicate operational, train communicate	Communication error between the train dispatcher and driver.
4	acoustic warnings, acoustic, warnings, crossing time forbidden, time forbidden, light acoustic warnings, failure respect light, drivers failure respect, level crossing time, respect light	Driver fails to adhere to acoustic and visual warnings.
5	inattention, inattention pedestrian crossing, inattention pedestrian, pedestrian crossing, inattentiveness driver user, inattention person, music inattention, near crossing inattention, oncoming train use, pedestrian crossing platform	Accidents caused at level crossings due to the inattention of persons other than the train driver. Ignored because it is not relevant to the study.
6	bearing, roller, bearings, roller bearings, heat load, axle, old, afferent wheel, overheating, wheelseat	Mechanical failure causing the incident. Ignored due to irrelevance to this study.

TABLE 3: A sample of identified themes and derived queries from aviation and railways datasets.

dataset	Theme	Queries
Aviation	Pilot exceeds angle of attack	What are some causes for pilots exceeding the angle of attack?
		What are consequences of exceeding angle of attack?
		What are circumstances that lead to exceeding angle of attack?
	Runway excursion due to loss of control	What are some causes of runway excursions?
		What pilot errors lead to runway excursions?
		Under what circumstances do runway excursions occur?
Railways	Communications failure	Can communications failure result in accidents?
		What are some causes of communications failure?
	Failed to perceive light and acoustic warning signals	What are some causes of failure to perceiving light signals?
		What are some causes of failure to perceiving acoustic warnings?
		What leads to drivers not respecting warnings?

a longer text or passage as query with a similar length document expected to be returned, or a shorter text returned from a shorter query). Semantic search with sentence-BERT embeds documents and queries into a vector space, where it is possible to match query embeddings with closely related document embeddings, with closeness dictated by semantic similarity rather than lexical similarity [25]. The sentence transformer model (sentence-BERT) is essential to enabling this process. BERT models are trained on large datasets (i.e., hundreds of thousands to millions of documents) and their training is carried out through a Masked Language Model (MLM), in which the model is asked to predict a masked word [31]. The model is asked to predict the masked word from both left and right ("bidirectional"), which improves context understanding [31]. Sentence-BERT differs from the standard BERT method in that it uses a specialized transformer model for sentences (the Siamese transformer, essentially two BERT networks with a pooling operation), which is more computationally efficient for sentence-level understanding [25]. It is possible to fine-tune sentence-BERT models on domain-specific datasets [35]. However, in prior studies, we have found pre-trained models perform sufficiently well [35], as such, we have chosen the pre-trained model for this study.

The ranked list of most relevant documents returned by the search can be reviewed by experts using human factors and systems engineering judgement to obtain information related to human errors, error producing conditions, and error mechanisms. An example of doc strings from the returned documents for the query "what leads to drivers not respecting warning?" are presented in Fig. 3. In some cases, the structured datasets used may not contain the entire reports content. For example, the field "Probable Cause" in the NTSB dataset is the only field with natural language, and at times is just a summary of the contents of the reports. If the details in the short summaries in the dataset are not sufficient (as in the second example presented in Fig. 3), experts may directly refer the specific document, which may contain more details than the database entries. Experts may also refine the theme or query based on the returned information and repeat the search, if needed. In the case of the aviation and railway datasets

....the braking being lately engaged by the driver.....The driver concerned had little experience. To interpret the signal, he was using his experience in the area i.e. the fifty or so times he had passed through there (as the investigation was able to establish).
There is a risk for drivers little experienced to incorrectly read a signal in a similar operational situation due to their expectations. The day of the accident, the driver carried out his third service after a long work interruption of almost 8 weeks. The attention deficit could easily be explained by the fact that the driver was returning from a long holiday.....explained by the weakening of automatic reflexes, the cognitive solicitations in relation to the activity and which maintain them being suspended during the holiday period. This effect is all the more marked when the reflexes have not yet fully taken root, which is the case for inexperienced drivers.....

The accident had its origin in a human failure due to the non-compliance of the signal instructions by the train driver.

The collision caused by a delay was due to the fact that the driver of the passenger train ignored a deactivated signal after not having used the braking system.....
...the driverâ€™s level of alertness was not optimal.....During the operational sequence, the driver of the passenger train arrived at signal D.11 showing a â€œDouble Yellowâ€ aspect which indicates to the driver that the following signal (signal B222) is to be considered closed. The driver is late to acknowledge the restrictive signal, that is to say after having passed the signal but within the 4 seconds time frame allowed. Beyond 4 seconds, an emergency brake would have been engaged.....

FIGURE 3: Top three most relevant documents returned by the semantic search for the query "what leads to drivers not respecting warning," where documents one and three are redacted to manage space. Text is printed exactly from source. The highlighted text shows the text that is relevant to the query.

in this research, we chose to look into the top five semantic search results for each entry and explore the detailed reports (if available and written in English) when the database entries are not sufficient to extract the needed information. While studying the search results, we first identified human errors and error producing conditions, and next tried to draw connections between them to understand the error mechanisms if it is not explicitly stated. To be conservative, we took an exhaustive approach, where we did not discard any error producing conditions and mechanisms even if they are only mentioned in one report. The resulting human errors, error producing conditions, and mechanisms are discussed in detail in the section 4.

3.2 Assessing the Cross-domain Applicability of Datasets

Assessing data for cross-domain applicability can not only help fill gaps in human error related insights in one domain with insights in another domain (e.g., when designing UAS's, data from self-driving car industry can have some human error-related insight that is useful) but also help practitioners think deeply about the elicited human errors. Additionally, qualitative assessments of human errors can be more helpful in deriving mitigations [45]. Thus, we propose a qualitative assessment for studying cross-domain applicability of human error related information extracted from historic reports. The first step to assess cross domain applicability is to select human errors that are common between datasets and build a graph (as shown in Fig. 4) with the aviation error in the left, railway error in the right, and the error producing conditions from both datasets in the middle. Next, the relationships between the error and error mechanisms (if

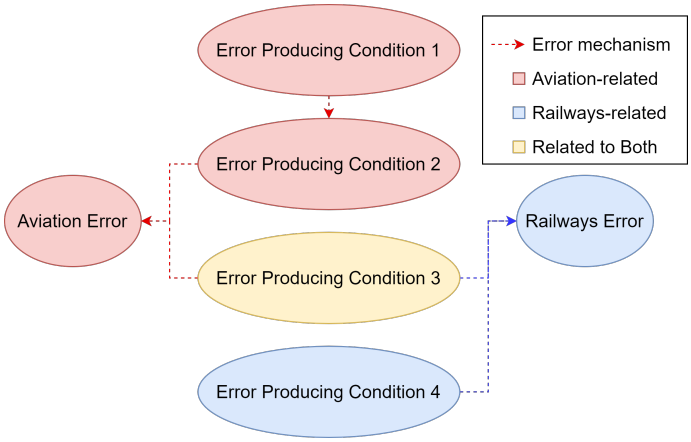

FIGURE 4: A generic representation of human errors common between domains with color shades representing the relationships between human error themes and error producing conditions and dotted arrows showing the error mechanisms.

feasible) are represented by color coding and connecting blocks as shown in Fig. 4. For example, the aviation hazard has error producing conditions 1, 2, and 3 mentioned in the reports. The error mechanism starts with error producing condition 1 leading to error producing condition 2, which then combines with error producing condition 3 to produce the error. We then aim to answer the following questions through the graph.

- Are there any overlapping human error mechanisms?
- Are there any error producing conditions that are common for both domains?
- If there is no overlap, are there any relationships between error producing conditions and unconnected errors
- If there are new connections, can we derive any new error mechanisms?

For human error themes that are not common between the two domains, we first omit any errors that are domain specific (e.g., failing to extend the landing gear or completing a prelanding checklist are errors specific to aviation and have no applicability in railways). Next, we use engineering judgement to contextualize the remaining human errors and error producing conditions to make them relevant to the domain of application. We finally explore if the contextualized human error related knowledge is applicable across domains.

4. RESULTS

A subset of the human errors, error producing factors, and error mechanisms extracted from the railways and aviation dataset using the proposed LLM-enabled pipeline are presented in Tables 4 and 5, respectively. We identified nine and fourteen human errors each from the railways and aviation datasets. Note that the number of human errors is less than the number of actual human error related themes identified. This reduction is expected because the topic modeling methodology extracts general themes,

TABLE 4: A subset of the Human Errors, Error Producing Factors, and Error Mechanisms from the Railways dataset.

Human Error	Error Producing Factors	Error Mechanisms
Misrepresenting signals	Task repetition, task infrequency, habitual driving, unfamiliar environment, inattention, poor signal design, poor regulations, poor safety culture	Task repetition can lead to habitual driving and task infrequency can lead to unfamiliar environment. Similarly, poor regulations can lead to poor safety culture. All of these can result in inattention, this with or without poor signal design can lead to misrepresenting a signal.
Not complying with signals	Distraction, poor alertness, inexperience, weakening automatic reflexes (due to working after a long break), inadequate training, reduced visibility, lack of situation awareness, poor task design (e.g., lack of communication protocols), poor signal design, poor interface design (e.g., warning system)	Poor interface design can result in distraction, leading to poor alertness. This can combine with poor signal and interface design, leading to not complying with signals. Inexperienced operators working after long breaks can lead to weakening of automatic reflexes, this combined with poor signal design can lead to not complying with signals. Reduced visibility can lead to poor situation awareness which can lead to not complying with signals.
Poor route planning	Poor safety culture and regulations, lack of situation awareness, poor workspace design, mismatch in mental model, lack of trust in automated system, high workload, inadequate training, organizational factors, poor interface design, poor interaction (communication channel) design, poor operating procedures	Poor safety culture and regulations and poor workspace design can lead to diminished situation awareness, leading to poor route planning. Inadequate training can lead to mismatch in the mental model. This combined with poor interface design can result in lack of trust in automated systems, resulting in them not being used as they should, increasing the workload. This combined with poor interaction design, safety regulations, and organizational factors can lead to poor route planning.
Late/no braking	Fatigue, poor shift design, poor workspace design, poor regulations, poor safety culture	Poor shift design can lead to fatigue, which can couple with poor workspace design, regulations and/or safety culture resulting in late/no braking.
Over speeding	Fatigue, poor alertness, intoxication, poor shift design, poor regulations, poor interface design, inattention	Intoxication can lead to poor alertness, leading to over speeding. Poor interface design can lead to inattention, which results in over speeding. Fatigue leads to poor alertness. This coupled with poor regulations can lead to drivers over speeding.

not final human errors, and expert interpretation is needed and desired to ensure quality of the final results. Additionally, the number of errors may change depending on the level of abstraction. For the purposes of this assessment, we maintain the high level of abstraction because it is needed for analyzing the cross-domain applicability (so similar errors can be compared and contrasted) of the human error related knowledge. We find numerous error producing conditions and at least one error mechanism for each human error, except for two human errors in the aviation dataset (e.g., last entry in Table 5) that has one error producing condition and no error mechanisms. In general, we find more error producing conditions per human error and more detailed error mechanisms in the railways dataset. This difference is mainly because the aviation dataset (or the NTSB reports) did not detail the human error producing conditions or human error mechanisms as the railways dataset did. This discrepancy is one of the motivations for studying the cross-domain applicability of these datasets (for which the results are detailed later in this section), so when data is not readily available or data quality is lacking in one domain, practitioners can use information from other domains with better data to fill gaps in information in their assessments.

The human errors presented in Tables 4 and 5 can be traced from the LLM output. For example, (1) in Table 2, the fifth topic is interpreted as "driver fails to adhere to acoustic and visual warnings;" (2) in Table 3, under the same theme, the query "What leads to drivers not respecting warnings?" is used to elicit further details; (3) in Fig. 3, the query results in documents describing incidents in which a driver did not respect a warning, with associated details highlighted; and (4) in Table 4, these details are used to inform expert analysis of the derived human error, "not complying with signals," with error producing factors including "poor alertness" (compare to the highlighted text in Fig. 3, "[driver's] level of alertness was not optimal") and error mechanisms including "inexperienced operators" (compare to the highlighted text in Fig. 3, "a risk for drive's little experience to incorrectly read a signal"). This process requires additional interpretation for the aviation dataset because the error producing conditions and error mechanisms were not explicitly stated. For example, if the text described an aircarft experiencing a sudden emergency and the pilot's response to it, we interpret the error producing conditions to be high stress. Similarly, if the accident happened during a high workload phase of flight (e.g., landing, takeoff), we include high workload into the error producing conditions even when it is not explicitly stated in the description.

TABLE 5: A subset of the Human Errors, Error Producing Factors, and Error Mechanisms from the Aviation dataset.

Human Error	Error Producing Conditions	Human Error Mechanisms
Exceeding angle of attack	Stress, high workload, distraction, unfamiliar environment (e.g., trees to clear during takeoff), intoxication	Unfamiliar operating environments can increase workload and stress, which can result in distractions, leading to exceeding the angle of attack. Intoxication can lead to poor judgement which can result in exceeding the angle of attack.
Failure to follow prelanding checklist	Time pressure, poor interface design, inattention, high workload	High workload can lead to inattention, this and/or poor interface design can result in pilots failing to follow prelanding checklists. Time pressure can lead to pilots failing to follow prelanding checklists.
Loss of spatial orientation	Inadequate training, inexperience, low visibility, high workload, distraction, maintenance issues (e.g., failed deicing equipment), stress	Maintenance issues can lead to high workload and stress, which can lead to distraction. This coupled with low visibility conditions can result in loss of spatial orientation. Inadequate training and inexperience coupled with high workload in low visibility conditions can lead to loss of spatial orientation.
Loss of directional control during landing/takeoff	Stress, operating conditions (e.g., wind), high workload, inexperience, distraction	High workload coupled with inexperience can lead to stress and distraction. This and/or adverse operational conditions can lead to directional control loss.
Inadequate preflight inspection	Poor judgement	N/A

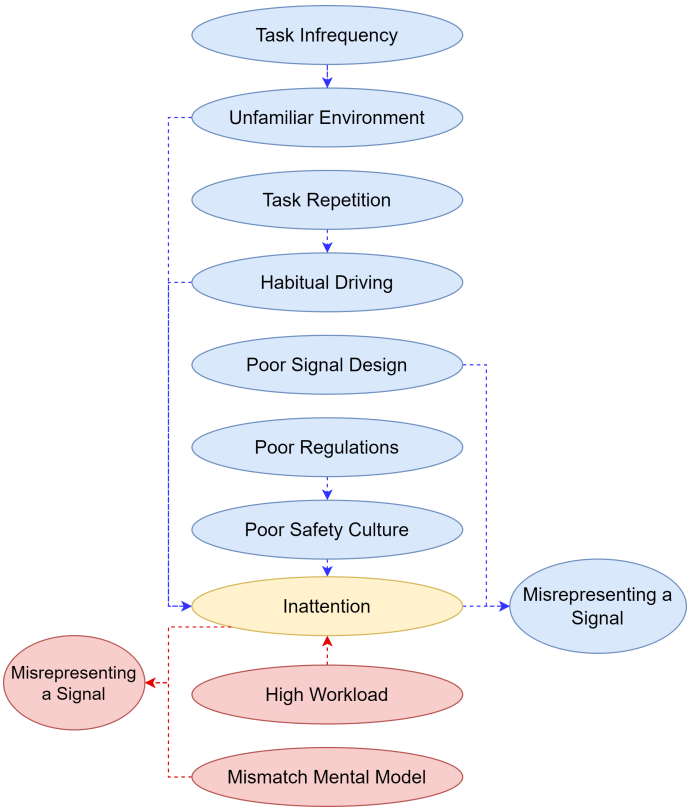

FIGURE 5: Cross-domain comparison for the error "misrepresenting a signal," where the red and blue color arrows represent aviation, and railways error mechanisms, respectively.

Note that for each of the error mechanisms in Tables 4 and 5, it is not necessary for all of the error producing conditions to be present for an error to occur. They instead show a general flow of how the error producing conditions interact to produce errors. For example, the first error mechanism for the human error "loss of spatial orientation" in Table 5 starts with maintenance issues leading to high workload and stress, which can result in distraction. When distraction is coupled with low visibility it can lead to loss of spatial orientation. This mechanism must be interpreted as having low visibility with or without distraction can lead to "loss of spatial orientation," while stress and high workload can lead to distractions with or without having maintenance issues. From a design perspective, designers may use the results from these assessment to inform human factors considerations early on. For example, if they are designing a system to automate route planning in railways, they may develop requirements to minimize error producing factors for the "poor route planning" error (e.g., the interface design should help operator attain high levels of situation awareness) in Table 4. Accounting for the second error mechanism, they may have strict training requirements to ensure that the operator can maintain an appropriate mental model and trust the system, so they use it appropriately in practice to ensure that their workload is not too high. They may also set interface design requirements to remind users to use the automated system to ensure that it is used when it should be. Having these considerations early on in design can help designers build safety into the system proactively.

To perform the cross-domain applicability study, we first identified common errors between the two domains, specifically looking for common themes rather than looking for an exact match, which resulted in three identified common errors, namely "failure to comply with operating procedure," "failure to maintain speed," and "misrepresenting a signal." Among the three human errors, "misrepresenting a signal" had one common error

producing factor (inattention, as seen in Fig. 5) while the others had none. Consequently, the mechanism related to the inattention factor for "misrepresenting a signal" had an overlap, while the mechanisms for the other errors had no overlap. However, all of the error producing conditions for each common error, even if they had no common connections between domains, were applicable for both domains. For example, all of the error producing factors that were unique to aviation domain (high workload and mismatch in the mental model) were applicable to the railways domain and vise versa for the "misrepresenting a signal error." However, engineering judgement must be exercised to account for the context of these conditions. For example, the factor habitual driving from the railways domain must be contextualized for the aviation domain (i.e., consider it as habitual piloting tasks) to ensure they are applicable in the aviation domain. Similarly, the mechanisms were also common for both domains even when no explicit overlaps were identified. Among the errors that were not common across domains, four of the six uncommon railways errors ("poor route planning," "poor signaling," "not complying with signals," and "failure to perceive signals") and related information (error producing conditions and mechanisms) were applicable for aviation. Three of the eleven uncommon aviation errors ("failure to communicate,"loss of control," and "failure to maintain clearance in low altitude flights") and related information were applicable for railways. As with the common errors, engineering judgement had to be used to contextualize the errors to make them applicable for the other domain. For instance, the railways error "poor signalling" can be taken in the context of air traffic control providing poor information. When this is contextualized, the error producing factors (lack of situation awareness, poor interface design, lack of trust in automated system, distraction, inattention, and stress) and the mechanisms ((1) poor interface design and lack of trust in automated systems can lead to poor situation awareness, leading to poor signaling and (2) distractions can lead to inattention, resulting in poor signaling) become relevant to the aviation domain.

5. DISCUSSION

In this research, we have demonstrated how the LLM interface in the MIKA toolkit can be used to extract human errors, error producing conditions, and human error mechanisms from incident reports from two domain datasets (aviation and railways). The results indicate that the human error related knowledge extracted can be valuable early on in design in helping with human considerations being built into the system. One of the challenges in considering the human elements in early design hazard assessment is that hazard assessment methods often rely on task analysis to identify human errors. Task analyses are often conducted later in design when system is designed using a variety of sources (e.g., expert interviews and surveys, past incidents, etc.). This approach allows designers consider the human without any task analysis, which makes it usable for early design hazard assessment approaches such as functional hazard assessment. This approach can also complement expert driven safety assurance approaches by helping them extract knowledge from historic data, rather than only relying on their judgement. For example, this approach can benefit simulation based hazard assessment tools that simulate human error propagation by helping experts setup their models by complementing their expertise with the knowledge extracted through this process. This approach has some implications for human reliability assessments as well. One of the challenges of human reliability assessment methods is identifying human failure events [45]. This approach identifies human failure event through the hazard themes. The human failure events along with human error mechanisms which are identified as part of this approach can complement the human reliability assessments.

Knowledge extracted using the presented approach is only as good as the data source. So, the results may not be comprehensive. For example, for a specific error, the approach may not find all possible error producing conditions or mechanisms. From a safety perspective, this can lead to poor considerations of hazards. As a means of overcoming this challenge, we have defined the approach to be fully expert driven where automation is used to aid with only interpreting documents. In other words, the hazard elicitation will only be complete once the experts interpret the results and complete any missing information. We see this as a strength of this approach rather than a weakness due to two reasons. First, the approach encourages designers to systematically think about the human considerations of the system more deeply early on, which can result in them considering factors that they might not have otherwise, minimizing the need for design changes and workarounds later. Next, the approach, through the historic data can help designers validate their assumptions, resulting in less uncertainty and more confidence in the analysis. Additionally, we have shown through a qualitative assessment that some of the errors have cross-domain applicability as long as an expert is accounting for domain specific assumptions and constraints. The results indicate there are causes of human errors that are common across domains (aviation and railways) as well as causes that are found in one domain dataset but not the other. This is consistent with prior research findings from extending human reliability methodologies from one domain to another.

To summarize, this research shows that LLMs via the MIKA toolkit can be used to elicit human errors, error producing conditions, and error mechanisms early in design. The findings in this paper indicate that it may be possible, for example while eliciting hazards during early design (i.e., for functional hazard assessment), to use lessons learned from an established domain with many documents available to inform a domain that has fewer historical incident reports to learn from. Typically, hazard elicitation requires expert interviews, which are invaluable but take time. With this approach of eliciting hazards using historical incident reports, including from other domains where needed, it may be possible to consider human errors earlier in the design process (while refining the set of hazards with expert input once available). This approach gives the flexibility for practitioners to elicit hazards at varying granularity (limited only by the data availability), tailored to their applications and assessments, which makes the approach usable at different design stages (e.g., component hazard assessment, system hazard assessment, preliminary safety assessment, etc.). All of these factors will be essential in assuring the safety of emerging operational concepts where autonomy and human machine interaction are central and critical.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have shown that human errors, extracted using a natural language processing pipeline from historical incident reports, can be applicable across domains (aviation to railways and vice versa). This is consistent with existing knowledge of human reliability, but has implications on the use of lessons learned from other domains to inform early design failure analysis activities such as functional hazard assessment. Moreover, the human errors identified in this study can be used to include human considerations in early design failure analysis for emerging aviation operational concepts. Future work will extend the demonstrated process to other datasets, notably to incident reports and lessons learned documents related to self-driving cars. Moreover, research efforts into the inclusion of a MIKA-based design assistant for functional hazard assessment are ongoing. Work remains to be completed to specify requirements for such an assistant, develop a prototype, and perform user studies.

ACKNOWLEDGMENTS

This research was funded by the System-Wide Safety project in the NASA Aeronautics Research Mission Directorate. The findings herein represent the research of the authors and do not necessarily the view of the U.S. Government or NASA. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the U.S. Government.

REFERENCES

[1] Walsh, Hannah S, Spirakis, Eleni, Andrade, Sequoia R, Hulse, Daniel E and Davies, Misty D. "SMARt-STEReO: Preliminary concept of operations." Technical Report No. NASA/TM-20205007665. National Aeronautics and Space Administration (NASA). 2020.

[2] "Scoping Document: S-18H Human Considerations for Safety Assessment Committee." https://standardsworks.sae.org/standards-committees/s-18h-human-considerations-safety-assessment-committee#. Accessed: 24/3/8.

[3] Demirel, Hasan Onan. "Modular human-in-the-loop design framework based on human factors." Ph.D. Thesis, Purdue University. 2015.

[4] Irshad, Lukman, Demirel, H. Onan and Tumer, Irem Y. "Automated Generation of Fault Scenarios to Assess Potential Human Errors and Functional Failures in Early Design Stages." *Journal of Computing and Information Science in Engineering* Vol. 20 No. 5 (2020): p. 051009. DOI 10.1115/1.4047557. URL https://asmedigitalcollection.asme.org/computingengineering/article-pdf/20/5/051009/6647557/jcise_20_5_051009.pdf, URL https://doi.org/10.1115/1.4047557.

[5] Irshad, Lukman and Hulse, Daniel. "Resilience Modeling in Complex Engineered Systems With Human-Machine Interactions." *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 86212: p. V002T02A024. 2022. American Society of Mechanical Engineers.

[6] Patriarca, Riccardo, Di Gravio, Giulio, Woltjer, Rogier, Costantino, Francesco, Praetorius, Gesa, Ferreira, Pedro and Hollnagel, Erik. "Framing the FRAM: A literature review on the functional resonance analysis method." *Safety Science* Vol. 129 (2020): p. 104827.

[7] Leveson, Nancy. "STPA (System-Theoretic Process Analysis) Compliance with MIL-STD-882E and other Army Safety Standards." (2016)URL http://sunnyday.mit.edu/compliance-with-882.pdf.

[8] McCormick, Frank, Graydon, Mallory, Neogi, Natasha, Miner, Paul and Maddalon, Jeffrey. "Safety Expertise and the Perils of Novelty." *2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC)*: pp. 1–10. 2023. DOI 10.1109/DASC58513.2023.10311181.

[9] Ramírez, Roberto Carlos, Adin, Iñigo, Goya, Jon, Alvarado, Unai, Brazalez, Alfonso and Mendizabal, Jaizki. "Freight Train in the Age of Self-Driving Vehicles. A Taxonomy Review." *IEEE Access* Vol. 10 (2022): pp. 9750–9762. DOI 10.1109/ACCESS.2022.3144602.

[10] Pan, Xing, Lin, Ye and He, Congjiao. "A Review of Cognitive Models in Human Reliability Analysis." *Quality and Reliability Engineering International* Vol. 33 No. 7 (2017): pp. 1299–1316. DOI https://doi.org/10.1002/qre.2111. URL https://onlinelibrary.wiley.com/doi/pdf/10.1002/qre.2111, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qre.2111.

[11] Williams, J.C. "A data-based method for assessing and reducing human error to improve operational performance." *Conference Record for 1988 IEEE Fourth Conference on Human Factors and Power Plants,*: pp. 436–450. 1988. DOI 10.1109/HFPP.1988.27540.

[12] Kirwan, Barry and Gibson, Huw. *CARA: A Human Reliability Assessment Tool for Air Traffic Safety Management — Technical Basis and Preliminary Architecture* (2007): pp. 197–214. DOI 10.1007/978-1-84628-806-7_13.

[13] Kirwan, Barry, Gibson, Huw, Kennedy, Richard, Edmunds, Jim, Cooksley, Garry and Umbers, Ian. "Nuclear action reliability assessment (NARA): A data-based HRA tool." *Safety and Reliability* Vol. 25 (2005): pp. 38–45. DOI 10.1080/09617353.2005.11690803.

[14] Gibson, Huw, A.M, Mills, S, Smith and Kirwan, Barry. "Railway Action Reliability Assessment, a railway- specific approach to human error quantification." 2014.

[15] Akyuz, Emre, Celik, Metin and Cebi, Selcuk. "A phase of comprehensive research to determine marine-specific EPC values in human error assessment and reduction technique." *Safety Science* Vol. 87 (2016): pp. 63–75. DOI https://doi.org/10.1016/j.ssci.2016.03.013. URL https://www.sciencedirect.com/science/article/pii/S0925753516300194.

[16] Hollnagel, Erik. *Cognitive reliability and error analysis method (CREAM)*. Elsevier (1998).

[17] Guo, Yundong, Sun, Youchao, Yang, Xiufang and Wang, Zongpeng. "Flight safety assessment based on a modified human reliability quantification method." *International Journal of Aerospace Engineering* Vol. 2019 (2019): pp. 1–12.

[18] Yoshimura, Kenji, Takemoto, Takahiro and Mitomo, Nobuo. "The support for using the cognitive reliability and error analysis method (CREAM) for marine accident investigation." *2015 International Conference on Informatics, Electronics & Vision (ICIEV)*: pp. 1–4. 2015. DOI 10.1109/ICIEV.2015.7334041.

[19] Chen, Jiayu, Zhou, Dong, Lyu, Chuan and Zhu, Xinv. "A method of human reliability analysis and quantification for space missions based on a Bayesian network and the cognitive reliability and error analysis method." *Quality and Reliability Engineering International* Vol. 34 No. 5 (2018): pp. 912–927. DOI https://doi.org/10.1002/qre.2300. URL https://onlinelibrary.wiley.com/doi/pdf/10.1002/qre.2300, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qre.2300.

[20] Walsh, Hannah and Andrade, Sequoia. "MIKA." URL https://github.com/nasa/mika.

[21] Andrade, Sequoia and Walsh, Hannah. "MIKA: Manager for Intelligent Knowledge Access Toolkit for Engineering Knowledge Discovery and Information Retrieval." *INCOSE International Symposium*, Vol. 33. 1: pp. 1659–1673. 2023. Wiley Online Library.

[22] "National Transportation Safety Board Aviation Investigation Search." https://www.ntsb.gov/Pages/AviationQueryV2.aspx. Accessed: 2024-03-01.

[23] "European Union Agency for Railways Accident Investigation." https://www.era.europa.eu/era-folder/accident-investigation. Accessed: 2024-03-01.

[24] Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." *arXiv preprint arXiv:2203.05794* (2022).

[25] Reimers, Nils and Gurevych, Iryna. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 2019. Association for Computational Linguistics. URL http://arxiv.org/abs/1908.10084.

[26] Lutz, Marc-Alexander, Schäfermeier, Bastian, Sexton, Rachael, Sharp, Michael, Dima, Alden, Faulstich, Stefan and Aluri, Jagan Mohini. "KPI Extraction from Maintenance Work Orders—A Comparison of Expert Labeling, Text Classification and AI-Assisted Tagging for Computing Failure Rates of Wind Turbines." *Energies* Vol. 16 No. 24 (2023): p. 7937.

[27] Cheong, Hyunmin, Li, Wei, Cheung, Adrian, Nogueira, Andy and Iorio, Francesco. "Automated extraction of function knowledge from text." *Journal of Mechanical Design* Vol. 139 No. 11 (2017): p. 111407.

[28] Ahmed, Faez, Fuge, Mark and Gorbunov, Lev D. "Discovering diverse, high quality design ideas from a large corpus." *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 50190: p. V007T06A008. 2016. American Society of Mechanical Engineers.

[29] Paradis, Carlos, Kazman, Rick, Davies, Misty and Hooey, Becky. "Augmenting topic finding in the NASA aviation safety reporting system using topic modeling." *AIAA scitech 2021 forum*: p. 1981. 2021.

[30] Paradis, Carlos, Kazman, Rick, Davies, Misty D and Hooey, Becky L. "Identifying Emerging Safety Threats through Topic Modeling in the Aviation Safety Reporting System: A COVID-19 Study." *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*: pp. 1–8. 2021. IEEE.

[31] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton and Toutanova, Kristina. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[32] Achiam, Josh, Adler, Steven, Agarwal, Sandhini, Ahmad, Lama, Akkaya, Ilge, Aleman, Florencia Leoni, Almeida, Diogo, Altenschmidt, Janko, Altman, Sam, Anadkat, Shyamal et al. "Gpt-4 technical report." *arXiv preprint arXiv:2303.08774* (2023).

[33] Fuchs, Jared, Helmerich, Christopher and Holland, Steven. *Transforming System Modeling with Declarative Methods and Generative AI*: DOI 10.2514/6.2024-1054. URL https://arc.aiaa.org/doi/pdf/10.2514/6.2024-1054, URL https://arc.aiaa.org/doi/abs/10.2514/6.2024-1054.

[34] Andrade, Sequoia R and Walsh, Hannah S. "Discovering a failure taxonomy for early design of complex engineered systems using natural language processing." *Journal of Computing and Information Science in Engineering* Vol. 23 No. 3 (2023): p. 031001.

[35] Walsh, Hannah S and Andrade, Sequoia R. "Semantic Search With Sentence-BERT for Design Information Retrieval." *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 86212: p. V002T02A066. 2022. American Society of Mechanical Engineers.

[36] Mbaye, Seydou, Walsh, Hannah S, Jones, Garfield and Davies, Misty. "BERT-based Topic Modeling and Information Retrieval to Support Fishbone Diagramming for Safe Integration of Unmanned Aircraft Systems in Wildfire Response." *2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC)*: pp. 1–7. 2023. IEEE.

[37] Mbaye, Seydou, Walsh, Hannah S, Davies, Misty, Infeld, Samantha I and Jones, Garfield. "From BERTopic to SysML: Informing Model-Based Failure Analysis with Natural Language Processing for Complex Aerospace Systems." *AIAA SCITECH 2024 Forum*: p. 2700. 2024.

[38] Feldman, Jolene, Barshi, Immanuel, Smith, Brian and Matthews, Bryan. "Reports of Resilient Performance: Investigating Operators' Descriptions of Safety-Producing Behaviors in the Aviation Safety Reporting System." *42nd International Symposium on Aviation Psychology*: p. 122. 2021.

[39] Barshi, Immanuel, Matthews, Bryan and Feldman, Jolene. "Extracting Lessons of Resilience Using Machine Mining of the ASRS Database." *22nd International Symposium on Aviation Psychology (ISAP)*. 2023.

[40] Sawyer, Michael, Berry, Katherine, Kinsella, Amelia, Hinson, R Jordan and Bynum, Edward. "Using Natural Language Processing to Identify Mental Health Indicators in Aviation Voluntary Safety Reports." (2024).

[41] Cankaya, Mehmet Burak. "Understanding Aviation Mental Health with Explainable Artificial Intelligence in Incident Reports." (2024).

[42] Darveau, Katherine E. "Automated Classification of Human Factors Aviation Operational and Safety Events: A Human-Machine Teaming Approach to Text Mining and Machine Learning." Ph.D. Thesis, Tufts University. 2021.

[43] Tikayat Ray, Archana, Bhat, Anirudh Prabhakara, White, Ryan, Nguyen, Van Minh, Pinon Fischer, Olivia and Mavris, Dimitri. "Examining the Potential of Generative Language Models for Aviation Safety Analysis: Case Study and Insights using the Aviation Safety Reporting System (ASRS)." (2023). DOI 10.20944/preprints202307.0192.v2. URL https://doi.org/10.20944/preprints202307.0192.v2.

[44] Cooper, Susan E, Ramey-Smith, AM, Wreathall, J and Parry, GW. "A technique for human error analysis (ATHEANA)." Technical report no. Nuclear Regulatory Commission. 1996.

[45] Levine, Camille S, Al-Douri, Ahmad, Paglioni, Vincent Philip, Bensi, Michelle and Groth, Katrina M. "Identifying human failure events for human reliability analysis: A review of gaps and research opportunities." *Reliability Engineering & System Safety* (2024): p. 109967.

[46] Blei, David M, Ng, Andrew Y and Jordan, Michael I. "Latent dirichlet allocation." *Journal of machine Learning research* Vol. 3 No. Jan (2003): pp. 993–1022.

[47] Teh, Yee, Jordan, Michael, Beal, Matthew and Blei, David. "Sharing clusters among related groups: Hierarchical Dirichlet processes." *Advances in neural information processing systems* Vol. 17 (2004).

[48] Mei, Qiaozhu, Shen, Xuehua and Zhai, ChengXiang. "Automatic labeling of multinomial topic models." *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*: pp. 490–499. 2007.