Advances in Technology Innovation, vol. 10, no. 4, 2025, pp. 340-357

Risk Management Framework-Based Failure Mode and Effect Analysis for AI Risk Assessment

Yunarso Anang[1,*], Lya Hulliyyatus Suadaa[1], Lutfi Rahmatuti Maghfiroh[1,2], Nori Wilantika[1], Masakazu Takahashi[2], Yoshimichi Watanabe[2]

[1]Department of Statistical Computing, Politeknik Statistika STIS, Jakarta, Indonesia

[2]Department of Computer Science and Engineering, University of Yamanashi, Kofu, Japan

Received 11 December 2024; received in revised form 11 June 2025; accepted 12 June 2025

DOI: https://doi.org/10.46604/aiti.2025.14609

Abstract

As artificial intelligence (AI) technologies continue to spread into human life, developers must ensure benefits while minimizing the risk of adverse impacts. This study aims to evaluate risks in real-world AI applications using the AI Incident Database. It employs Failure Mode and Effect Analysis and the National Institute of Standards and Technology AI Risk Management Framework to identify failures, their causes and effects, and assess how current systems address them. A total of 100 incident reports were analyzed. The findings indicate frequent failures in autonomous systems and biased predictions. Seven cases were classified in the highest risk categories, including those involving physical harm and loss of life. Over 80% failures originated from algorithmic flaws or poor data quality. The method employed successfully evaluates the risks in current AI applications, revealing critical gaps in risk management and emphasizing the urgent need for targeted safeguards and proactive mitigation strategies.

Keywords: AI risk assessment, failure mode and effect analysis, FMEA, AI incident database, NIST's AI RMF

1. Introduction

Artificial intelligence (AI) systems have become integral to modern society, impacting everything from healthcare to autonomous vehicles. However, as AI applications accelerate, so do the associated risks. Ensuring the safety and reliability of AI systems is essential to prevent unintended or even intentional consequences and protect users. AI systems should have been developed to benefit people, society, and the ecosystem while minimizing the risk of adverse effects. However, despite the undeniable AI's remarkable advancements and continuing rapid expansion into new application domains, as with most emerging technologies, there are many cases where AI systems have gone wrong, and the number of incident reports is still growing [1]. A proper method or guideline is required to minimize the risk of adverse effects.

Risk management, especially in manufacturing industries and critical missions such as the Apollo space program, has become a formal science since the 1950s. Its guideline has been published in ISO/IEC 31000:2018 as Risk Management—Principles and Guidelines [2]. As for AI systems, the United States National Institute of Standards and Technology (NIST) released the Artificial Intelligence Risk Management Framework (AI RMF 1.0) [3]. Finally, one month later, in the same year, a guideline for risk management, particularly for AI, was published in ISO/IEC 23894:2023 Artificial Intelligence—Guidance on Risk Management [4]. Both guidelines guide how organizations that utilize AI, involving development, production, deployment, or use of products, systems, and services, can manage risk specifically related to AI. It also describes processes of implementing and integrating AI risk management effectively.

* Corresponding author. E-mail address: anang@stis.ac.id

However, particularly in the risk assessment process, as stated by Xia et al. in their study, current frameworks, including the one provided by NIST and the ISO/IEC, lack clear guidance on how to adapt them to diverse contexts [5]. They also fail to provide a concrete and structured approach to presenting the potential failures, risks, and mitigation strategies, including their measurement. This lack of clarity can make it difficult for organizations to address identified risks effectively.

Especially in manufacturing industries, Failure Modes and Effects Analysis (FMEA) has long been used as a familiar process analysis tool in a step-by-step approach to identify all possible failures in a design, a manufacturing or assembly process, or a product or service [6]. FMEA determines the risk priority level of each failure mode based on how severe the effect is, how frequently it occurs, and how easily it could be detected. The purpose of FMEA is to provide information for designers, developers, or decision-makers to take action to eliminate or reduce failures. FMEA has also been used in software engineering and development for various applications such as manufacturing, safety, and automation [7-8].

Back to the AI systems, an initiative by an industrial/non-profit cooperative to collect reports on real-world harms related to AI systems has been initiated [9]. The work aims to enable companies to implement (design, develop, or deploy) AI systems to avoid or mitigate incidents occurring from the AI system. The AI Incidents Database (AIID) supports various research and development use cases with faceted and full-text searches on over 1,000 incident reports. Several researchers, such as Wei and Zhou [10], have used the reports to analyze the incidents occurring in AI systems.

This research aims to assess the risk in the implementation of AI systems based on the incidents that occurred in the past and provide countermeasures to address the failures. FMEA assessed failure modes, potential causes, and risk priority levels based on the level of consequence, likelihood, and detectability, followed by recommended actions. The AIID was used as the data source for FMEA. Finally, the NIST's AI RMF was used as the guideline to assess the current state of the risk management of the various AI systems reported.

This paper is organized as follows: Section 2 outlines the background and related works. Section 3 briefly describes the FMEA, the AIID, the NIST's AI RMF, and the assessment process. Section 4 describes the results and discussion, while Section 5 concludes the paper with a summary and remarks for future work.

2. Background and Related Works

In this section, the related studies are described. The beginning studies, which become the background and motivation of this research, and the latter studies, which provide hints of references for the method and guidelines used in this research, are described.

The study concerning the use of AI systems, particularly in the ethical context, was conducted by Jobin et al. [11]. The study investigates whether a global agreement on the principles of "ethical AI" exists among private companies, research institutions, and public sector organizations. The results reveal a global convergence of emerging ethical principles, including transparency, justice and fairness, non-maleficence, responsibility, and privacy. Kaun discussed two litigation cases about fully automated decision-making (ADM) in public services in Sweden [12]. The research shows how different stakeholders were in conflict over what ADM is and does. Heinrichs examined whether AI and ADM aggravate discrimination issues [13]. The author argues that the use of AI/ADM can increase the issue of discrimination due to the opacity of AI/ADM, which threatens the moral deliberation of understanding about discrimination. The author stated that algorithms can detect hidden forms of discrimination.

Brecker et al., in their research, show that it remains challenging to assess AI systems to mitigate risks arising from biased, unreliable, or regulatory non-compliance [14]. The research highlights seven areas of concern in AI assessment, from ethical compliance to regulatory gaps and socio-technical limitations. Chanda et al., in their study, also investigate several cases of

failure of AI systems employing machine learning and deep learning [15]. The research focuses on the origins of the failure related to omission and commission errors in the inputs, processing logic, and outputs. Wen et al., in their research, state that despite the prevalence of AI ethical problems, especially in facial recognition technology, most companies are constructively unprepared to respond adequately to the public [16]. Four big technology companies' responses are deflection, improvement, validation, or preemption.

The rapid development of AI has led to growing concerns about its capability to behave responsibly in making decisions, as evidenced by the increasing number of failure reports and studies of AI systems failing. This motivates Xia et al. to investigate existing guidelines or frameworks of risk assessment, particularly for AI systems provided or used in the industry, governments, and non-government organizations [5]. In their research, Xia et al. analyzed over 16 frameworks and assessed their effectiveness and limitations in the field of AI risk assessment. This study's findings can assist relevant stakeholders in selecting an appropriate framework for their AI risk assessment. However, it should also offer clear guidance on extending or adapting the framework to suit various contexts.

The U.S. military has used FMEA as a tool for risk assessment in several industrial domains, including software applications, since the beginning of the 1940s. Haapanen and Helminen conducted a literature study to clarify the practical use of software failure mode and effects analysis in a safety-critical software-based automation application in a nuclear power plant [7]. Takahashi et al. applied FMEA to 20 existing drug-manufacturing computerized systems (DMCS) to derive standard failure modes [8]. Using the list of standard failure modes, a method for risk management has been developed for several categories of DMCS, including in-house, off-the-shelf, or a combination of them. After conducting an experiment involving a less-experienced engineer, the result shows that risk management has been achieved similarly to that of an experienced engineer.

Finally, regarding the application of FMEA in AI systems, Li et al. applied FMEA to an AI system to assess the risk to fairness by adding four additional columns: user groups, unfairness, fairness risk, and fairness mitigation [17]. Three approaches have been introduced to maximize the fairness of an AI system: normative, procedural, and algorithmic. The paper presents how it could explicitly identify user groups, unfairness, risk, and mitigation considerations related to fair AI, which may differ from risks related to safety in AI. As a risk assessment method, compared to alternatives, such as Hazard and Operability Study and Fault Tree Analysis, FMEA is structured, numerical, and also supports not only post-failure investigations but also anticipation by prioritizing risks and mitigation strategies [18]. Compared to the other methods, which are rigid and process-focused, FMEA is more flexible and better suited for more general risk analysis using a simple structured table. This nature makes it suitable for this research, whose aim is to assess the risk of AI systems based on the incident reports and to suggest countermeasures to address the failures.

FMEA can be applied to both new and existing systems. FMEA is classified as a semi-quantitative method combining both quantitative and qualitative methods. While FMEA is typically done through brainstorming processes, the availability of supporting data on failure and incident reports is inevitable. McGregor's initiative, AIID, for AI systems serves as an educational tool not only to raise awareness about the harms of AI but also to foster a deeper understanding of its potential benefits [19]. Still, it is also a tool that can be utilized to understand and anticipate these risks. It is already collecting more than a thousand reports of AI systems causing safety, fairness, or other real-world problems. Wei and Zhou conducted a content analysis on AIID to examine how AI ethics issues occur in the real world [10]. From the reported database, 13 application areas have been identified as practicing unethical AI, with language/vision models, intelligent service robots, and autonomous driving taking the lead. Issues in ethics appear in 8 different forms, including physical safety, racial discrimination, and unfair algorithms.

3. Methods and Dataset

This section describes the methods and the dataset used in this research. First, the application of FMEA as the primary risk assessment technique is described, detailing how it was adapted for use in the context of AI systems. Second, the dataset used for the analysis is introduced, which consists of incident reports collected from a publicly available AI incident database. This dataset provides real-world examples of AI system failures, which serve as the basis for identifying potential failure modes and assessing associated risks. Lastly, the overall workflow of the risk assessment process is presented, including data processing, failure mode identification, risk prioritization, and the formulation of mitigation strategies. This structured approach ensures a comprehensive and systematic evaluation of AI-related risks.

3.1. Failure Mode and Effect Analysis (FMEA)

FMEA is a method for risk assessment. FMEA follows a step-by-step approach. First, it identified all possible design, process, or final product failures. "Failure modes" refer to how a function or something might fail. "Failures" are any errors or defects, especially those affecting the end user, that can be actual or potential. Besides identifying the possible failures, another approach in FMEA is analyzing the effect, which means studying the consequences or effects of those failures. The FMEA's basic steps include definition, preparation, execution, and documentation. According to IEC 60812:2006 Analysis Techniques for System Reliability – Procedure for FMEA, a typical FMEA may include the following items: (1) Function/item; (2) Failure mode; (3) Failure causes; (4) Failure effects; (5) Detection mode; (6) Compensation provisions; (7) Severity; (8) Probability of occurrence; and (9) Comments or recommendations. These items are collected, identified, and determined typically using a table. To fit the purpose of this research, FMEA was modified. Besides the standard items, some items have been added to associate the failure mode and effects with the principle of risks in AI RMF.

To tailor the traditional FMEA for AI systems, several modifications have been introduced to align with the AI RMF. These adaptations were designed to capture unique characteristics and lifecycle stages relevant to AI risk contexts. There are two additional columns:

(1) Lifecycle Phases: Maps the failure mode to the relevant AI lifecycle phase (e.g., Plan and Design, Collect and Process Data, Operate and Monitor).
(2) Trustworthy Characteristics: Identifies which of the AI RMF's seven characteristics (e.g., reliability, fairness, transparency) is impacted by the failure mode.

More details of the items and the workflow are described in Subsection 3.4.

3.2. AI Incident Database (AIID)

AIID is a database that collects harms or near-harms realized in the real world by deploying AI systems. It aims to prevent or mitigate bad outcomes by learning from experience. As outlined in their Editor's Guide, the AI incident is defined as an alleged harm or near-harm event to individuals, property, or the environment, where an AI system is suspected to be involved. AI is a system of machinery that can perform human intelligence functions, such as reasoning, pattern recognition, and understanding natural language. Machine learning is a subset of AI.

AIID is a crowd-sourced, collaborative collection of reports database. The essential items include information about the report title, the author or submitter, incident date, image URL, incident ID (automatically assigned for a new incident), and text that contains the remaining details about the incident reports. The pre-processed and searchable data is available online, while the raw data is also downloadable for further analysis. Besides regular content analysis, AIID provides a taxonomy-based analysis based on original incident reports, which can be tailored to specific usages.

A taxonomy-based analysis or a taxonomic system of the AI incident reports is a comprehensive classification workflow paired with the ontology structure of the incident, including the various factors and technical causes of the implicated systems [20]. Currently, there are two taxonomy-based classifications: Center for Security and Emerging Technology (CSET) and Goals, Methods, and Failures. The CSET refers to the AI Harm Taxonomy characterization of AI incidents, classifying harms relevant to the public policy community [21]. First, to understand the characteristics of AI harm or potential harm, the harm is grouped into tangible and intangible. Then, the additional categories of harm are defined, such as physical or psychological harm, financial loss, property damage, detrimental content, bias, differential treatment, and violation of privacy.

Currently, there are two versions of the CSET: version 0 and version 1. This research mainly used classified data from version 0 since version 1 is still relatively new, and the amount of data is less than that of version 0. However, since some of the newer incident reports only exist in version 1 classification, the classified data from this version is also incorporated.

3.3. AI Risk Management Framework (AI RMF)

2023 is a milestone in risk management for AI, and the AI RMF was published [3]. The initiative to develop the framework started in July 2021 and has been improved by approximately 400 comments from 240 global organizations. Since every organization is different, the challenge is treating the risks most relevant to the organization, considering its specific context.

AI RMF outlines seven characteristics of a trustworthy AI system. "Valid and Reliable" is a necessary condition of trustworthiness as the base for the other five characteristics on top of it, from "Safety to Fairness", while "Accountable and Transparent" is one characteristic that relates to all other characteristics. Approaches that enhance AI trustworthiness can reduce adverse AI risks.

Risk is the composite measure of the degree of impacts or consequences and the likelihood of an event occurring. The impact or consequence can be positive, negative, or both, resulting in opportunities or threats. When considering the negative impact, the risk is the magnitude of the harm. An example of how the negative impact is assessed is based on who the adverse risks contribute to. The potential harm of AI systems can be categorized into three categories: Harm to people, an organization, and an ecosystem.

AI risks or failures that need to be better defined are difficult to measure quantitatively or qualitatively. Some challenges in risk measurement include those related to third-party software, hardware, and data; risks at different stages of the AI lifecycle; risks in real-world settings; obscurity, which can be a result of the nature of AI systems lacking transparency or documentation during development or deployment, or inherent uncertainties in AI systems and human baseline. On top of that, it can be challenging to systematize the baseline metrics for comparison for AI systems that are intended to replace or augment human activity because AI systems can perform various tasks and exhibit different behaviors compared to humans.

The OECD has established a framework for categorizing AI lifecycle activities based on five crucial socio-technical dimensions. People and Planet, Application Context (where in the original paper written as Economic Context), Data and Input, AI Model, and Task and Output, each with properties relevant for AI policy and governance, including risk management [22]. Based on these dimensions, AI RMF outlines the lifecycle of the AI systems, including the audience or actors related to each phase of the lifecycle, which includes planning and design, collecting and processing data, building and using models, verifying and validating, deploying and using, operating and monitoring, and using.

While AI RMF does not provide a concrete or structured way to present the potential failure, risk, and mitigation, including their measurement, in this research, the AI trustworthy characteristics and the AI lifecycle stages defined in AI RMF are used as the guideline to assess the current state of AI systems risk management reported in the AIID, along with the FMEA analysis.

3.4. Workflow of the Risk Assessment

This section describes the overall workflow of this research, as shown in Fig. 1. The process begins with Step 1, which involves collecting AI incident reports from a publicly available database to serve as the foundation for risk analysis. In Step 2, the collected incidents are analyzed to identify failure modes, their causes, effects, and detection mechanisms. Step 3 focuses on determining the severity, likelihood, and detectability levels of each failure mode to assess risk priority. In Step 4, the analysis is extended by mapping the incidents to relevant trustworthy AI characteristics, lifecycle phases, and involved actors. Finally, Step 5 presents a discussion of the findings and offers recommendations for improving AI system reliability and risk management practices.

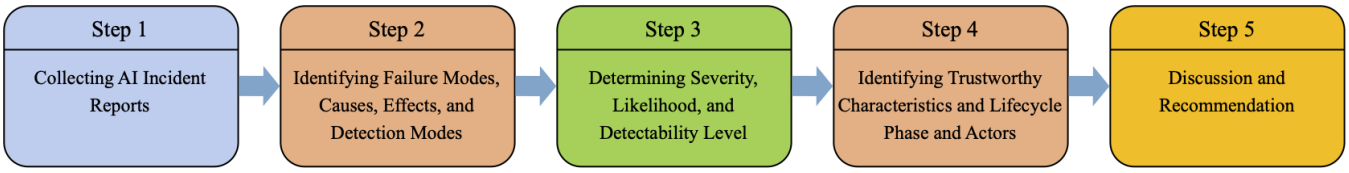


Fig. 1 Workflow of the Risk Assessment

Step 1: Collecting AI Incident Reports

The AI incident reports used in this research are the database snapshot downloaded from the AIID website. The database snapshot contains all "tables" from the AIID that use MongoDB. This research used the pre-classified data using AI Harm Taxonomy characterization stored in the "classifications" table. The table was converted into a worksheet from its original "bson" format. The table was then filtered to include only the CSETv0 and CSETv1 pre-classified reports. The following data have been used in the risk assessment: full description, named entities, incident ID, severity, harm type, near miss, loss of life, and AI applications.

Step 2: Identifying Failure Modes, Causes, Effects, and Detection Modes

This step, along with the following one, constitutes the core of the FMEA process applied in this research. Before conducting the analysis, a standardized template for the FMEA table was developed to ensure consistency and clarity in documenting failure modes, causes, effects, and detection methods. The structure of this template is presented in Table 1, which serves as the foundation for systematically assessing and prioritizing risks.

The table is filled with the incident reports, with each row representing a distinct incident. Each column corresponds to a specific attribute or detail of the incident, such as identification, functions, failure mode, causes, detection of failure, prevention, effects of failure, and risk priority, as outlined in Table 2.

Table 1 Template of the FMEA Table

Description of Unit			Description of Failure				Effects of Failure			Risk Priority			
FMEA ID	AIID incident ID	Function	Failure Mode	Failure Causes	Detection of Failure	Prevention	People	Organization	Ecosystem	Severity	Likelihood	Detectability	Risk Priority Number

"Failure modes" in FMEA means the ways or modes in which something might fail, and failures or any errors or defects, especially those that affect the customer and can be potential or actual. Therefore, for convenience, in this research, failure modes were taken from the description of the reported incident in the AIID, which consists of the failures as well as the ways or modes in which the failures occurred. This approach allows for a practical mapping of real-world AI system failures into the structured FMEA framework. By doing so, the study ensures that both the nature and context of the failures are captured, enabling a more comprehensive risk assessment.

Table 2 Group of Columns in FMEA Table

Group of Columns	Description
Description of Unit	Contains the identity of each incident report. The FMEA ID is a unique number starting from 1. The AIID incident ID is taken from the report. The Functions are filled with "AI Applications" from the report.
Description of Failure	Contains the failure mode, failure causes, detection mode, and prevention. The Failure Mode is taken from the Full Description of the report. The Failure Causes, Detection of Failure, and Prevention will be identified in Step 3.
Effects of Failure	Contains the effects or the consequences of the failure. This research divides the effects into three categories: People, Organization, or Ecosystem, indicating to whom the adverse risks contribute. These items will also be identified in Step 3.
Risk Priority	Contains the degree or level of risk obtained from multiplying the Severity, Likelihood, and Detectability levels.

In addition to these columns, two more columns will be introduced later in Step 4 to enrich the dataset: Trustworthy Characteristics and Lifecycle Phase and Actors. These columns are designed to establish a clear connection between each recorded failure and the relevant components of the AI RMF. By incorporating these dimensions, the analysis can more effectively align observed incidents with trustworthiness principles and lifecycle stages, enhancing the interpretability and relevance of the findings.

Step 3: Determining Severity, Likelihood, and Detectability Level

In this step, the level of severity, likelihood, or probability of occurrence, and detectability are determined based on several data points from the incident report. Severity is the scale of impact or seriousness of the failure, ranging from 1 for no harm to 10 for hazardous without warning. Table 3 lists the scales used to measure Severity. The value is determined from the item "Failure" from the AI incident report, with additional considerations from other items: "Harm type" and "Near miss". Failure of an AI incident with no or negligible effect is categorized as a low severity scale. On the other hand, failures that cause damage to humans, financial, property, social, political, and even cause loss of life are categorized on a high severity scale.

Table 3 FMEA scale for Severity

Level	Class	Meaning
10	Hazardous without warning	Safe system operational failure with high severity compromising safety without warning
9	Hazardous with warning	Safe system operational failure with high severity compromising safety with warning
8	Very High	System inoperable with destructive failure without compromising safety
7	High	System inoperable with equipment damage
6	Moderate	System inoperable with minor damage
5	Low	System inoperable without damage
4	Very Low	System operable with significant degradation of performance
3	Minor	System operable with some degradation of performance
2	Very Minor	System operable with minimal interference
1	None	No effect

Next, Likelihood refers to the estimated probability that a failure associated with an AI incident will occur. In this research, likelihood is quantified on a scale from 1 to 10, where a score of 1 indicates a remote or unlikely event, and a score of 10 represents a very high or almost inevitable event. The specific criteria and definitions for each level of this scale are detailed in Table 4 lists the scales used to measure Likelihood.

Table 4 FMEA scale for Likelihood or Probability of Occurrence

Level	Class	Meaning
9-10	Very high	Failure is almost inevitable
7-8	High	Failure is repetadly occurred
4-6	Moderate	Failure is occassionally occurred
2-3	Low	Failure is relatively few
1	Remote	Failure is unlikely

Then, Detectability refers to the ability to detect the failure of AI incidents. In this research, detectability ranged from 1 for "almost certain," or the failure almost certainly can be detected by current control before being exposed to users, to 10 for "almost impossible," or no known controls available to detect the failure before exposure of the system or device that uses AI

to users. Table 5 lists the scales used to measure Detectability.

Table 5 FMEA scale for Detectability

Level	Class	Meaning
10	Almost Impossible	No known controls available to detect the failure mode
9	Very Remote	Very remote likelihood current controls will detect failure mode
8	Remote	Remote likelihood current controls will detect failure mode
7	Very Low	Very low likelihood current controls will detect failure mode
6	Low	Low likelihood, current controls will detect failure mode
5	Moderate	Moderate likelihood current controls will detect failure mode
4	Moderately High	Moderately high likelihood current controls will detect failure mode
3	High	High likelihood current controls will detect failure mode
2	Very High	Very high likelihood current controls will detect failure mode
1	Almost Certain	Current controls almost certain to detect the failure mode. Reliable detection controls with similar processes.

Lastly, the Risk Priority Number (RPN) is a key metric used to evaluate the overall risk associated with a specific failure mode. It is calculated by multiplying three factors: Severity, Likelihood, and Detectability. A higher RPN value indicates a greater level of risk, signaling the need for more immediate or significant corrective actions to mitigate potential failures and enhance system reliability.

Step 4: Identifying Trustworthy Characteristics, Lifecycle Phase, and Actors

To assess the current state of the risk management of each AI application or AI system reported in AIID, in this step, the characteristics of trustworthiness and the phase and actors of the AI lifecycle related to each function in each incident are identified.

Step 5: Discussion and Recommendation

Finally, based on the FMEA table and the results of the descriptive analysis, the discussion and the recommendation are developed. This process involves interpreting the identified failure modes, their associated risk levels, and patterns observed across the dataset. The resulting discussion highlights key findings, while the recommendations aim to address critical risk areas and propose strategies for improving system reliability.

4. Results and Discussion

In the first step, the incident reports were collected from AIID. This research used the AIID snapshot of the "January 15, 2024" version, downloaded from their website. At the time, there were 92 incident records found with CSET v0 classifications and only 62 records with CSET v1 classifications. There were 53 records with CSET v1 classifications, which were also found in the CSET v0 classified records, while only 8 records with CSET v1 classifications were not found in the CSET v0 classified records. Accordingly, all 92 records with CSET v0 specifications plus 8 records with CSET v1 classifications have been chosen as the data. One hundred incident reports of CSETv0 and CSETv1 annotated records have been extracted from the "classification" table. Of the 100 incidents, five records with Incident IDs 4, 21, 29, 30, and 42 have been excluded due to the absence of AI-related harm or vague information. Values of Incident ID, AI Applications, and Full Description extracted from the "classification" table were put in the FMEA table, as shown in Table 1 in the FMEA incident ID, Function, and Failure Mode columns, respectively.

The next step was the identification of failure modes, causes, effects, and detection modes for each incident. Two independent groups—each consisting of two researchers with academic and professional backgrounds in computation, statistics, and data analysis—analyzed the classified AIID reports separately. All four analysts had prior experience working with qualitative coding and risk assessment frameworks. Before the formal annotation process began, both groups were introduced to the FMEA framework through a structured internal briefing that included examples, detailed definitions of failure

modes and scoring criteria, and a walkthrough using sample AIID reports. During the annotation phase, two groups independently analyzed the data from the classified AIID to identify the causes, effects, and detection modes based on the available data from the AIID. The "Effects of Failure" were also determined based on the description of the incident, which was divided into three audiences: people, organization, and ecosystem.

Next, in the third step, each group assigned scores for severity, likelihood, and detectability (on a scale of 1 to 10), based on the information extracted from the reports. These values were then used to calculate the RPN for each failure mode. After both groups completed their independent annotations, a reconciliation process was conducted. In this stage, the two groups reviewed and compared their respective annotations and scoring results. Any discrepancies were discussed, and a consensus was reached through deliberation. The final dataset used for analysis consisted of these consensus annotations. Excerpts of the filled FMEA table are shown in Table 6.

Table 6 FMEA Table from AI Incident Database (excerpts)

Description of Unit			Description of Failure				
FMEA ID	AIID incident ID	Functions	Failure Mode	Category of the Failure Mode	Failure Causes	Detection of Failure	Prevention
1	1	["content filtering", "decision support", "curation", "recommendation engine"]	The content filtering system for YouTube's children's entertainment app, which incorporated	Content filter failure	Algorithmic filters and human reviewers failed to screen out inappropriate material	Only visual inspection after it is viewed	Video labeling by content creator
2	2	["robotics"]	On December 5, 2018, a robot punctured a can of bear spray in Amazon's fulfillment center in	False object detection by robot	The robot failed to detect objects with active and hazardous ingredients.	No hazardous object detection feature	None
3	5	["Robotic Surgery"]	Reports of robotic surgeries resulting in injury and death between 2000 and 2013, as found in the Manufacturer		Due to system/hardware error (62%). The remainder is due to the inherent risk of surgery or human error.	None	None
4	6	["comprehension", "language output", "chatbot"]	Microsoft chatbot, Tay, was published on Twitter on March 23, 2016. Within 24		The automated tweet feature had been manipulated by Twitter users	Only visual inspection after it is posted	None
5	7	["AI content creation", "AI content editing"]	Wikipedia bots meant to help edit articles through artificial intelligence clash with each other,		A clash among Wikipedia bots caused the bot to undo the other's edits repeatedly. The whole situation has been described as a "bot-on-bot editing war".	The problem can be detected by investigating the change logs.	None
6	8	["traffick flow forecasting", "autonomous driving"]	Uber's autonomous vehicles have been recorded running red lights on two occasions	Autonomous vehicle crash	The operator, Uber, claimed the system's error did not cause it but blamed it for human operator error, while the driver reported that the fault was of the AI system.	If there is, the failure can only be detected by the human operator.	Visual inspection by the driver
7	9	["data processing", "data prediction"]	A value-added measurement-based algorithm used to calculate the effectiveness of		The algorithm only covers two objects	Human reports	None
9	11	["risk assessment", "crime projection"]	An algorithm developed by Northpointe and used in the penal system is shown to be	Systems produce discriminatory outcomes (race, gender, religion)	Lack of high-quality (unbiased and non-discriminative) data training	It can be detected by analyzing the results or the reports.	None
10	12	["Natural language processing"]	The most common techniques used to embed words in natural language	Systems produce discriminatory outcomes (race, gender, religion)	Lack of high-quality (unbiased and non-discriminative) data training	It can be detected by analyzing the results or the reports.	None

Table 6 FMEA Table from AI Incident Database (excerpts) (continued)

Description of Unit		Effects of Failure			Risk Priority				NIST AI RMF Characteristics	
FMEA ID	AIID incident ID	People	Organization	Ecosystem	Severity	Likeliho-od	Detecta-bility	Risk Priority Number	Phases/Actors	Trustworthy Characteristics
1	1	Exposing children to videos that include sex, drugs, violence, profanity, and conspiracy theories			5	3	9	135	Build & Use Model, Verify and Validate	Valid and reliable
2	2	It had a secondary impact on people caused by the active and hazardous ingredients, which caused several workers to be exposed to the fumes from the spray, having trouble breathing and a burning sensation in the eyes and throat.			6	2	9	108	Use or impacted by	Safe, Valid, and Reliable
3	5	Injuries from burns from sparks emitted by the machine, robotic arms becoming dislodged in the patient, and the surgeon losing control of the machine or the machine powering			7	5	10	350	Use or impacted by	Safe, Valid, and Reliable
4	6	People who use Twitter feel unpleasant as a result of this behavior.			4	5	9	180	Use or impacted by	Explainable and Interpretable
5	7			Two notable cases between Darnkessbot and Xqbot led to 3,629 edited articles between 2009-2010 and between Tachikoma and Russbot, leading to more than 3,000 edits. The failures have occurred across articles in 13 languages on Wikipedia, with most occurring in Portuguese and German.	2	5	9	90	Use or impacted by	Explainable and Interpretable
6	8	While there were no injuries or collisions, it may lead to an incident that caused collision and damage to property, or lead to human injuries and life lost.			2	3	5	30	Operate and monitor	Safe, Valid, and Reliable
7	9	Teachers who were evaluated as not being effective will be affected.	The school will lose some of its teachers due to the wrong evaluation.		2	3	8	48	Build & Use Model, Verify and Validate	Valid and reliable, Fair - with harmful bias managed, accountable and transparent

Table 6 FMEA Table from AI Incident Database (excerpts) (continued)

Description of Unit		Effects of Failure			Risk Priority				NIST AI RMF Characteristics	
FMEA ID	AIID incident ID	People	Organization	Ecosystem	Severity	Likeliho-od	Detecta-bility	Risk Priority Number	Phases/Actors	Trustworthy Characteristics
9	11	According to the report, 40% of the predictions were incorrect. Since then, the system has been used in Broward County, Florida, to help judges make decisions surrounding pre-trial release and sentencing post-trial; the sentence may be inaccurately judged. Also, the system is likely to produce racially skewed results, according to a review by ProPublica.	It may influence the trust in the courts and legal system.	It may influence trust in automated systems in highly impacted systems like courts and legal systems.	3	5	8	120	Collect and Process Data, Build and Use Model, Verify and Validate	Valid and reliable, Fair - with harmful bias managed, accountable and transparent
10	12	The effect may affect people depending on the application.	The effect may affect the organization depending on the application.	The effect may broadly destroy the ecosystem if not correctly addressed.	3	5	8	120	Collect and Process Data, Build and Use Model, Verify and Validate	Valid and reliable, Fair - with harmful bias managed, accountable and transparent

The most often found failures of all processed incident reports are related to systems producing prediction results biased by race, gender, or religion. In several incidents, biased prediction results occurred due to flaws in the algorithm used or a lack of diversity in the dataset used for training. For example, when users search for a particular name of a public figure, Google suggests a specific religion or ethnicity in its autocomplete. However, in some incidents, the discriminatory prediction results occurred on purpose. A health system was found to include a race multiplier in the algorithm for estimating kidney function. Researchers found in their study that the race multiplier resulted in an underestimated risk of African-American patients, leading to inequitable chances of being placed on a kidney transplant waiting list [23]. Besides that, other failures that are also often found are related to autonomous vehicle crashes, including car accidents, self-driving shuttle accidents, and even a plane crash, resulting in human physical injury or loss of life.

Fig. 2 illustrates a portion of each function of the incident data. There is no dominant function where an incident can contain multiple functions. However, functions related to facial recognition, decision support, recommendation engines, image recognition, and image classification appear more frequently than other functions in the reported AI incidents. From the investigation, more than half of the incidents were caused by algorithm problems, and almost 30% were caused by data problems. The result is in line with typical AI applications, as stated by Hamzah-Cherif et al. in their research, that the inappropriate number and the poor quality of the training data may produce the issue in the algorithm, thus influencing the accuracy of the result [24].

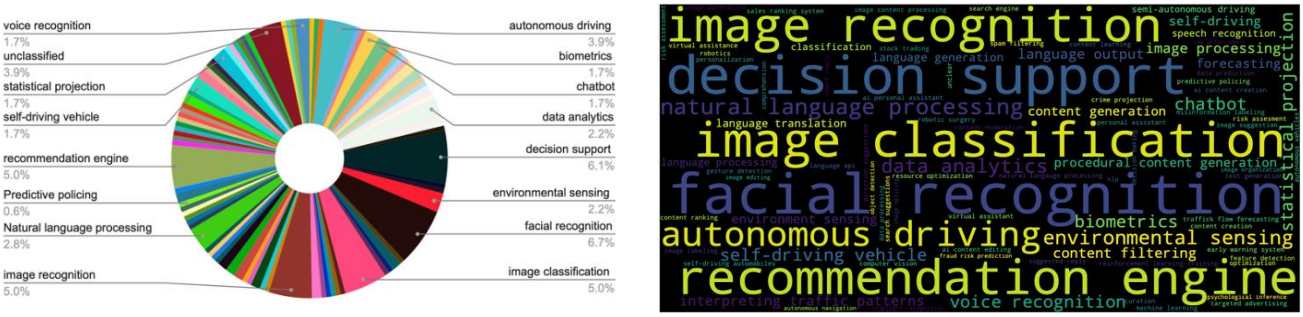


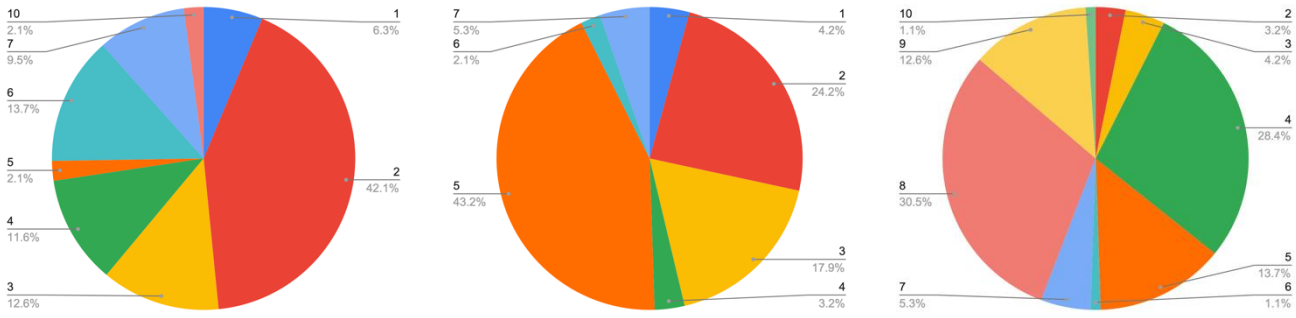
Fig. 2 Distribution of functions related to each failure mode (an incident may have more than one function)

Based on the incident reports, 35.79% of incidents could only be detected through human reporting, 35.79% through analysis of results, data, or models, and approximately 20% through both methods. One of the most notable findings from this study is that 87.37% of the incidents analyzed had no documented prevention method. This striking absence suggests that many organizations either do not engage in proactive risk mitigation or fail to institutionalize and communicate such efforts.

Several factors may contribute to this gap, including the lack of standardized risk assessment frameworks, insufficient regulatory pressure, limited organizational expertise in AI safety, the rapid pace of AI deployment, and a prevailing focus on performance over robustness. The implications are significant: without preventive strategies, organizations are more likely to respond reactively to failures, increasing the likelihood of harm, reputational damage, and regulatory scrutiny. This reactive posture contrasts sharply with established engineering practices, such as FMEA and the Risk Management Framework (RMF), which emphasize early identification and mitigation of potential failure points. As Kotonya and Sommerville [25] argue, preventive planning and early intervention are essential in complex system development to reduce downstream impacts and ensure system dependability.

Fig. 3 shows the distribution of the number of incidents for each level of severity, likelihood, and detectability. The severity level has a value between 1 and 10. The severity level means the higher the level number, the more severe the incident. While the likelihood level means the higher the level number, the greater the chance that the incident will occur. An analysis of the severity and likelihood of incidents reveals that most events are not highly severe—42.1% are at level 2 on a 10-point scale. However, the chance that these incidents happen is relatively high, with 43.2% at the likelihood level 5. Only 4.2% of incidents have a likelihood level of 1, or a failure is unlikely. This shows that even if the damage caused by each incident is small, they happen often enough to create long-term or widespread effects. In FMEA analysis, Carlson has emphasized that frequently occurring minor failures can point to systemic design weaknesses if not addressed early [26].

In addition, the detectability level represents the likelihood that process controls will detect the existence of a defect before the subsequent process or before exposure to a client. This level ranges from 1 to 10. Level 10 means it is almost impossible to detect the upcoming incident, and level 1 means current controls are nearly certain to detect the failure mode. The analysis of detectability results in many failures that are difficult to detect in time (Fig. 3 (c)). Around one-third of incidents are at detectability level 8, meaning they are hard to catch before causing harm. Only 1% of incidents are at level 10 (the hardest to detect), and none are at level 1, where a failure is almost guaranteed to be detected. This means that current monitoring systems may not be strong or advanced enough to find problems early.



(a) Severity (b) Likelihood (c) Detectability

Fig. 3 Number of Incidents by the Level of Severity, Likelihood, and Detectability

Furthermore, the relationship between the severity, likelihood, and detectability of failure was examined. It is often intertwined, but only sometimes straightforward. Generally, the severity of failure refers to the scale of failure effect, and the likelihood of failure refers to the probability or chance of failure. Detectability, on the other hand, refers to the ease or difficulty of identifying that failure once it has occurred by process controls before the subsequent process or exposure to a client. This relationship is shown in a Sankey chart in Fig. 4. In many cases, hazardous failures without warning and high-impact failures

are unlikely to occur and may be harder to detect. However, some of them are moderately detectable. Most failures with low and moderate probability occurrence are both hard to detect and easy to detect. This suggests that some failures may not receive as much attention in the failure detection controls.

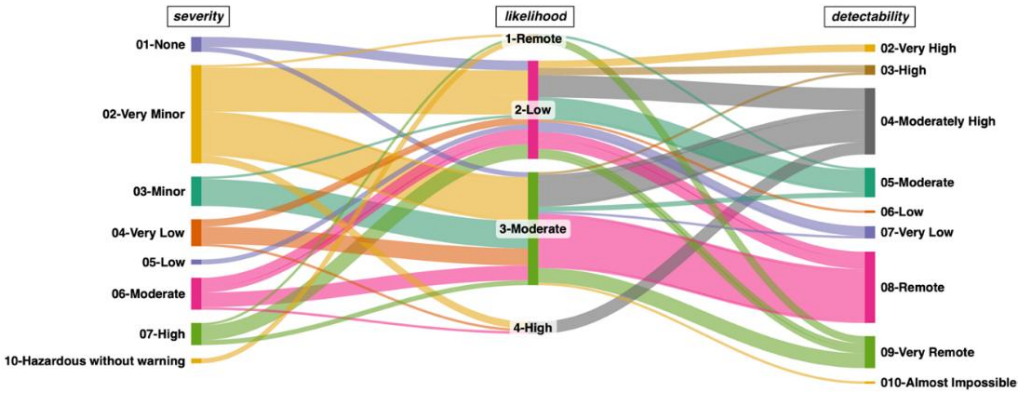
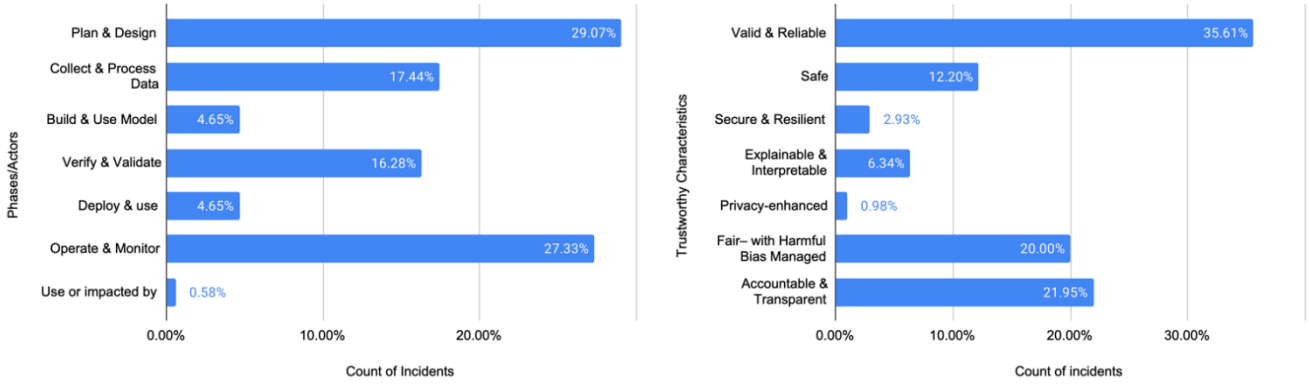


Fig. 4 Relationship between the severity, likelihood, and detectability of failure

In the fourth step, the characteristics of the current AI risk management according to the NIST AI RMF were determined as shown in Table 6. According to AI RMF, an AI system has several key dimensions surrounding its lifecycle stages. In this research, the potential impacts and risks of the incident phases have also been identified. Each incident may have one or more possible effects and risks in phases. Of all identified incidents, almost half have one identified phase, while the remaining have two or three phases each in a similar proportion.



(a) Phases/Actors (b) Characteristics of Trustworthy

Fig. 5 Characteristics of Current AI System Risk Management

From all identified phases in incidents, as shown in Fig. 5(a), "Plan and Design" and "Operate and Monitor" are the most frequently identified phases, with 50 and 47 incidents, respectively, or 29.07% and 27.33% of all incidents. This means that 29.07% of the failure modes of all the incidents examined in this study can be identified during the early stage of the AI lifecycle, the "Plan and Design" phase. In the AI lifecycle stages, activities conducted during the "Plan and Design" phase encompass auditing and assessing the impact of the AI system to be developed from legal and ethical perspectives. Yet, in 27.33% of incidents, failure could only be detected after the AI systems were operated.

Besides the "Plan and Design" and "Operate and Monitor" stages, most failure modes are detected in the "Collect and Process Data" and "Verify and Validate" phases. This is due to the functional problems of most incidents, which are related to algorithms and training data. This data tells us that risks often appear either very early or only after the AI system has already been deployed. It also means that managing risk must happen throughout the entire development cycle, not just at the beginning or end. This supports the NIST AI Risk Management Framework, which recommends continuous assessment and improvement throughout the AI lifecycle [3].

Building on the mapping of incidents to AI RMF lifecycle phases, analysis is conducted on the distribution of incidents and identified targeted risk management strategies for each phase. In the "Plan and Design" phase, ethical impact assessments and stakeholder engagement are critical to anticipate societal and legal implications. During the "Collect and Process Data" and "Verify and Validate" phases, technical interventions such as adversarial robustness testing, bias audits, and validation against edge cases can mitigate common failure modes. For the "Operate and Monitor" phase, continuous monitoring and incident response protocols are emphasized. Finally, by advocating for post-deployment audits and feedback loops, iterative improvements could be achieved. These additions aim to provide actionable guidance aligned with the AI RMF lifecycle, enhancing the utility of the framework introduced in this research for practitioners.

Besides phases/actors, the characteristics of the trustworthiness of the incident were also identified. Creating AI people can trust involves finding the right balance among its various characteristics, depending on its use. However, this balancing act often comes with trade-offs. Organizations may need help with tough decisions as they try to find this balance, considering the specific context in which decisions are made. More than 35% of incidents have validity and reliability issues. The following most significant issues were fairness, accountability, and transparency. Safety was the next frequent issue, reaching more than 12%. The detailed illustration is shown in Fig. 5(b).

While the AI RMF offers a comprehensive, principle-based structure for evaluating the trustworthiness of AI systems, such as fairness, robustness, and transparency, it does not provide detailed mechanisms for identifying and prioritizing specific failure points within AI components or processes. FMEA complements AI RMF by offering a systematic, bottom-up approach to identifying failure modes, analyzing their causes and effects, and prioritizing them based on severity, occurrence, and detectability. By combining AI RMF's strategic guidance with FMEA's operational rigor, the approach of this research enhances the effectiveness of AI risk assessment and management. This integration ensures that both abstract trustworthiness goals and concrete failure risks are addressed in a unified framework.

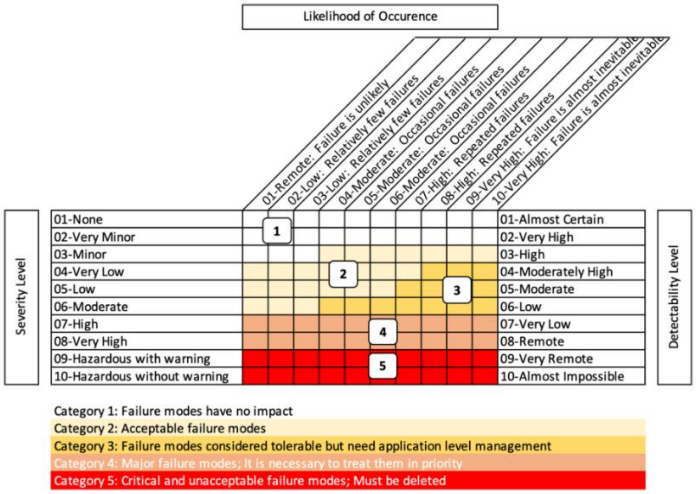


Fig. 6 Severity–Likelihood–Detectability Matrix

Furthermore, the risk priority was examined to see which incident needed more attention. Meriem and Abdelaziz created a severity–likelihood–detectability matrix incorporating relationships between those three risk-composing measurements in one matrix, to obtain the categories for prioritizing the risks [27]. In order to prioritize the risks, each risk priority is categorized into five levels, from Category 1: No impact to Category 5: Critical. The matrix, with modification, is shown in Fig. 6.

Seven incidents have the risk priority of Categories 4 and 5. Of those with risk priority of Category 5, there are three incidents: one which happened to Boeing 737 Max caused by pilots not being able to override the MCAS' actions; one which occurred to Tesla's autonomous car caused by flaws in the autopilot algorithm; and one which occurred in a factory in India caused by a malfunction of the autonomous robot. All these incidents cause loss of life. According to the suggestion of this category, the functions need to be fixed before the application can be used again.

Table 7 AI Incidents with Risk Priority of Categories 4 and 5

Category	Functions	Failure		Effect of Failure	
		Failure Mode	Failure Causes	People	Organization
4	["Robotic Surgery"]	Reports of robotic surgeries resulting in injury and death between 2000 and 2013 as found in the Manufacturer and User Facility Device Experience (MAUDE) database, a database of both voluntary and mandatory reports of mishaps. Within the 14-year span, there are 8,091 recorded malfunctions resulting in 1,391 injuries and 144 deaths.	Due to system/hardware error (62%), the remainder is due to the inherent risk of surgery or human error.	Injuries from burns from sparks emitted by the machine, robotic arms becoming dislodged in the patient, and the surgeon losing control of the machine or the machine powering down unexpectedly.	
4	["autonomous driving","self-driving vehicle"]	A robot at an office building in Washington, DC, ran itself into a water fountain. The robot, named Knightscope K5, was developed as a security robot that uses facial recognition and a variety of sensors to detect criminals. The reasons the robot fell into the fountain are unclear.	The algorithm falsely detects a criminal	May harm physical properties. May injure human's physical health/safety if falsely detected as criminal	Trust lost in security robot
4	["computer vision","autonomous driving","self-driving vehicle"]	On February 14, 2016, a Google autonomous test vehicle was involved in a low-speed collision with a bus in Google's hometown of Mountain View, CA. The self-driving car, a Lexus RX450h SUV, was attempting to navigate around an obstruction by merging toward the middle of a wide lane on El Camino Real, while a bus was approaching from the rear. The car and its test driver expected that the bus would slow down and allow the merge. However, the bus continued, apparently not expecting the self-driving car to attempt the merge, resulting in a low-speed collision.	The autonomous vehicle algorithm expected that the bus to give way	May injure human's physical health/safety	Trust lost on autopilot car
4	["stoplight recognition","semi-autonomous driving"]	A Tesla Model 3 misidentified flags with "COOP" written vertically on them as traffic lights.	Flaws in autopilot algorithm	May injure human's physical health/safety	Trust lost on autopilot car
5	[""]	A factory robot at the SKH Metals Factory in Manesar, India, pierced and killed 24-year-old worker Ramji Lal when Lal reached behind the machine to dislodge a piece of metal stuck in the machine. The robot is pre-programmed to weld together sheets of metal, and had dropped a piece of metal. When Lal reached to dislodge the piece of metal, he was pierced by a welding arm and electrocuted, dying as a result.	The robot dropped a piece of metal, and the worker had apparently moved too close to the robot while adjusting the metal	A worker was killed	Trust lost in factory robot
5	[]	A Boeing 737 crashed into the sea, killing 189 people, after faulty sensor data caused an automated maneuvering system to repeatedly push the plane's nose downward.	Pilots were supposed to be able to override MCAS' actions, but they weren't during the incident	Death	Trust Lost
5	["autonomous navigation","semi-autonomous driving","object detection","classification"]	A Tesla Model 3 on Autopilot mode crashed into a pickup on a California freeway, where data and video from the company showed that neither Autopilot nor the driver slowing the vehicle until seconds before the crash.	Flaws in autopilot algorithm	Death	Trust lost on autopilot car

The following recommendations are proposed: (1) for the Boeing 737 Max case, future designs should ensure that automated systems can always be overridden by pilots, with clear alerts and simplified interface logic. Additionally, pilot training must include thorough simulation-based scenarios for handling such system behavior; (2) for Tesla's autonomous car case, developers should improve multi-sensor fusion (e.g., combining radar, lidar, and camera data) and implement stricter

real-world validation of edge cases. Regular third-party safety audits should also be required before deploying updates, and (3) for a robot in the Indian factory case, industrial robots should include advanced fail-safe mechanisms, such as emergency shutdowns triggered by proximity sensors or human detection systems. Safety zones must be enforced using physical barriers and programmable safety logic. The remaining incidents with risk priority of Categories 4 and 5 are shown in Table 7.

Subsequently, from the AIID, there are functions derived from each AI system. By clearly defining the relations between functions, failures, and their modes, the FMEA can be used to identify the appropriate corrective actions. For example, most reported incidents involving high severity levels involve some autonomous or autopilot functions. To lower the risk of the failure's effect, the AI systems' developers can then focus on improving the functions' algorithm.

While this study is based on empirical data from the AIID, such sources may underrepresent lower-severity-oriented failure modes. Following the approach of Spreafico and Sutrisno [28], integrating synthetic failure generation, particularly for social sustainability issues like fairness, inclusivity, and transparency, can help address this gap. By simulating anticipated failures, the FMEA can be extended to capture a broader and more socially relevant risk landscape. This hybrid approach enhances both the comprehensiveness and ethical responsiveness of AI risk assessments, supporting more inclusive and sustainable system design.

Finally, while the AIID serves as a valuable repository for tracking AI-related failures, its reliance on publicly reported incidents introduces several limitations. These include a selection bias toward high-profile or media-sensationalized cases, limited representation of incidents in critical infrastructure sectors, such as healthcare, energy, and transportation, where proprietary constraints or national security concerns often prevent disclosure, and underreporting from regions with weaker regulatory or journalistic frameworks. As a result, the dataset may not fully capture the breadth or systemic nature of AI risks, particularly in domains where failures could have cascading societal impacts. Moreover, the voluntary nature of contributions can skew the dataset toward consumer-facing applications, leaving gaps in industrial or embedded AI systems. The lack of standardized definitions and reporting protocols further affects consistency and comparability across cases.

Importantly, the relationship between AI incidents and AI risk is not always linear or unidirectional. While some incidents directly reflect known risks, others may act as precursors or amplifiers of broader systemic vulnerabilities. In this sense, incidents can both result from and contribute to the evolution of AI risk. Recognizing this dynamic interplay allows for a structural distinction between "AI" as a technological paradigm and "AI systems" as context-specific implementations. Such a distinction enables a more granular analysis of incidents to understand how individual system failures may signal latent risks in the broader AI ecosystem. This structural approach supports more robust risk assessments by revealing patterns that transcend individual systems and by informing domain-sensitive mitigation strategies.

As noted by Agarwal and Nene [29], these gaps underscore the urgent need for structured, domain-sensitive reporting schemas to enhance data quality, ensure broader applicability in AI risk assessment, and support more equitable and comprehensive governance frameworks.

5. Conclusions

This study investigated various failure modes in AI systems by analyzing incidents from an AI system incident database. Using the FMEA framework, the research identified the causes and effects of failures, assessed their severity, likelihood, and detectability, and evaluated the associated risk levels. The analysis was further enriched by mapping each failure to the AI lifecycle phases, involved actors, and trustworthy characteristics as defined by the NIST AI RMF. The methodology was applied to consumer AI systems but is designed to be adaptable across domains. Key conclusions from this study are as follows:

(1) Characterization of AI Failures: AI system failures differ from traditional hardware failures. While hardware failures often stem from physical degradation, software—and by extension, AI system—failures are more abstract and functional in

nature. A failure mode in an AI system is defined as a way in which a component fails to meet its intended function or requirements.

(2) Applicability of FMEA to AI Systems: The FMEA framework is effective in identifying and prioritizing AI system risks. It provides a structured approach to assess failure modes based on severity, likelihood, and detectability, offering a quantifiable risk profile for each incident.

(3) Insights into Trustworthiness: By aligning failure modes with the NIST AI RMF, the study highlights how failures impact trustworthiness characteristics such as reliability, safety, and accountability. This alignment supports more targeted risk mitigation strategies.

(4) Domain-Agnostic Methodology: Although the study focused on consumer AI incidents, the FMEA-based approach is inherently adaptable to other high-stakes sectors such as healthcare, critical infrastructure, and defense. Incorporating domain-specific data sources—like regulatory filings and audit reports—can enhance the contextual relevance of the analysis.

(5) Recommendations for Risk Mitigation: The findings suggest that proactive identification and mitigation of high-risk failure modes can significantly improve the reliability and trustworthiness of AI systems. This includes better design practices, operational safeguards, and continuous monitoring.

As part of future work, further analysis of the incidents with high-risk priority numbers due to their severity, likelihood, and detectability needs to be conducted to provide recommendations for corrective actions to lower the risk. Additional methods and techniques for risk assessment and management shall be incorporated.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] A. Drapkin, "AI Gone Wrong: An Updated List of AI Errors, Mistakes and Failures," https://tech.co/news/list-ai-failures-mistakes-errors, 2024.

[2] ISO/IEC 31000:2018 Risk Management — Principles and Guidelines, ISO/IEC, 2018.

[3] "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," National Institute of Standards and Technology (U.S.), Gaithersburg, MD NIST AI 100-1, 2023.

[4] "ISO/IEC 23894:2023 Information Technology — Artificial Intelligence — Guidance on Risk Management," ISO/IEC, 2023.

[5] B. Xia, et al., "Towards Concrete and Connected AI Risk Assessment (C^2AIRA): A Systematic Mapping Study," Proceedings of the IEEE Conference on AI Engineering – Software Engineering for AI (CAIN 2023), IEEE Press, pp. 104-116, 2023.

[6] N. R. Tague, The Quality Toolbox, 2nd Ed., Milwaukee, ASQ Quality Press, 2005.

[7] P. Haapanen and A. Helminen, "Failure Mode and Effects Analysis of Software-Based Automation Systems," Radiation and Nuclear Safety Authority (STUK), Technical Report STUK-YTO-TR-190, 2002.

[8] M. Takahashi, R. Nanba, and Y. Fukue, "A Proposal of Operational Risk Management Method Using FMEA for Drug Manufacturing Computerized System," Transactions of the Society of Instrument and Control Engineers, vol. 48, no. 5, pp. 285-294, 2012.

[9] S. McGregor, "Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 17, pp. 15458-15463, 2021.

[10] M. Wei and Z. Zhou, "AI Ethics Issues in Real World: Evidence from the AI Incident Database," Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS-56), IEEE Press, pp. 4923-4932, 2023.

[11] A. Jobin, M. Ienca, and E. Vayena, "The Global Landscape of AI Ethics Guidelines," Nature Machine Intelligence, vol. 1, no. 9, pp. 389-399, 2019.

[12] A. Kaun, "Suing the Algorithm: The Mundanization of Automated Decision-Making in Public Services Through Litigation," Information, Communication & Society, vol. 25, no. 14, pp. 2046-2062, 2022.

[13] B. Heinrichs, "Discrimination in the Age of Artificial Intelligence," AI & SOCIETY, vol. 37, pp. 143-154, 2022.

[14] K. Brecker, S. Lins, and A. Sunyaev, "Why it Remains Challenging to Assess Artificial Intelligence," Proceedings of the 56th Hawaii International Conference on System Sciences, pp. 5242-5251, 2023.
[15] S. S. Chanda and D. N. Banerjee, "Omission and Commission Errors Underlying AI Failures," AI & SOCIETY, vol. 37, pp. 937-960, 2024.
[16] Y. Wen and M. Holweg, "A Phenomenological Perspective on AI Ethical Failures: The Case of Facial Recognition Technology," AI & SOCIETY, vol. 39, pp. 1929-1946, 2024.
[17] J. Li and M. Chignell, "FMEA-AI: AI Fairness Impact Assessment Using Failure Mode and Effects Analysis," AI and Ethics, vol. 2, no. 4, pp. 837-850, 2022.
[18] L. T. Ostrom and C. A. Wilhelmsen, Risk Assessment: Tools, Techniques, and Their Applications: 2nd Ed., Hoboken, NJ: John Wiley & Sons, 2019.
[19] M. Feffer, N. Martelaro, and H. Heidari, "The AI Incident Database as an Educational Tool to Raise Awareness of AI Harms: A Classroom Exploration of Efficacy, Limitations, & Future Improvements," Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23), no. 3, pp. 1-11, 2023.
[20] N. Pittaras and S. McGregor, "A Taxonomic System for Failure Cause Analysis of Open Source AI Incidents," Proceedings of the SafeAI 2023 Workshop, vol. 3381, pp. 17-28, 2023.
[21] M. Hoffmann and H. Frase, "Adding Structure to AI Harm," Center for Security and Emerging Technology (CSET), 2023.
[22] OECD, "OECD Framework for the classification of AI systems," OECD Digital Economy Papers, No. 323, 2022.
[23] S. Ahmed, et al., "Examining the Potential Impact of Race Multiplier Utilization in Estimated Glomerular Filtration Rate Calculation on African-American Care Outcomes," Journal of General Internal Medicine, vol. 36, no. 2, pp. 464-471, 2021.
[24] S. Hamza-Cherif, L. F. Kazi Tani, and N. Settouti, "Improving Healthcare Communication: AI-Driven Emotion Classification in Imbalanced Patient Text Data with Explainable Models," Advances in Technology Innovation, vol. 9, no. 2, pp. 129-142, 2024.
[25] G. Kotonya and I. Sommerville, Requirements Engineering: Processes and Techniques, Chichester: John Wiley & Sons, 1998.
[26] C. Carlson, Effective FMEAs: Achieving Safe, Reliable, and Economical Products and Processes Using Failure Mode and Effects Analysis, Hoboken, NJ: John Wiley & Sons, 2012.
[27] A. Meriem and M. Abdelaziz, "Combining Model-Based Testing and Failure Modes and Effects Analysis for Test Case Prioritization: A Software Testing Approach," Journal of Computer Science, vol. 15, no. 4, pp. 435-449, 2019.
[28] C. Spreafico and A. Sutrisno, "Artificial Intelligence Assisted Social Failure Mode and Effect Analysis (FMEA) for Sustainable Product Design," Sustainability, vol. 15, no. 11, article no. 8678, 2023.
[29] A. Agarwal and M. J. Nene, "Addressing AI Risks in Critical Infrastructure: Formalising the AI Incident Reporting Process," 2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, pp. 1-6, 2024.